Optimal Leadership Development for Professionals

Jaason Matthew Geerts

Sidney Sussex College

This dissertation is submitted for the degree of Doctor of Philosophy

Faculty of Education

University of Cambridge

15 May, 2018

A dedication:

For my father for his example and love;
For my mother for her support and love;
For my brother for his courage and love;
For my sister for her confused enthusiasm and love;
And for my (late) Opa for his unwavering faith and love.
I hope this makes you proud.

"if I were a wise person, I would do my part."

In the Bleak Midwinter,
quoted by the Queen, 25 December, 2012

I am a researcher, and this is (the beginning of) my part …

DECLARATION OF ORIGINALITY

This dissertation is my own original work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other university or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit of 80,000 words.

Jaason Geerts

ACKNOWLEDGEMENTS

When Emerson wrote, “What lies behind us and before us are tiny matters compared to what lies *within* us,” he overlooked the most important finale, “and who is around us through it all.”

It is with that level of tremendous gratitude that I would like to acknowledge the invaluable support of the following:

First, my family, of course, which is why this is dedicated to them.

Thank-you to my supervisor, Panayiotis Antoniou, as well as Peter Gronn, Philip Stiles, Jan Vermunt, Stevie Agius, and the brilliant power couple: Amanda Goodall and Andrew Oswald.

Thanks to my mentor and friend, Helen Taylor. I owe a lot of this to you.

I appreciate the guidance and support of my college tutors, Iain Black, Jo Craigwood, and Berry Groisman. I am also grateful to Emma Rixon, Freya Villis, Nidhi Singal, Helen Duncan, John Harding, and my “favourite people” for making sure that I had a room I loved each year: Angela Parr-Burman, Karolyn Duke, Suzanne Flack, and Annette Secker.

I appreciate the dedication of my research collaborators: Liam Sanio, Nina Vahra, Chidera Ota, and the technical support from Eric Ferreira, Anna Blakney, Kitty Malone, and Karyn Pickles.

The contributions of Subject Matter Experts have added tremendously: Brig Greville Bibby, Charlie Brown, Maj Gen Julian Free, Wendy McDonald, Dr Bryce Taylor, and Maj Gen Peter Williams.

I would be remiss not to thanks those who made this happen in the first place: Joe Brisbois, Greg Rogers, Moira McQueen, and Amy Rebecca (07/2011).

I would like to sincerely thank the Cambridge Home and European Scholarship Scheme for funding me all these years. The next round is on me.

Finally, to my friends who were most involved in this part of my life: Stephen Young, Alice Williams, Ryan Geerts, Stephen Young, Alexander Hackmann, Aline Hackmann, Alex Ross, Professor Andy Evans, and Stephanie.

I hope that whatever ‘tiny matters’ are before us in the future, we will remain as blessed as we are now.

ABSTRACT

OPTIMAL LEADERSHIP DEVELOPMENT FOR PROFESSIONALS

Jaason Geerts

Leadership development is a widespread and burgeoning global enterprise, as well as a rapidly growing field of academic study. An estimated $50 billion is spent on leadership programmes annually (Kellerman, 2012) and yet, there is a large degree of confusion regarding what is known regarding optimal approaches, especially those that are tied to organisational outcomes. There is further confusion in terms of the evidence to reinforce such claims, as well as effective forms of measuring leadership, particularly after interventions. The aim of this dissertation is to address those two topics, as well as to assess the current state of literature in terms of leadership development for professionals.

A novel methodology was employed called a systematic evidence analysis (SEA), which isolates multiple data sets and involves several stages and layers of analysis. This study involved three separate, but related literature reviews to generate these data sets. The first was a systematic review of leadership development for professionals in multiple domains that identified 56 studies. The second was a review of existing literature reviews on leadership development for physicians that included one non-systematic and six systematic reviews. The third was a systematic review of leadership development for physicians that included 25 studies. A validated instrument, the Medical Education Research Study Quality Instrument (MERSQI), was applied to each of the 25 aforementioned studies to critique their quality. Categories of evidence groupings were then devised based on commonalities among the included studies' designs. The categories of evidence are: strong, good, moderate, limited, and anecdotal. Further stages of analysis involved investigating two of the conclusions from the best available studies in detail, as well as developing a prototype theoretical model of leadership development and evaluation.

The results are that the overall quality of literature is quite low. None of the 25 studies qualifies as strong evidence, two are good evidence calibre, four are moderate, and the remaining 19 are either of either limited or anecdotal quality. The overall mean was in the anecdotal calibre range. Likewise, there were common flaws in the seven literature reviews that were analysed, including failing to tier the findings and conclusions according to the quality of evidence. Conclusions from the strong and moderate evidence studies include that workshops followed by videotaped

simulations with expert feedback can improve observable leadership behaviour and contribute to self-awareness. Action-learning is effective in enabling participants to achieve organisational and benefit to patients/clients outcomes, among others. Leadership development has been found to lead to a variety of individual outcomes, such as increased confidence, self-efficacy, and career advancement.

Further analysis revealed that Knowles's (1984) principles of adult education is perhaps the most common educational theory applied to leadership development design. This thesis adapted and expanded his theory by adding two principles, as well as providing examples from the included studies. A second finding was explored in detail, which is the collection of factors before, during, and after interventions that facilitate or inhibit the application of leadership following programmes. These are important not only to enhance the impact of programmes, but to avoid common pitfalls that led several programmes to fail. The beginnings of a theoretical model are offered concerning the cardinal and complementary functions of different developmental activities, which can maximise their utility, especially in reference to specific programme objectives. Another product of the systematic evidence analysis is an outcomes-based prototype theoretical model of leadership design and evaluation. Finally, elements of quality research design and evaluation are presented, as is an overarching proposal to ameliorate the thin nature of the evidence in the field.

The conclusions suggest that the state of the literature in the field needs to be improved. This can be done through a combination of stronger individual study and literature review research designs, better reporting, and tiered findings and conclusions based on the quality of the evidence. Outstanding specific gaps in, or extensions of, the knowledge base are included. This thesis provides a clear and transparent elucidation of what is known in terms of optimal leadership development for professionals and the evidence to reinforce it, which can potentially inform practitioners and serve as the foundation for further research. Similarly, those designing and delivering programmes can potentially use aspects of the two conclusions explored, as well as the two theoretical models, to guide their interventions. The intention is that doing so could increase the impact of programmes, as demonstrated by improved outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CEO – Chief Executive Officer

MLCF – Medical Leadership Competency Framework

NHS – National Health System (UK)

ROI – Return on Investment

PPEs – Post-Programme Evaluations

GMAC – Graduate Management Admissions Council

WHO – World Health Organisation

SEA – Systematic Evidence Analysis

SLR – Systematic Literature Review

MULTI SLR/MULTI – Review of leadership development in multiple professional domains

HEE – Health Education England

HEE SLR/HEE – Review of leadership development for physicians

EMD SLR/EMD – Review of extant literature reviews of leadership development for physicians

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

MERSQI – Medical Education Research Study Quality Instrument

MSF – Multi-Source Feedback

NOS – Newcastle-Ottawa Scale

CASP – Critical Appraisal Skills Program

PDP – Personal Development Plans

RCT – Randomised controlled trial

GDP – Gross Domestic Product

MD – Medical doctor

PICOS – Participants, Interventions, Comparison, Outcomes, and Study Design framework

NR2GP – Nonrandomised, two-group designs including a control group

SGPP – Single group pre and posttest

SGCS – Single group, cross-sectional or posttest only

CV – Curriculum vitae

BEME – Best Evidence Medical Education

JSCSC – Joint Services Command and Staff College (UK)

IOM – Institute of Medicine (US)

MES – Medical Engagement Scale

CMO – Chief Medical Officer

IHQ – Index of Hospital Quality

OPTIMAL LEADERSHIP DEVELOPMENT FOR PROFESSIONALS

JAASON GEERTS

1 Chapter One: Introduction

Leadership development is a widespread and burgeoning global enterprise, as well as a rapidly growing field of academic study. There is some debate surrounding whether leaders are born or bred;[1] however, without discounting the effect of natural qualities or abilities, there is an increasing belief that leadership can be improved (Goodall & Stoller, 2017; McCall & Morrison, 1988; Pfeffer, 2016) and that development programmes are to some extent effective in enabling people to become better leaders (Husebø & Akerjordet, 2016; Komives, Nance, & McMahon, 1998; McAlearney, 2010; Rose, 2015; Rosenman, Shandro, Ilgen, Harper, & Fernandez, 2014; Sanfey, Harris, Pollart, & Schwartz, 2011). The evidence for this claim is that leadership manifestly *is* being taught in military academies, business schools, international corporations, and other institutions around the world on a large scale (Solansky, 2010). For reasons that will be discussed further on, this list has now grown to include healthcare centres. Physicians are taking on leadership roles with increasing frequency in response to an imminent need in the field, prompting a need for leadership development specific to the healthcare domain (McKimm & Swanwick, 2011; Straus, Soobiah, & Levinson, 2013). Despite the substantial and proliferating number of leadership development programmes and expanding body of research, there are significant gaps in the academic literature regarding the most effective ways to design, deliver, and evaluate these interventions. This dissertation investigates what is known regarding optimal principles of leadership development, as well as the evidence to support it, with a focus on physicians and other professionals.

The term "**professional**" in this sense is not limited to the traditional spheres of academia, medicine, law, and clergy whose members have been typically distinguished by their education, esoteric knowledge, complex skill, and prestige (Freidson, 1983). There is a lack of consensus concerning the definition of the professions, as well as the defining characteristics and attributes of their members (Freidson, 1983). This debate becomes more complex if one agrees with Freidson that these are evolving historical concepts, not static ones. An expanded definition of professionals for this study is included in chapter two; however, suffice to say that

[1] Two interesting studies involving identical twins (who share 100 per cent of their DNA) suggest that more than two-thirds of the variance in leadership role attainment is attributable to non-shared environmental factors, with the genetic factor (or heritability) accounting for only 30 – 32 per cent (Arvey, Rotundo, Johnson, Zhang, & McGue, 2006; Arvey, Zhang, Avolio, & Krueger, 2007).

it extends to people whose careers involve occupying a leadership role in a corporate organisation, such as chief executive officers (CEOs) and military officers.

As will be explained in more detail in chapter four, there are several benefits to focusing on physicians. Leadership development for doctors is a blossoming enterprise and yet, although the medical profession is long-standing and widely respected, formal training of doctors as leaders is relatively new in the past few decades and the body of research, while growing, is limited (Dine, Kahn, Abella, & Shea, 2011; Ireri, Walshe, Benson, & Mwanthi, 2011; Lee, 2010; McAlearney, 2010). A review of the established literature demonstrates a clear need in this domain for specific evidence supporting optimal practices to fill research gaps and address practical questions. From a conceptual point of view, as explained in chapter four, studying doctors is valuable because of the parallels between leadership in healthcare and leadership in other domains. Many physician leaders operate within clear organisational structures and face demands similar to professionals in other domains to make decisions in high-pressure, high-stakes environments (B. Taylor, 2010). The findings from this thesis are intended to be transferrable to other organisations and professional domains. Finally, medical leadership development is directed toward a clear ultimate goal of improved patient outcomes; whereas in other domains, such as business, there can be a range of goals, including increased profit, benefit to clients, etc, without a single, universally-accepted one.

Since modern-day healthcare is often delivered by complex teams and physicians typically work with multiple teams that are frequently forming and changing (B. Taylor, 2010), some consider the process of **attribution** in medical leadership development challenging. Despite this complexity, there are several reasons why benefits to patients represents the ultimate outcome for healthcare leadership development. The first reason is that the ultimate purpose of leadership development generally is not personal development alone; it is *application* to the workplace (Edmonstone, 2013; Raelin, 2011). Similarly, the definition of leadership in chapter two stipulates that leadership is not an individual enterprise; it necessarily involves leaders and team members working together. This suggests that measuring the impact of leadership development interventions should not be restricted to individual-level outcomes. To take the two points mentioned together, this indicates that the ultimate goal of leadership development research and programme evaluation is to demonstrate sustained improvement at the team, organisational, and, in the case of healthcare, clinical, levels (Edmonstone, 2011; Nakanjako et al., 2015). This priority is reflected in Husebø and Akerjordet (2016) and Weaver et al. (2014) focusing their reviews on evidence of leadership development impact on patient outcomes.

There are also a few ways to address the challenges associated with tracing attribution to individuals and teams working in complex environments. The first is that there are outcomes one can measure such as improving workplace satisfaction (Kirkpatrick Level 4a), which are *indirectly* related to benefit to patients, since there is evidence correlating the two (Jeon, Simpson, Chenoweth, Cunich, & Kendig, 2013). Second, especially for programmes that allow participants to select their own goals, in accordance with the modified set of principles of adult learning explained in chapter six, each participant can identify clinical outcomes herself which she can reasonably aspire to improve. This will vary according to specialty, role, realm of influence, and specific needs of her workplace. For example, a physician CEO might select implementing a hospital-wide policy change for quality improvement as level 4a and 4b outcomes; whereas, a resident may select improving one clinical outcome on his ward. Regardless, each delegate is able to identify quantifiable outcomes that can be used as programme goals. This can equally be satisfied by choosing action learning projects, as evidenced in many of the HEE included studies. Finally, as contended above, no matter how complex the workplace environments, benefit to patients is the ultimate outcome of healthcare (Lee, 2010; B. Taylor, 2010). By extension, this suggests that it should be the ultimate outcome for healthcare leadership development. Therefore, despite the complex nature of healthcare, benefits to patients should be kept at the forefront of programme and individual goals; and yet, as will be demonstrated in chapter five, much of leadership development evaluation is restricted to individual-level outcomes.

The study of physician leadership development is therefore ideal based on a practical need in the field and research community, conceptual parallels to other professional domains, and an accepted and quantifiable desired preeminent outcome.

The use of the term "**optimal**" in the title of this thesis is a deliberate choice and is preferable to "effective." Although most programmes (though not all, as will be demonstrated in chapter six) are well evaluated in terms of effectiveness, there are definite opportunities to improve their quality and yield. For example, in the Satiani et al. (2014) study, though all participants stated that they would recommend the programme to others, many provided negative feedback regarding the specific sessions and assignments. This suggests that while the enterprise itself was considered valuable, there is room to ameliorate. This study examines leadership development programme outcomes at the individual, organisational, benefit to patient/clients, and economic levels, seeking to answer the question, "What are the best outcomes that interventions can successfully achieve?" To explore the different aspects of "optimal," this thesis also explores factors related to programme samples such as stage of career or level of seniority and mixed versus single professions, faculty (internal versus external), programme details (such as their length, location, and structure), and developmental activities (such as lectures and coaching). These factors are discussed in detail in the statistical analysis findings section of chapter six. Given that professionals' time comes at a premium, maximising the impact of leadership interventions is crucial (Fernandez et al 2016), which requires ensuring that these interventions are as efficient and beneficial as possible. For this reason, the central goal of this thesis is to identify evidence of optimal, beyond just effective, leadership development.

1.1 Background to Leadership Development: Investment and Number of Programmes

There is a plethora of leadership development programmes being offered worldwide (Collins & Holton III, 2004). Kellerman (2012) estimates that, annually, $50 billion is spent on them, which is nearly half the amount of money spent of cancer treatment around the world (QuintilesIMS Institute, 2017). This intensifies the pressure to ensure that these programmes are as effectual and efficient as possible. The number of programme providers and consultancies is constantly increasing (Sahlin-Andersson & Engwell, 2002). In addition to the amount of money invested, there are significant time commitments devoted to planning, delivering, and undertaking leadership development programmes, which carry with them a significant opportunity cost for those involved. McAlearney et al. (2005) postulate that this trend is a direct result of heightened appreciation of the importance, perceived effectiveness, and feasibility of leadership development (Sonnino, 2016). In addition to available private and corporate programmes, leadership development in medicine is being instituted nationally in many countries. Canada, the United States, Denmark, and the UK, for example, have recently introduced formal leadership learning objectives for physicians in medical schools and hospitals. In the UK, for example, every medical school is now required by law to ensure that its students have demonstrated all the outcomes in the Integrating the Medical Leadership Competency Framework (MLCF) by the time they graduate (Collins & Holton III, 2004). Furthermore, professional standards at all levels in the National Health System (NHS) now include a reference to leadership (McKimm & Swanwick, 2011). The number of leadership development programmes and the amount of time and money invested in them is increasing in many domains, including the field of healthcare.

1.2 Lack of Research on Effectiveness

Despite the prevalence of and surge in the number of programmes and the prescriptions of the vast leadership industry, relatively little is known about the impact of leadership development programmes (Hannum & Bartholomew, 2010; Ireri, Walshe, Benson, & Mwanthi, 2011; Klimoski & Amos, 2012; Straus et al., 2013), or about optimal principles of their design, delivery, and evaluation (Goodall & Stoller, 2017; Pfeffer, 2016). Powell and Yalcin (2010) describe the situation as centring around a lot of discussion and advice, particularly in popular literature, with very little information grounded in empirically-based, scientific research (D. V. Day & O'Connor, 2003; D. V. Day & Sin, 2011). Stanford professor Jeffrey Pfeffer (2015) adds that most of the available information is "wonderfully disconnected" from organisational realities, rendering it useless for sparking improvement. Kellerman (2012) adds that because what she calls "the leadership industry" is so bereft of empirical evidence, it is impossible to confirm that "this massive, expensive, thirty-plus-year [leadership development] effort has paid off" (p. 168). Beer et al. (2016) describe the poor return on investment (ROI) in leadership development as the "great training robbery" and Gilpin-Jackson and Bushe (2007) note that there is evidence that the overall proportion of leadership transferred to the workplace is low. Likewise, Pfeffer (2015) adds,

> It is not just that all the efforts to develop better leaders, decades of such effort notwithstanding, have failed to make things appreciably better. I realised that much of what was and is going on almost certainly, although sometimes inadvertently and unintentionally, makes things much worse (p. 5)

This lack of an established credible evidence base has led some commentators to question the relevance and the worth of the yield of such programmes (Blume, Ford, Bladwin, & Huang, 2010; K. E. Watkins, Lysø, & deMarrais, 2011) and has generated scepticism, causing many managers to view leadership development as a low priority (Avolio, Avey, & Quisenberry, 2010).

There are further gaps in the research findings in terms of the individual aspects of leadership development and its benefits. First, there is a paucity of evidence regarding what specific knowledge or which capabilities might enhance individual, team, or organisational performance (Allio, 2005; Ardts, Velde, & Maurer, 2010; Collins & Holton III, 2004; DeRue & Wellman, 2009; Ireri et al., 2011; Straus et al., 2013). Likewise, there are few empirical studies available that outline which developmental activities, individually and collectively, are effective (Allen & Hartman, 2008; Collins & Holton III, 2004; Suutari & Viitala, 2008). It is also largely unclear in what ways interventions impact on organisational performance, especially since leadership development often features a combination of formal and informal packages (Galli & Muller-Stewens, 2012). Another gap in the research is related to effective metrics and approaches to measuring leadership, particularly after programmes are complete. Hartley and Benington (2010) suggest that much of the existing research features cross-sectional designs, which preclude establishing causal or correlational links between interventions and outcomes or ruling out alternative explanations. As mentioned above, given the relative newness of formal leadership development for doctors, it is not surprising that there is an insufficient body of work in this area, despite its rising popularity (Dine et al., 2011; Ireri et al., 2011; McAlearney, 2010). The result is a selection of interventions for physicians that Ireri et al. (2011) describe as "scanty and ad hoc" (p. 18), Leslie et al. (2005) assert is "sporadic and rudimentary" (p. 766), and Satiani et al. (2014) state that "what passes as leadership development in some hospitals and medical schools is a hodge-podge of classes and lectures lacking coherence, logical progression, comprehensiveness, and relevance" (p. 542). The episodic nature of the instruction leads the latter authors to conclude that such programmes are rarely successful in developing effective physician leaders. Taken together, these research deficiencies demonstrate a need to justify investment in leadership development by evaluating how such programmes impact individual and organisational effectiveness and outcomes (Galli & Muller-Stewens, 2012).

A further challenge is that the information in the literature concerning the effectiveness of leadership development programmes is equivocal and at times conflicting, as will be shown in detail in chapter five. Many meta-analyses and individual studies report that programmes are effective (Collins & Holton III, 2004; Frich, Brewster, Cherlin, & Bradley, 2014; Zhang, 1999), but others indicate that certain programmes in their sample failed miserably, citing effect sizes that ranged from -1.39 to 2.10 (Collins & Holton III, 2004). For example, Ireri et al. (2011) state that many doctor managers in their study claimed that leadership training added little to their existing knowledge. Likewise, Mabey and Thompson (2001) report that only 19 per cent of the companies in their survey achieved their leadership development objectives, while 37 per cent performed poorly or did not succeed at all. It is difficult for readers of these studies to determine why there is such an effect size range and to ascertain whether this effect can be attributed to the quality of the individual programmes, differing measurement metrics, or other factors, such as organisational culture. A further point of confusion is that many authors suggest that the programmes they studied were successful but provide no objective data to support those claims (Guskey, 2002; Hartley & Benington, 2010). Therefore, despite widespread belief in the benefits of leadership development as a concept, the impact of individual programmes on desired outcomes, particularly beyond the individual-level and in the long-term, as well as what knowledge, specific capabilities, and intervention designs are most effective and in what ways, is still largely unclear.

1.3 Lack of Evaluation and Consequences

While the limited empirical evidence supports leadership development programmes in general, the clear inadequacies in the research are partly attributable to the fact that the majority of these programmes are not being evaluated effectively (Collins & Holton III, 2004; Hartley & Benington, 2010; Ireri et al., 2011). Indeed, many programmes are not being evaluated at all (Amagoh, 2009; Groves, 2007; Van Aerde, 2013; Vardiman, Houghton, & Jinkerson, 2006). Avolio (2005), for example, estimates that fewer than ten per cent of organisations that invest in leadership development ever actually evaluate the programmes in terms of performance outcomes. Collins and Holton (2004) and Allio (2005) suggest that this is the case because many organisations either do not devote sufficient funding to long-term evaluation of programmes or blindly assume that leadership development interventions translate into positive organisational outcomes (Russon & Reinelt, 2004). Another possible reason for the lack of evaluation is that people can develop evaluation fatigue or frustration, especially if past instances of gathering such information have been burdensome and time-consuming, without demonstrating clear benefits. Similarly, MacPhail et al. (2015) acknowledge that although formal assessment in their study would have improved the evaluation, it was viewed as a disincentive or hurdle to participation, so it was not done. Further postulations are that evaluation can be risky politically, whether because of differing stakeholder priorities or worries that negative feedback regarding a flagship programme might result in budgetary cuts or professional discredit for those who designed them (Hartley & Benington, 2010; C. Mabey & Finch-Lees, 2008). It is also possible that for some organizations, the true purpose of leadership development lies in its latent functions, whereby such programmes serve as an aspect of branding and institutional prestige, a recruiting tool, a required medium for advancement, or to whet participants' appetite for future interventions, such as an MBA. In such cases, the intricacies of the programmes would be secondary, and evaluation would run the risk of challenging a source of corporate pride or strategy. While there are clearly multiple possibilities as to the cause, it is apparent that much leadership development is either poorly evaluated or not evaluated at all.

Kellerman (2012) asserts that the majority of leadership programmes that are evaluated rely totally on one subjective measure: whether or not participants are satisfied with the programme. These measures, called Post-Programme Evaluations (PPEs), provide no evidence regarding the transfer of learning to the workplace. While this review will show that Kellerman's appraisal is not entirely accurate, at least in terms of programmes described in the academic literature, there is certainly a lack of objective evidence of the transfer of learning and her assessment reflects the exaggerated perspective that exists regarding the dearth of research in the field. This thesis' unique methodology, to be described later in this chapter, was designed to clarify the state and scope of evidence. The breakdown, frequency, and effectiveness of different ways of evaluating leadership development programmes will be discussed in detail in chapters five and seven. As mentioned earlier, many claims have been made that programmes were effective without supporting evidence to legitimise these conclusions, while other evaluations are vague about how "effectiveness" was defined (Guskey, 2002; Hartley & Benington, 2010). Leslie et al. (2005) add that the same fuzziness exists in some studies that draw correlations between leadership development in medicine and quality of care. As intimated earlier, part of the impediment facing practitioners and researchers alike is that there is no agreed-upon metric or outcome measures for assessing leaders' effectiveness (Clarke, 2012; Fallesen, Keller-Glaze, & Curnow, 2011) or the impact of leadership programmes on organisational outcomes (Allen & Hartman, 2008; Collins & Holton III, 2004; Dexter & Prince, 2007). Another challenge in quantifying leadership outcomes and isolating correlative and causal links is that leadership development often takes place in complex, uncontrolled environments in which various developmental activities are used together (Guskey, 2002; Sanfey et al., 2011). Therefore, the complex nature of the phenomenon itself and the lack of standardised metrics offer additional challenges to measuring the impact of leadership development programmes.

Overall, this lack of evaluation of leadership development programmes inhibits the collection of valuable information that could aid researchers and programme designers in optimising interventions. Given the significant and growing investment in leadership development, Bolden (2005) asserts that it "seems crazy" to design and deliver programmes based on insubstantial evidence (p. 48). An illustration of this trend is Klimoski and Amos's (2012) study of 48 elite Graduate Management Admissions Council (GMAC) MBA programmes, which suggested that few were guided by a well-articulated, research-based pedagogical framework or effectively assessed whether programme components translated into desired outcomes. Boaden (2006) concludes that this lack of evaluation results in leadership development becoming "sporadic, haphazard, and illogical" (p. 9), a notion supported by Leslie et al. (2005). It also exposes programmes to the danger of stagnation through the repeated use of ineffective or suboptimal means by not identifying problem areas (Rousseau, 2006), not to mention wasted time for the participants. This lack of assessment and publication of programme evaluations also precludes organisations from sharing wisdom in order to evolve concomitantly. Since leadership development is widely considered to be a source of competitive advantage in business, it could be argued that another explanation for this lack of publicised evidence is an aversion to supplying one's competitors with information. The fact that many organisations are not collecting and analysing data on their programmes even for their own purposes, however, erodes the steel of this argument. Although there can be apprehension towards evaluation, there is a common interest among participants, providers, and organisations in demonstrating the impact of leadership development programmes and their return on investment (ROI) (Beer et al., 2016). As will be shown throughout this dissertation, particularly in chapter seven, one effective way of demonstrating this impact is by linking evaluation metrics to performance outcomes at the individual, team, and organisational levels. The consequence of the aforementioned gaps in the research, therefore, is that while leadership development programmes are numerous, their designs are seldom based on credible research or thoroughly assessed, decreasing the likelihood that their effectiveness is being optimised.

1.4 Need for the Proposed Study and Significance to the Field

1.4.1 Importance of Leadership

A recent survey of 5,561 executives from 109 countries identified the improvement of leadership development as the most important human resources priority for organisations around the world (DeRue & Wellman, 2009). One likely explanation is the all-too-frequent pervasive leadership failures across professional domains (Pfeffer, 2016). A second related explanation is the widespread belief in the importance of effective leadership. In healthcare, "clinical leadership" is described as the core business of everyday medical care and public health and is critical to staff engagement, improved clinical, financial, and operational performance, as well as the delivery of high-quality care (CMO Clinical Advisor Alumni, 2012; Dine et al., 2011; Jeon et al., 2013; Kim & Thompson, 2012; Squazzo, 2009). Bruce Barraclough, Clinical Lead and Chair of the World Health Organisation (WHO) Patient Safety Curriculum Guide, agrees, writing that effective leadership is the essential ingredient necessary to acquire the resources, improve quality, address risks, and provide the safest and best possible care in the complex environment of modern day healthcare (in Taylor, 2010). Finally, Jones, McCay, and Keogh (2011) suggest that in the UK effective leadership is central to implementing National Health Service (NHS) reforms, which explains why physician leadership was prioritised in the 2008 NHS review (Darzi, 2008; Horton, 2008). With these points in mind, Straus, Soobiah, and Levinson (2013) conclude that the role of the physician leader simply cannot be overemphasised. Therefore, effective leadership is considered to be tied to a range of positive clinical, financial, and operational outcomes in the field of healthcare.

1.4.2 Importance: Evidence Correlating Leadership and Outcomes

A limited number of studies suggest that effective leadership translates into identifiable organisational outcomes. For example, BusinessWeek's world's "Best Companies for Leadership" (BusinessWeek/Hay Group, 2010) consistently outperformed others in sales growth and value creation over one, three, five, and ten year periods (Thomas, Jules, & Light, 2012). Research also suggests that there is a strong connection between effective leadership and increased employee satisfaction (Doran et al., 2004; Gagnon et al., 2006; Hayes, 2007; Jeon et al., 2013; Artz, Goodall, & Oswald, 2016), including physician job satisfaction and well-being (Shanafelt et al., 2015), employee retention (A. Baker & Goodall, 2017; Doran et al., 2004; Gagnon et al., 2006), employee motivation, commitment, and a sense of shared purpose (Bolden, 2005), and customer satisfaction (Doran et al., 2004; Gagnon et al., 2006; Hayes, 2007; Jeon et al., 2013). Further research suggests that employee job satisfaction is believed to positively influence organisational performance (Bryson, Forth, & Stokes, 2017; Combs, Liu, Hall, & Ketchen, 2006; Jiang, Lepak, Hu, & Baer, 2012; A. Oswald, Proto, & Sgroi, 2015; Peccei, Van de Voorde, & Van Veldhoven, 2013; Van de Voorde, Paauwe, & Van Veldhoven, 2012). Similarly, there is a reported correlation between physicians' job satisfaction and resulting performance to patient outcomes (Halbesleben & Rathert, 2008). In another case, Mannion et al. (2005) found that a key point of divergence between high and low-performing hospitals in England was leadership and management orientation. Sarto and Veronesi (2016) suggest a link between physician leadership and financial and resource management, as well as quality of care.

Furthermore, there is a growing number of reports in the literature tracing connections between medical leadership development programmes and significantly improved patient safety (Edmonstone, 2011; Jeon et al., 2013; McAlearney, 2010). A McKinsey report describes a quality improvement initiative in a dozen UK hospitals, which led to as much as a 30 per cent drop in lengths of stay, mortality rates, and costs (Mountford & Webb, 2009). Husebø and Akerjordet (2016) also note in their review that researchers behind two quasi-experiments reported a significant decrease in clinical error rate following a team-based intervention. Similarly, others have produced strong evidence that leadership interventions can positively impact on a variety of patient outcomes (Kunzle, Kolbe, & Grote, 2010; Strasser et al., 2008; Weaver, Dy, & Rosen, 2014). Finally, Spurgeon et al. (2011) claim there is increasing evidence to suggest that healthcare organisations in which doctors are more engaged with maintaining and enhancing the performance of the organisation as a whole perform better financially and clinically. The correlation between effective leadership and improved individual and organisational outcomes demonstrated in this small number of studies provides an indication of the kind of work required to advance this field.

Building on the notion that leadership is considered important for effective, high-quality and cost-effective medical care (Edmonstone, 2011; Jeon et al., 2013), McAlearney et al. (2005) state that "developing physician leaders in medicine is *essential*" (p. 11, original emphasis). Ireri et al. (2011) agree, affirming that "leadership development for frontline leaders is critical to the sustainability of the healthcare industry" (p. 18). Martins (2010) suggests that without structured training in leadership, there is a the risk that aspects of doctors' practice will be left to trial and error or remain undeveloped, resulting in underperformance, which could ultimately jeopardise patient safety (B. Taylor, 2010). Thus, while there is a growing appreciation of the value of effective leadership and leadership development, further research is needed to ensure that such programmes are empirically-based and optimised.

Given the various, clear gaps in the academic literature and the importance of leadership and its development, this thesis intends to address the questions of what is known regarding optimal leadership development for professionals and the evidence that exists to reinforce this. As will be described briefly below and in detail in chapter two, this will be done by way of a systematic evidence analysis (SEA).

1.4.3 Background to Medical Leadership

Understanding the social and historical context of a research phenomenon is important. Focusing on leadership development for physicians is particularly timely as the Canadian Royal College of Physicians asserts that the medical profession is at a turning point in its history, principally due to the two interrelated concepts of quality control and leadership (B. Taylor, 2010). The current situation in medical care demands a new kind of physician leader for three main reasons: the increased use of medical technology and budget concerns, ensuring patient safety, and doctors assuming senior leadership roles (Lee, 2010). An expanded explanation of these factors is included in the appendix on page 334 for the readers' convenience.

1.4.4 Need for Doctor Leaders and Development Programmes

Lee (2010) argues that the changing landscape of the field of medicine requires a fundamentally different approach along with "a new breed of leaders" (p. 52). Taylor (2010) asserts that this is necessary to guarantee the "well-being of the profession and certainly that of the patient" (p. 3). Shah et al. (2013) and McKenna and Pugno (2006) echo this point, the latter asserting that given the current state of healthcare, the need for physician leaders is urgent: "Clinically trained administrators who govern the human and financial resources within healthcare organisations" (quoted in Murdock & Brammer, 2011, p. 52). There are many reasons behind the increased need for effective medical leadership, which are explained fully in the appendix on page 331. A few such impetuses are the number of preventable errors that harm patients, inconsistent diagnoses and treatment, and unsustainable costs (Maccoby, Norman, Norman, & Margolies, 2013). Physician administrators can help decrease these occurrences by implementing systemic protocols such as the inclusion of a surgical checklist, a measure that is now required nation-wide in the UK. Leadership development need not focus exclusively on administrators, however; in fact, Bohmer (2012) suggests that working doctors exercise the most influence over the key processes and microsystems necessary to significantly improve overall health system performance, medical outcomes (eg error rates), and terminal outcomes (eg readmission and mortality rates). One argument for the usefulness of having doctor leaders at the highest levels of trusts and hospitals is that they best understand the inherent tension between cost and patient welfare and can anticipate the potential impact of policy changes (Bohmer, 2012). Moreover, doctors are also in a position to guide politicians to keep health delivery and funding structures focused on patient well-being, providing a strong common purpose for approaching the current challenges facing healthcare systems (Darzi, 2008). Consequently, having effective physician leaders at the administrative level and in clinical settings is seen as the key to preventing medical errors and meeting much-needed targets in healthcare organisations such as the NHS.

Although leadership development programmes are thought to be effective across industries at developing organisational leaders (McAlearney, 2010), Day (2007) and Ireri et al. (2011) suggest that physician leaders often do not have access to the training and support that they need, especially when taking on managerial roles. As an illustration of this point, McKinsey & Company reported that in the UK there are significant skills and knowledge deficits among middle and senior management NHS staff, compared to their counterparts in industry and private health care (Ireri et al., 2011). This situation is not restricted to the UK either; the WHO has identified a deficiency in leadership capacity of many developing countries as a key reason for failure to meet their Millennium Development Goals (CMO Clinical Advisor Alumni, 2012). The result of this phenomenon is that doctors tend to build leadership capability though ad hoc, on-the-job learning, which is not sufficient given the changing demands of the field (Blumenthal et al., 2014). Van Aerde (2013) suggests that simply creating formal leadership positions for physicians within organisations without providing development opportunities is equally insufficient, which is supported by Satiani et al.'s (2014) assertion that the performance of doctor leaders in new roles is often mediocre or worse. Additionally, Ackerly et al. (2011) argue that placing physicians in leadership roles without adequate preparation can result in a loss in confidence in them and limit career development for those who underperform in these roles, or, most concerningly, lead to mismanagement of systems. For these reasons, it is clear to medical leadership proponents that evidence-based, programmatic approaches to clinical leadership development is required at various stages of physicians' careers (Swanwick & McKimm, 2012).

1.4.5 Evidence-Based Pedagogies

Building on the above points, Dugan (2011) concludes that there is a need for "high impact learning pedagogies empirically-proven to make a difference in leadership development" (p. 81). He adds that educators who are versed in leadership theory and "learning pedagogies known to leverage leadership development" (p. 18) are also required. Klimoski and Amos (2012) suggest that although university educators engage in considerable teaching about leadership in business schools and elsewhere, much of this teaching has not yet been subjected to rigorous empirical tests, especially with context in mind (Schyns, Tymon, Kiefer, & Kerschreiter, 2013). Thus, a helpful starting point would be to collect and generate empirical evidence surrounding effective interventions (Klimoski & Amos, 2012). Johnson et al. (2012) add that it is important to identify the conditions under which leadership development is most likely to initiate behaviour change, which is addressed in chapter six. Bolden (2005) argues that this kind of information could guide the design of leadership development programmes and enable the improvement of the quality and precision of current programmes. Edmonstone (2013) suggests that this is also the best way to ensure that the significant investments made in healthcare leadership development, which Kellerman calls into question, yield the best possible benefit. Overall, the evidence demonstrates that, despite the recognised importance of medical leadership, development opportunities are scarce and often unfit for purpose.

1.4.6 Measurements

A final need is for credible metrics to measure the impact of leadership development programmes at the individual, organisational, benefit to patients/clients, and economic levels. This includes the identification of short-term results for funders (Russon & Reinelt, 2004), long-term career development outcomes at various levels of analysis (Hiller, DeChurch, Murase, & Doty, 2011), and indicators of organisation-level performance (Collins & Holton III, 2004; Russon & Reinelt, 2004). It is not for lack of appreciation of their importance that these research gaps exist, since there is a relatively recent movement in favour of evidence-based leadership development (Hamlin, 2010; Klimoski & Amos, 2012), which involves designing interventions based on the highest calibre research available, as an essential part of human resource development. The aforementioned authors suggest that there is widespread interest in putting empirical research into action. Just as physicians and patients alike are unlikely to opt for non-evidence-based healthcare, leadership development should be no different. Finally, Russon and Reinelt (2004) advocate weaving evidence-based insights into an explicit programme theory that maps out how and why leadership development interventions are meant to generate particular outcomes. Many voices are therefore echoing the need for developmental goals, programme components and activities, and forms of measurement that are theory-driven, empirically supported, and consistently evaluated in terms of various levels of outcomes.

1.4.7 Basic Assumptions and Central Research Question

The basic assumptions of this study are that leadership can be, to at least some extent, learned and developed; the impact of development programmes can be measured; evidence-based programmes and measurement tools are more likely to yield better outcomes; and that research linked to performance outcomes is required. Therefore, the principal research question of this thesis is **how is leadership development for professionals made optimal?** The full explanation for the choice of professionals and doctors is provided in the sample section of chapter four.

1.4.8 Research Sub-Questions

The key research sub-questions that arose from the over-arching question are outlined below in Table 1.1.

Table 1.1

Research Sub-Questions

#	Research Sub-Question
1 (Background)	What is the **current state** of the leadership development literature regarding available information relating to professionals what is its calibre?
2	What evidence is available regarding **optimal** leadership development for professionals? This refers to programme components, such as length, developmental activities, such as lectures, facilitators, such as internal versus external, and professional characteristics of the participants.
3	What evidence is available in terms of effective ways of **measuring** leadership, particularly following interventions. This refers equally to effective approaches measurement, as well as to which post-programme outcomes are achievable.
4	What insights can be drawn regarding the nature of leadership in terms of it being generic versus contextual? This refers to **nuances** of the extent to which leadership development transfers naturally among different countries, professions, organisations, teams, roles, and levels of seniority of participants.

Answering the research sub-questions above is intended to provide a richer understanding of the phenomenon than currently exists and to enhance transferability of the findings and conclusions to other contexts.

1.4.9 A Unique Methodology: Systematic Evidence Analysis (SEA) and Its Inception

Although the central research question for this study has remained constant from the beginning of the thesis work, the sub-questions and methodology have evolved as it progressed. Background reading on the topic provided a number of revelations, three of which prompted the formulation of the sub-questions relating to optimal leadership development, measurement, and the generic versus contextual nature of leadership and its development. First, it became clear that until now there have not been adequate answers to these sub-questions despite their centrality in the field. Second, preliminary research revealed that these questions are intimately connected. For example, it would be of limited value to offer a set of principles of optimal leadership development without (a) addressing measurement of outcomes (begging the question: “optimal” in what respect?); (b) discussing the calibre of research from which the principles arise (evoking the question: how can it be trusted?); and (c) exploring the extent to which principles can be confidently applied to other situations and contexts. Third, a significant portion of leadership development research, referring equally to reviews and individual studies, lacks the credibility required to elicit confidence in the results. This point came to light in the initial literature review, which consisted of a systematic review of leadership development for professionals in multiple domains, called MULTI.

Although each of the 56 included studies in this review offered findings and conclusions, they were of varying credibility. Without a systematic and transparent way of evaluating the calibre of studies, it is challenging to make strong conclusions, despite considerable sample size. This revelation sparked the final research sub-question relating to the current state of the literature and made it clear that a novel methodology was needed.

Although the author of this thesis gained access to data and personnel for several leadership development programmes at respected institutions, such as the UK Defence Academy, given the current state of the literature, it seemed unclear how a new empirical data set would fit with, reinforce, expand upon, nuance, or contradict a predominantly equivocal knowledge base. Although there is a significant body of literature on leadership development, exponentially more so when one delves into popular literature, it difficult to ascertain exactly what is known and on what evidence that knowledge is based. For this reason, the decision was made to proceed with a systematic evidence analysis (SEA), a novel methodology which is equipped to answer the key research questions mentioned above, while at the same time providing the ability to comment critically on the state of the literature. Although this study was open to and allowed for discovering innovative ideas within the published literature, that was not its exclusive focus.

The SEA approach begins by identifying data sets, in this case, four: the information in the background reading of non-empirical studies from the HEE and MULTI review, as well as the included studies in the MULTI, EMD, and HEE reviews.

Using these data sets, the SEA methodology was designed to serve four **functions** through multi-level, iterative analysis:

1) To systematically answer the central research question regarding optimal leadership development, as well as identifying the supporting evidence that reinforces it
2) To critically evaluate the manner in which research is being done in the field and make suggestions on how to improve it (in terms of meta-analyses, systematic reviews, and individual studies)
3) By addressing the two points above, to comment critically on the current state of the literature; and,
4) To use the overall data set to explore further relevant topics through deeper-level analysis. In the case of this thesis, these in-depth analytical steps concerned the two conclusions explored and the prototype theoretical model, which emerged from the conclusions from the best available evidence as worthy of further investigation. At

this stage, another review of the overall data set was made through the lens of these three topics.

Approaches to analysing each of the data sets differed. The background articles were primarily used to get a sense of the key issues and gaps in the research. MULTI is a systematic literature review and that was particularly useful for the raw data findings and the conclusions explored. Extant Medical Doctors (EMD) is a review of existing literature reviews on leadership development for doctors. "Extant" in this sense is meant to distinguish that collection of reviews from the third review undertaken for this dissertation, which is a systematic literature review of leadership development for doctors called HEE. The analysis of the extant reviews informed the design of the final review by identifying effective ways of conducting literature reviews in the field and clarifying what knowledge exists in these sources. Content from the extant reviews was compared to the thesis's raw findings, conclusions from the best available evidence, conclusions explored, and the implications for research. Finally, a third review was undertaken, named HEE, because it was done in collaboration with a Health Education England (HEE) fellow in medical education. HEE included a unique feature based on the need for transparency regarding research credibility that to date was found to be lacking. This feature involved applying a validated instrument to critique the calibre of each of the included studies, presenting the full findings in the text, and basing the analysis and conclusions in a tiered manner on the best available evidence. These conclusions formed the heart of this study. This feature will be described in more detail in chapter two.

The four functions of the SEA methodology together were intended to establish a clear and solid foundation of evidence, a result that the most common methodologies of case studies, surveys, or quasi-experiments could not accomplish in as extensive a manner.

In addition, the open-ended nature of the SEA methodology allowed for the formulation of two unexpected contributions to the empirical base. The first is the "conclusions explored" described in chapter six, which consist of an in-depth exploration and extension of two of the findings from the best available studies. The first conclusion explored is how Knowles's (1984) principles of adult learning, the most common theory mentioned in the included studies, apply to leadership development for professionals. Further research uncovered a second, related educational theory, Dale's (1969) Cone of Experience, which describes how different developmental activities serve different key functions. Analysing this thesis's included studies through the lens of Knowles's and Dale's work provided a novel adaptation and application of these two theories, along with two new principles of adult learning. The second conclusion explored produced a novel set of factors in the design, delivery, evaluation, and follow-up of leadership development programmes that are thought to contribute to the transfer of learning. This collection of factors emerged from studies' assertions of best practice, as well as from articles that claimed that the programmes they analysed had failed. Taken together, these points provide key insights for researchers and practitioners alike. Without a meta approach to the topic, the stock and usefulness of such a grouping would have been much more limited. Finally, although this study was not intended to produce a theoretical model, a set of research-verified procedures emerged which fit into a sequence that suggested a prototype for a theoretical model of designing and implementing leadership development. It is possible that a different methodology would not have allowed for this kind of learning unless one researched theoretical models specifically. A systematic evidence analysis enabled this study to address a series of clear needs in the field and to facilitate the detailed, literature-based exploration of further relevant topics using multiple data sets.

The results of this study therefore have potential implications for research, policy, and practice.

1.4.10 Potential Benefits: Research

This thesis has the potential to be beneficial to research in many ways. First, the background research and the application of the validated instrument to the HEE included studies revealed that overall, the calibre of research in the field needs improvement. This thesis offers critiques and examples of individual study strengths and weaknesses based on a large total sample size, as well as recommendations for improving the quality of research that could be applied to future studies. Second, the findings and analysis provide examples of much-needed, effective outcome metrics at the individual, team, organisational, clinical/benefit to clients, and economic levels that could be used by other authors in their work. Similarly, the feature study described in chapter seven, which is the only randomised controlled trial in the included studies, demonstrates how key research components such as economic outcome metrics can be incorporated successfully. Third, by clarifying precisely what is known and based on what evidence, recommendations for topics to be explored in future research spring from a more precise foundation than the more general claims cited in the introduction. Therefore, this thesis offers guidance for future research in terms of individual studies, effective components of quality leadership development research, along with examples, and a clearer sense of what specifically is known and what merits further investigation.

A fourth potential benefit relates to the results of the analysis of extant reviews, which revealed several common weaknesses, the most significant of which was lack of clarity regarding the calibre of evidence to reinforce their conclusions. The application of a validated instrument to assess the credibility of each included study in the HEE review and the grouping of the studies into tiered categories based on commonalities among their designs, along with recommendations regarding ways of improving the instrument, could serve as a resource to improve the quality and transparency of future meta-analyses and reviews. As described in the discussion, this could also be used to encourage better research at the individual study level. Fifth, the testing of the use of statistical analysis to investigate the relationships among variables in the study designs and programme components could identify ways to address the pervasive failure of reviewers to analyse these connections. Contrasting the HEE review to the others (MULTI and those in EMD, all of which share common traits and differences from HEE) demonstrates the gap between the current state of the literature and the kind of research that will advance the field in the future. This study therefore can provide suggestions for improving the calibre of research, including the use of a validated instrument to evaluate individual study quality and statistical analysis, at the meta-analysis and review level.

The final set of potential benefits for research offered by this study relate to the in-depth exploration of the conclusions of the best available evidence, which are described in chapter six. First, the two conclusions explored offer a novel and extensive set of points, along with references and detailed examples, on two key topics, which can be subject to further research and potentially contribute to theory development. Similarly, the prototype theoretical model of optimal leadership development provides an opportunity to test its effectiveness, including against other models. The discussion section of chapter seven offers possible explanations for the thin evidence base in the field, as well as a set of suggestions for improving its quality. Finally, perhaps the most important potential benefit of this study is that amidst a great deal of skepticism and confusion regarding the state of literature in the field, this thesis clarifies in a systematic and transparent way what knowledge is supported by good evidence and which areas require more robust investigation.

As such, the implications for research derived from this thesis could inform academics and those in organisations alike in analysing the outcomes of programmes. By addressing key gaps in the research using a comprehensive methodology, this thesis is intended to generate findings with the potential to inform further research and extend to other contexts.

1.4.11 Potential Benefits: Policy and Practice

The findings from this study also have the potential to benefit policy and practice. First, programme providers could use the best available evidence outlined in the conclusions to influence the design, delivery, and evaluation of leadership interventions and to refine existing programmes. As will be discussed in chapter five, the analysis of this study did not detect patterns across different levels of seniority or varying professional domains, which increases the potential for the findings to be generalised across different contexts. The methodology, including the three separate literature reviews, broadens the knowledge base, as do the points made in the two conclusions explored. These conclusions explored can further inform programme providers and those investing in leadership development as to how to maximise impact by basing programmes on the revised and expanded principles of adult learning described in chapter six and ensuring that the organisational culture is conducive to the transfer of leadership learning to avoid common pitfalls. As well, this study highlights points drawn from studies of programmes that claimed failure, offering information that may help practitioners avoid similar results. Chapter six explores these two conclusions more extensively, with full references and examples, than has been done previously in the literature. Similarly, the theoretical model prototype presented in chapter six could be used by providers as a guide to design new, or re-evaluate existing, programmes. Therefore, the conclusions from the best available evidence, conclusions explored, and the prototype theoretical model presented in this thesis offer credible resources which providers and stakeholders can consult to plan new or improve existing interventions. The findings can help guard against the danger of insular thinking and stagnation by stimulating reflection on existing practice and introducing new ideas, potentially leading to improved, evidence-based curriculum and practices (Boaden, 2006; Klimoski & Amos, 2012; Pradarelli, Jaffe, Lemak, Mulholland, & Dimick, 2016; Straus et al., 2013). The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines state that systematic reviews are being used increasingly by healthcare providers to inform best practice patient care (Liberati et al., 2009). This process could benefit individual organisations while at the same time increasing the evidence base in the field when findings are made public. Gronn (2002) suggests that publishing this kind of information widens the potential for sharing and adapting beneficial practices, avoiding cultural and professional insularity, which he suggests, is in “no-one’s interest" (p. 1065).

Another potential benefit is that reflecting on descriptions of the optimal functions of aspects of leadership development in the first conclusion explored and examples of effective post-programme outcomes can encourage organisations to be intentional about the purpose and role of leadership development, particularly when this concept is identified as an integral part of organisational strategy. This intentionality could include recognising the important role of measuring the impact of leadership development, which many organisations neglect altogether. In healthcare, McAlearney and Butler (2008) explain, this practice involves explicitly outlining how leadership development can contribute to strategic goals of improved efficiency and quality of care. These considerations also apply to the strategic choice or sponsoring of participants for programmes based on different organisational needs. Evidence-based programmes can also enhance the credibility of the field of leadership development as a whole (Russon & Reinelt, 2004).

The conclusions and implications for practice apply as much to through-career leadership development as to individual interventions. The implications for practice and research could potentially influence organisational funding decisions as well as assist independent bodies in deciding which research needs to prioritise. Finally, the analysis of the three reviews can potentially unveil insights into the extent to which leadership development is generic or contextual, an aspect that been largely unexplored in research to date. Information of this kind could influence the nature of programmes offered, such as those that are in-house or external and domain-specific or open, and could suggest the extent to which findings from individual programmes can be generalised to other contexts. This research can therefore benefit policy and practice by offering credible data that providers can draw upon to improve or enhance their programmes in related and separate contexts.

1.4.12 Generalisability

Several aspects of this thesis's methodology are intended to enhance the generalisability of the study. First, the face validity of selecting professionals as the sample increases the potential to extend the findings to other contexts and domains, as will be explained in detail in chapter four. Second, the relatively large sample size of 72 unique included studies provides a wealth of data and examples that have been woven into the findings, conclusions, discussion, and implications for research and practice. Third, the transparency of the methodology, including the publication of critiques of individual studies and reviews, enables readers to judge for themselves the applicability to their own contexts. Fourth, part of this study focuses on physicians, which in the context of the HEE review, could be considered a case study. Several researchers, including Lincoln and Guba (1985) and Stake (1995), assert that when individual cases are described in detail and analysed critically, they can be applied to other situations and other domains (Boaden, 2006; Geertz, 1973). It could be said that a second case in this thesis is the review focused on professionals in a more general sense. The comparison of these two cases is intended to shed light on different approaches, as well as the extent to which leadership development is generic or contextual. This study's design provides a deeper level of analysis than a single methodology can offer and its generalisability is thought to be enhanced by the face validity of professionals, the large sample, the nature of the methodology, and the comparison of cases from multiple perspectives.

1.5 Thesis Structure

This thesis begins with a description of the background to the design and the literature review, followed by the raw findings. It then proceeds to the analysis, conclusions, conclusions explored, theoretical model prototype, and implications for practice and research. Finally, the discussion is presented and concludes with future possibilities.

Chapter two describes the methodology of the original literature review (MULTI), the resources that guided the design of the HEE review, and the instrument used to critique the HEE included studies, called the Medical Education Research Study Quality Instrument (MERSQI). Next, the analysis and findings of the extant reviews of leadership development for doctors (EMD), which served as the thesis's literature review, are outlined, followed by an identical critique of MULTI. This analysis provided the initial clarification in terms of the existing knowledge base and informed the design of the HEE review based on the conclusions regarding effective approaches to conducting systematic reviews. Finally, chapter two details the definitions used for this study.

Chapter three includes the study's underlying philosophical framework, including a description of an alternative to either of the two extreme ontological and epistemological positions.

Chapter four begins with the justification for the sample choice of professionals, followed by a treatment of the unique features of medical leadership. The last section of this chapter is a description of the methodology of the HEE systematic literature review, which forms the heart of this study.

Chapter five begins with the application of the assessment instrument (MERSQI) to the HEE-included studies, followed by an explanation and description of the tiered calibre groupings for the included studies, which steered this thesis's analysis and conclusions. The raw data of the three SLRs are presented section-by-section alongside the findings from the studies that qualified as good and moderate calibre in the HEE review. The last section of the chapter outlines the statistical analysis applied to the HEE included studies.

Chapter six identifies the conclusions from the best available evidence, based on the most credible HEE included studies. The second section of the chapter discusses the two conclusions explored that surfaced based on the analysis.

Finally, **chapter seven** outlines the implications for practice and research. The latter describes features of effective and ineffective research design and identifies areas of need for further investigation. Next, it presents details of one of the included studies, which utilised three helpful elements in a way that no other included study in the three reviews did. This study is described as an example of how to successfully implement key research elements, demonstrating how these can be applied to future studies. The third section of the chapter is a critique of the evaluation instrument (MERSQI) with suggestions for revisions to optimise its usefulness. The discussion follows, postulating why the leadership development evidence base is so thin and offering ideas for how to improve the situation. The discussion also summarises the answers to each of the research sub-questions sequentially and identifies the limitations and strengths of the study.

1.6 Chapter Conclusion

To summarise, organisations and individuals are investing enormous amounts of time and money in leadership development despite substantial gaps in the research. These gaps include which developmental goals, programme components, and measurements are optimal for producing desired results, how such programmes translate into outcomes at various levels, and how best to measure leadership and the effects of programmes. Furthermore, the evidence supporting answers to the aforementioned questions is often regarded with a good deal of suspicion and confusion. In addition, the question of the extent to which leadership development translates across contexts has not yet been adequately answered. While it is generally believed that such programmes are effective and/or provide latent benefits, there is widespread interest in justifying investment in leadership development by ensuring that programmes are empirically-informed and produce measured performance results. Although resources are available, many leaders, whether or not in formal leadership positions, lack the preparation necessary to succeed, given the complexity of organisations and industries today. Some of the evidence is conflicting and there are many reports of programmes that failed, which suggests that although most interventions are evaluated positively, it is indeed possible to get it wrong. For many reasons, this is a situation that can no longer be afforded (Rowland, 2016). Stakeholders want to know what works optimally and what evidence there is to substantiate those claims. The systematic evidence analysis methodology is intended to provide a clear and transparent treatment of the phenomenon by highlighting what is known and on what evidence it is based. Focusing on professionals is a useful sample case to analyse leadership development, since its external validity and translatability to other contexts is thought to be high.

2 Chapter Two: Literature Review

This chapter outlines the two literature reviews and other background steps that informed the design of the HEE systematic literature review (SLR). As an inversion of a traditional PhD structure, the overall PhD methodology is described in this chapter and the HEE SLR methodology is outlined in chapter four, since the latter is the culmination of the preceding steps and forms the heart of this study and its conclusions. As described in the introduction, this thesis's methodology, called a systematic evidence analysis, centres primarily on one non-systematic and two SLRs. This chapter begins with a description of the methodology of the original PhD SLR focused on leadership development for professionals in multiple domains (MULTI). The articles that served as resources to guide the design of the HEE SLR are explained, followed by the instrument that was applied to critique the credibility of the HEE SLR included studies. The chapter then details the review of extant literature reviews on physician leadership development (EMD). This section includes the analysis and critique of those reviews, highlighting their strengths and shortcomings. The next section presents the application of the same analysis to the MULTI SLR. The final preliminary stage was to combine the previous steps to pinpoint the key elements of optimal approaches to conducting systematic reviews on leadership development, which informed the choice of design for the HEE SLR. As will be described near the end of the chapter, the findings and conclusions from MULTI and the EMD SLR were compared with those of the HEE SLR to form a robust analysis and presentation of the conclusions of the best available current literature. The findings of MULTI, EMD, and HEE were then combined to produce the final conclusions, conclusions explored, prototype theoretical model, discussion, and implications for research and practice.

2.1 Original SLR (MULTI) Methodology

The first stage in the PhD methodology was conducting the original PhD SLR (MULTI), as depicted in Figure 2.2. The research protocol was devised with guidance from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Liberati et al., 2009) and the Cook and West strategy for conducting systematic reviews in medical education (2012). The search of scholarly literature for MULTI was guided by two specialist librarians from the University of Cambridge: one from the Faculty of Education and the other from the Faculty of Medicine. It was limited to articles published in English in peer-reviewed academic journals in the period from 2005 to 2015 using four electronic databases: Business Source Complete, ABI, ERIC, and Pubmed/Medline. The keywords used in all the searches were: "lead*" AND ("educat*" OR "develop*" OR "teach*" OR "taught" OR "train*"), each allowing for variations (eg “educating”). The results of the search process are summarised in Figure 2.1 below. The initial search produced 9,745 citations. Before continuing, it should be mentioned that the initial sample of nearly ten thousand articles was expected. The main search term, “lead*,” as a homograph, attracts a profusion of hits based on unrelated topics, such as “lead poisoning,” and these based on the term’s colloquial use, for example, “*leading* research in agriculture.” Given this large sample, details of the reasons for excluding each of the more than nine thousand articles were not recorded.

Since multiple databases were used, there was some overlap and duplicate articles were removed. Studies which focused on secondary school, undergraduate, military officer cadets, or medical students were excluded on the basis of not being directly relevant to the current study’s focus on adult professionals. Articles from the fields of primary or secondary education and nursing were also excluded due to the extensive amount of literature available in these areas that is also not specifically relevant to leadership development for professionals. This is similar to the exclusion criteria that Frich et al. (2014) employed in their literature review. Relevance of the publications was then assessed based on titles and abstracts. Given the inclusion criteria listed below, more than 1,300 articles were consulted beyond the abstract, which is described as “full paper assessment” in Figure 2.1 below. This was often necessary to determine whether physicians were included in the sample, since the samples are commonly described as healthcare professionals or a variation, or whether the programme was evaluated. After reviewing the bibliographies of the relevant articles, studies not identified in the initial search which met the inclusion criteria were added. Finally, empirical studies were separated from non-empirical articles, with the latter consulted as useful background information.

Specifically, studies between 2005 and 2015 were analysed as part of the MULTI review of empirical studies, provided that, in addition to the above, their:

- Designs focused on leadership development interventions, programmes, or individual developmental activities used for leadership development (eg coaching)
- Designs involved some form of evaluating the effectiveness of the programme/intervention, rather than simply presenting a model or theory or a description of a pilot programme that had not yet been evaluated
- Sample groups were adults
- Study focus was not on one individual capability, such as the paper by Mumford et al. (2007), which studied creativity in leadership.

The final sample of 56 empirical studies formed the nucleus of the MULTI literature review and each was analysed extensively. The details of each study were recorded using structured data entry according to the following codes: author name, publication year, whether or not they tested hypotheses, published research questions, data collected (quantitative, qualitative, or both), methodology, methods and their details, sample size, control group (if applicable), gender split, mean age, level of seniority and role of the participants (eg senior managers), domain (eg healthcare), selection criteria for the programme (eg nominated by supervisor), programme location, number of sites, name, and goals, in-house or external, the length and structure of the programme (eg six months with one day-long session every month), topics addressed, developmental activities or sources of learning involved (eg coaching), raters, type of data collected from among subjective numbers (eg self-ratings), subjective descriptions of benefits, and objective statistics on the effects of performance, and when data was collected (eg six months after the programme), and outcome measures and reported benefits (eg Post-Programme Evaluations, promotions) according to the Kirkpatrick model to be described below. These categories and codes are depicted in Table 2.1 below.

Table 2.1

Coding for the MULTI SLR

Categories	Codes
Study details	Author name
	Publication year
	Whether they tested hypotheses
	Research questions
	Data collected: quantitative, qualitative, or both
	Methodology
	Methods and their details
Sample	Size
	Control group size (if applicable)
	Gender split percentage
	Mean age
	Level of seniority and role (e.g. senior managers)
	Domain (e.g. healthcare)
Programme	Selection criteria (e.g. nominated, applications)
	Location
	Number of sites
	Name of the programme
	Programme goals
	In-house or external
	Length and structure (e.g. six months with one day-long session every month)
	Topics addressed
	Developmental activities (e.g. coaching)
	Faculty: internal, external, or mixed
Measurements	Raters (self, supervisor, peer, subordinates, facilitator, statistics)
	Type of data collected (subjective descriptions, self-reported numbers, and objective statistics)
	When data was collected (pre, baseline, during, post, post-post)
	Outcome measures and reported benefits (e.g. Post-Programme Evaluations, promotions received following the programme)

The list of codes presented above were applied to each of the 56 included studies.

Many of the more than 300 non-empirical articles were consulted as background information for the introduction and major theme. For example, relevant meta-analyses and literature reviews published in academic journals were carefully reviewed, including those by Burke and Day (1986), Collins and Holton (2004), Straus, Soobiah, and Levinson (2013), and

Frich et al. (2014). These were not analysed in the same way as the empirical studies at this point.

Figure 2.1 below illustrates the process of the MULTI literature search, which progressed from a predictably large initial sample to the two categories of relevant sources: the 56 empirical studies and the more than 300 non-empirical studies consulted for background information.

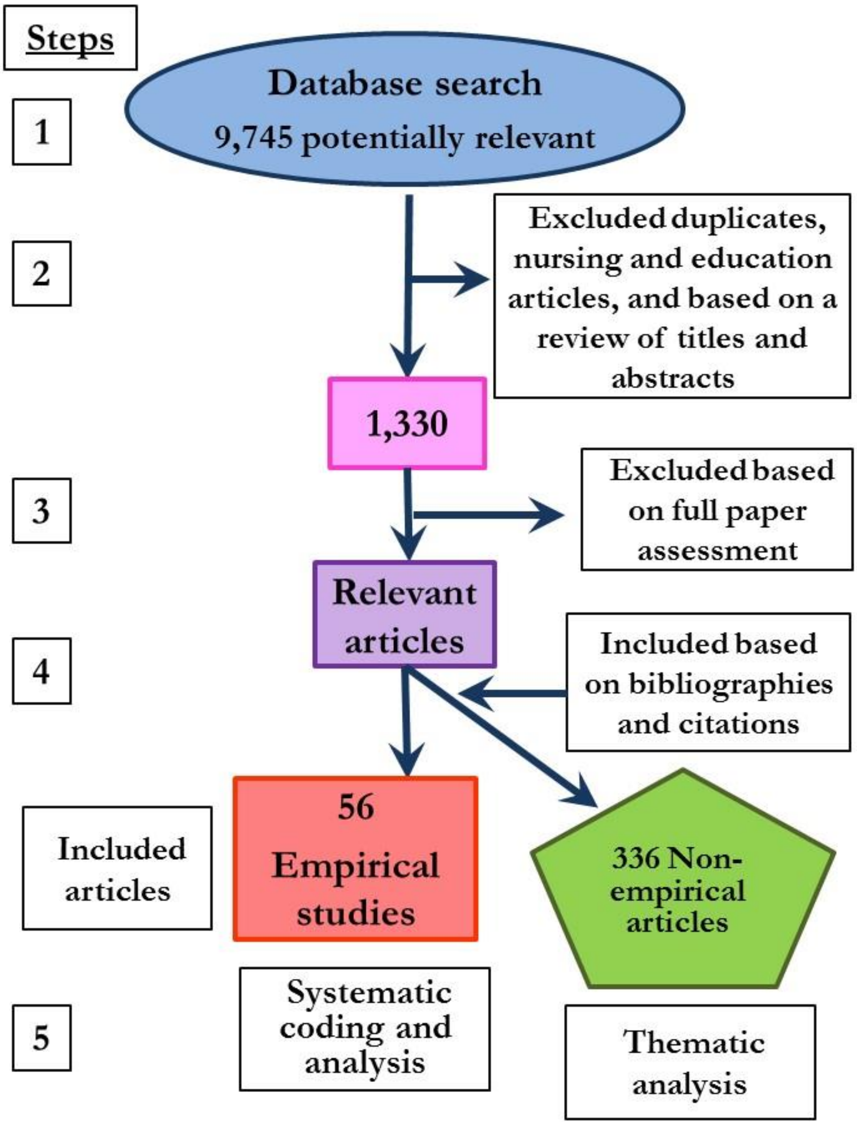

Steps
Database search
9,745 potentially relevant
1
2
Excluded duplicates, nursing and education articles, and based on a review of titles and abstracts
1,330
3
Excluded based on full paper assessment
Relevant articles
4
Included based on bibliographies and citations
Included articles
56
Empirical studies
336 Non-empirical articles
5
Systematic coding and analysis
Thematic analysis

Figure 2.1 Original SLR (MULTI) Literature Search

The process of the MULTI literature search is depicted above, progressing from a predictably large initial sample to the two categories of relevant sources: the 56 empirical studies and the more than 300 non-empirical studies consulted for background information.

As touched on previously, the choice was made before the analysis to incorporate Kirkpatrick and Kirkpatrick's (2006) four-part model categorising the reported outcomes of training evaluation, following the example of Frich et al. (2014) and Straus et al.'s (2013) systematic reviews, and other studies. The original model was adapted for this study by adding an objective measure for Level 3 (3b), along with Frich et al.'s (2014) separation of Levels 4a and 4b outcomes. In this model:

- **Level 1** refers to participants' satisfaction with the programme, most commonly in the form of Post-Programme Evaluations (PPEs).
- **Level 2a** involves changes in participants' attitudes or perceptions, such as increased engagement and aspirations to lead.
- **Level 2b** groups the changes in participants' knowledge and skills together, which tend to be reported using those terms.
- **Level 3a** denotes self-reported changes in participants' behaviour.
- **Level 3b** refers to objective changes in participants' behaviour. This can involve outcomes such as promotions or improved Multi-Source Feedback (MSF) results (pre and post).
- **Level 4a** refers to organisational impact, such as developing or implementing a new programme (subjective and objective).
- **Level 4b** refers to benefits to patients (in the case of healthcare) or clients (subjective and objective), such as a decrease in patient mortality.

This model is summarised in Table 2.2 below.

Table 2.2

A Modified Version of Kirkpatrick's (2006) Training Evaluation Model

Level	Details
1	Participant satisfaction with the programme/intervention, useful mainly for quality control
2a	Changes in participants' attitudes or perceptions
2b	Changes in participants' knowledge and skills
3a	Changes in participants' behaviour (subjective)
3b	Changes in participants' behaviour (objective)
4a	Organisational change (subjective and objective)
4b	Benefits to clients or patients (subjective and objective)

Above is a depiction of Kirkpatrick and Kirkpatrick's (2006) model that categorises post-programme outcomes at the individual, organisational, and clinical/benefit to clients levels.

As mentioned previously, the first stage in the overall PhD methodology after the background reading was the original literature review, as shown below in Figure 2.2.

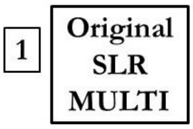


Figure 2.2 PhD Methodology, Stage 1.

As mentioned previously, the findings of the MULTI SLR is presented in chapter five of this thesis.

2.2 Systematic Literature Review Resources and Guides

The next step in preparing to undertake the HEE review was devising a research protocol. A committee of researchers and healthcare professionals, led by the author of this thesis, was formed to strengthen the quality of the protocol. The committee members' professions are listed in the appendix on page 348. The committee agreed that a systematic literature review was the appropriate approach to answer the research question by collating the best available evidence in the academic literature. This process's transparent, reproducible, and scientific nature is said to minimise bias and strengthen the credibility of a review's findings and conclusions (S. Green et al., 2011; Husebø & Akerjordet, 2016; Liberati et al., 2009). Hartley and Hinksman (2003) assert that systematic reviews are considered more rigorous that typical social sciences literature reviews, which are often based on narrative and subjective judgments. They affirm that systematic reviews are the kind of research that is needed in the field. The Kirkpatrick model was again selected to categorise the reported outcomes.

The next stage of preparation for the HEE SLR was to consult guides on conducting systematic literature reviews and meta-analyses, which was done with the MULTI review, as depicted in Figure 2.3 below. Three resources were selected: the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Liberati et al., 2009) and the *Cochrane Review Handbook for Systematic Reviews of Interventions* ('Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0', 2011) served as the main sources of guidance, supported by the Cook and West (2012) strategy for conducting systematic reviews in medical education.

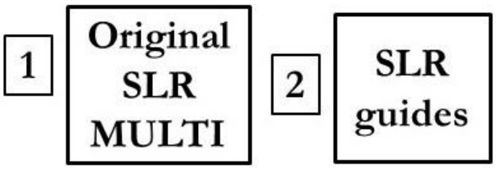


Figure 2.3 PhD Methodology, Stage 2.

Consulting documents of this nature is important, since the rigour and reliability of reviews are largely based on their protocol and methodology (Shamseer et al., 2015). PRISMA is evidence-based and provides step-by-step guidance on high calibre, transparent reporting of systematic reviews, which can also be used to direct the design of reviews. It is useful as the basis for many types of research, but particularly the evaluation of interventions (PRISMA, 2015), which made it suitable for a review of leadership development programmes. PRISMA is also widely endorsed by author's guidelines of academic journals (Tao et al., 2011), as well as by the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network's reporting guidelines (UK EQUATOR Centre). Similarly, the Cochrane Handbook focuses on the effects of interventions and offers empirical evidence-based direction on making methodological decisions that are systematic, informed, and explicit ('Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0', 2011). Finally, Cook and West (2012) provide a concise and practical guide to the conduct and reporting of systematic reviews, particularly for medical education research, which was used as an additional resource.

Alternatives and additional resources were also considered, such as the *Systematic reviews and meta-analyses: a step-by-step guide* by the Centre for Cognitive Ageing and Cognitive Epidemiology research group (2013). Use of this source however was deemed unnecessary, since the authors declare that much of the guidance in their document derived from the "excellent and extensive" Cochrane Review Handbook. The three sources listed above were therefore selected for their comprehensive nature, widespread credibility, and specific applicability to educational interventions.

2.3 Included Study Credibility Critique Instrument

As was mentioned in chapter one, to assess the current state of the literature and isolate and clarify transparently which is the best available evidence, it was decided that this study would require an instrument to critique the calibre of the HEE SLR included studies. It has already been mentioned that many stakeholders view the quality of medical education research as inadequate (Husebø & Akerjordet, 2016; Reed et al., 2007). One issue cited is that there are common deficiencies in study design and poor reporting of study details that impair readers' ability to learn from articles' conclusions (Straus et al., 2013). To address this problem, Reed et al. (2007) developed an instrument to measure the methodological quality of education research studies called the Medical Education Research Study Quality Instrument (MERSQI). MERSQI includes ten items pertaining to six domains of study quality: design, sampling, type of data (subjective or objective), validity, data analysis, and outcomes (see Table 2.4 and Table 2.5 below). Each of these aspects is scored on an ordinal scale and the points are summed to produce a total score. Each domain has a maximum score of three and the maximum overall score is 18, with a minimum of 4.5. Like this study's reviews, Reed et al. also used Kirkpatrick's hierarchy of outcomes and attributed the highest score to outcomes that benefit patients (Level 4b). The authors tested the instrument's validity and reliability extensively and thus, it was chosen to enhance the transparency and credibility of the HEE review and its conclusions.

To explain, a research instrument is said to be validated when it has been tested for reliability and validity relevant to the population to be studied (Dowrick, Wootten, Murphy, & Costello, 2015). The former determines the internal consistency, that is whether random error is minimal, and that an instrument produces stable results and has high reproducibility (Dowrick et al., 2015). Validity refers to whether an instrument accurately measures what it intends to measure. Face validity is whether the instrument appears to be accomplishing this. Content validity refers to key stakeholders confirm that the instrument investigates the most important aspects of the phenomenon in question. Construct validity considers the relationships between the instrument and theoretical concepts or constructs, including variables that are not directly observable such as pain or anxiety.

Several alternatives to MERSQI were considered, including the Newcastle-Ottawa Scale (NOS) (Wells et al., 2016), but because the NOS is not specific to medical education interventions, it was felt that MERSQI is more specific and appropriate. As well, unlike the MERSQI, the NOS does not include numerical score components, which serve to increase transparency and minimise bias. Another option that was considered was the Critical Appraisal Skills Programme (CASP), an independent organisation that offers a variety of checklists, including the Qualitative Research Checklist (2013). These tools were rejected in favour of having one instrument that could be applied to quantitative, qualitative, and mixed methods studies. Finally, the Cochrane Assessment Tool for Nonrandomised Studies of Interventions (Sterne, Higgins, & Reeves, 2014), which analyses risk of bias was also considered. Although that instrument has excellent points that informed the analysis of the HEE review at various stages, it is overly detailed in some regards and not specific enough in others to form the basis of the HEE analysis. Husebø and Akerjordet used this tool in their review and concluded that all studies were exposed to a high risk of bias (2016), a conclusion shared by Straus et al. of the studies in their review (2013). Focusing exclusively on risk of bias to measure studies' credibility overlooks other important considerations, ones that MERSQI addresses. It was also felt that the analysis in the HEE SLR included many steps that could identify and explain bias in the included studies while also considering the aspects that Straus et al.'s approach, with its purely quantitative focus, leaves out. Thus, MERSQI was selected as the best instrument to evaluate the credibility of the HEE SLR included studies based on its specific nature and numerical score components.

For the readers' convenience, a colour coding system was developed, as outlined in Table 2.3 below, to illustrate the results of the analysis. Green always denotes the most credible and red the least credible result, including if relevant variables were not reported. If there were three possibilities, yellow denotes the middle level of credibility and if there were four possibilities, purple denotes a level of credibility that is higher than yellow but less than green. Thus, in a four-option scenario, level of credibility in order of highest to lowest would be represented by green, purple, yellow, and red.

This system is used consistently throughout the remainder of the presentation of the HEE review.

Table 2.3

Colour Coding for MERSQI

Two possibilities		Three possibilities		Four possibilities	
Highest	Green	Highest	Green	Highest	Green
Lowest	Red	Middle	Yellow	Second best	Purple
		Lowest	Red	Third best	Yellow
				Lowest	Red

The colour coding system above, which was applied to the HEE included studies, was also applied to the EMD reviews, which will be described further on in this chapter.

Table 2.4 and Table 2.5 below present the MERSQI instrument, which was applied to each study in the HEE SLR, with the colour coding system added.

Table 2.4

The MERSQI Instrument with Colour Coding Applied (1/2)

Domain	Item	Item score	Maximum domain score
1. Study design	Single group cross-sectional or single group post-test only	1	3
	Single group pre and post-test	1.5	
	Non-randomised, two-group	2	
	Randomised controlled experiment	3	
Sampling			
2. Institutions	One	0.5	3
	Two	1	
	>Two	1.5	
3. Response rate	<50% or Not reported	0.5	
	50 – 74%	1	
	≥75%	1.5	
4. Type of data	Assessment by study subject	1	3
	Objective measurement	3	
Validity of evaluation instruments' scores			
5. Internal structure	Not reported	0	3
	Reported	1	
6. Content	Not reported	0	
	Reported	1	
7. Relationships to other variables	Not reported	0	
	Reported	1	

Table 2.5

The MERSQI Instrument with Colour Coding Applied (2/2)

Data analysis			
8. Appropriateness	Data analysis inappropriate for study design or type of data	🟥 0	3
	Data analysis appropriate for study design or type of data	🟩 1	
9. Sophistication	Descriptive analysis only	🟥 1	
	Beyond descriptive analysis	🟩 2	
10. Outcomes	Satisfaction, attitudes, perceptions, opinions, general facts (Level 1 and 2a)	🟥 1	3
	Knowledge, skills (Level 2b)	🟨 1.5	
	Behaviours (Level 3a and 3b)	🟪 2	
	Patient/healthcare outcome (Level 4b)	🟩 3	
Total			18

Note. Green = most credible. Purple = second most credible. Yellow = third most credible. Red = least credible result, including if relevant variables were not reported.

Above is the colour coding system as applied to the MERSQI instrument.

2.4 Review of the Extant Reviews (EMD)

As mentioned previously, another step taken prior to the HEE SLR was a review of existing literature reviews on leadership development for doctors, called the extant medical doctors SLR or EMD SLR. (To reiterate, "extant" in this sense is used only to distinguish this collection from the HEE SLR). This endeavour served two purposes: first, critiquing the strengths and weaknesses of those reviews helped inform the design of the HEE, as advised by Cook and West (2012). This step allows authors to benefit from the work of others, while ensuring that they are filling a meaningful gap in published reviews and adding significantly to the current knowledge (D. A. Cook & West, 2012). The second purpose of this stage was to be able to relate the findings of the HEE SLR to other evidence, as recommended in the PRISMA guidelines (Liberati et al., 2009).

To be included, the reviews had to focus on leadership development interventions for doctors, but their study samples did not need to be physician-exclusive. Likewise, they were included if their reviews featured interventions with leadership as a primary focus, but they did not have to be restricted to these. For example, in the Steinert et al. review (2012), only 14 of

48 studies they identified had leadership as a primary focus. In this case, only the leadership-focused interventions were compared to the HEE SLR findings and conclusions. Six systematic reviews and one non-systematic literature review were identified. The reviews were coded for the same items that were used to analyse the HEE SLR. Where the findings and conclusions of the EMD SLRs agreed with those of the HEE SLR, citations were added, and when there were important contrasts, these were highlighted and discussed as well. In addition to this level of analysis, the EMD SLRs were coded for the following variables, as depicted in Table 2.6 below: number of researchers, number of databases included in the literature search, year range for included studies, whether or not they included only journals in English, whether they restricted their search to peer-reviewed articles only, target population, number of included articles, and to what extent they reported the full sample details. The reviews' critiques of the developmental activities, programme components, and programme and participant evaluation were noted, as was their absence, when applicable. The analysis of the reviews included which, if any, assessment instrument the researchers used to critique the studies, whether they rated and ranked the credibility of the included studies, and if they included in the publication a chart that presented those ratings and rankings. In terms of outcomes and analysis, it was considered whether the reviews used the Kirkpatrick model or another, whether they critiqued the included studies' reported outcomes or took them at face value, and if they analysed the correlations among variables. It was investigated whether the SLRs report only positive outcomes, possibly indicating selective reporting bias, or if they also describe negative or nuanced reports. Finally, the reviews were examined to determine if they present tiered conclusions according to the credibility of the included studies. As has been mentioned previously, this question is one key to filling the knowledge gaps in the field with credible evidence.

Table 2.6

Coding for the EMD SLRs

Categories	Codes
Reseachers	Number of researchers
Literature Search	Number of databases
	Year range for included studies
	English-only articles?
	Peer-reviewed articles only?
	Target population
	Number of included articles
Reporting	Are all sample details provided?
Critiques	Developmental activities?
	Programme components?
	Programme and participant evaluation?
Study Credibility Critique and Rank	Credibility assessement instrument
	Rate/rank the credibility of studies
	Chart depicting the ratings and rankings?
Outcomes and Analysis	Use of Kirkpatrick outcome levels?
	Critiques of reported outcomes
	Analysed correlations among variables?
Selective Reporting Bias	Are only positive outcomes reported?
	Descriptions of negative/nuanced reports?
Conclusions	Tiered conclusions?

As seen above, the codes were applied as part of the analysis to the eight extant reviews (seven EMD and MULTI).

2.5 Results of the Critique of the Extant Reviews

Table 2.7, Table 2.8, and Table 2.9 below outline the key components of the analysis of each review in the EMD SLR, as well as the original MULTI SLR and the HEE SLR, according to the same variables. The details of the HEE SLR are described in chapter four. The six systematic reviews identified in the EMD SLR are:

- **Hartley and Hinksman** (2003)'s report for the NHS Leadership Centre that targeted medical leadership development from the period of 1997 – 2003. The authors did not specify whether their sample was physician-only or not. The number of included studies is unclear and the authors did not report using an assessment instrument to critique the studies' calibre.
- **Steinert, Naismth, and Mann** (2012) focused on physicians from the period of 1980 – 2009. They identified 48 studies, although only 14 featured leadership

as a primary focus of the interventions, and they also did not report using an assessment instrument.

- **Straus, Soobiah, and Levinson** (2013) targeted physicians in Academic Medical Centres (AMC's) in studies between 1948 and 2011. They identified ten studies and assessed the calibre of studies using the Newcastle-Ottawa Scale (NOS) and the Critical Appraisal Skills Program (CASP) worksheet for qualitative articles.
- **Frich et al.** (2014) focused on physicians from the period of 1950 – 2014. They identified 45 studies and did not report using an assessment instrument.
- **Rosenman et al.** (2014) focused on interdisciplinary Health Care Action (HCA) teams from the period of 1990 – 2012. They identified 45 studies, with only ten per cent of the included studies featuring leadership as a primary focus and only two studies (four per cent) assessing leadership behaviors as a primary outcome. The reviewers applied MERSQI and Cook and Beckman's elements of validity to critique the included studies.
- **Husebø and Akerjordet** (2016) targeted multi-professionals in acute hospital settings from the period of 2000 – 2009. They identified 12 studies and used the Cochrane Collaboration's tool for assessing risk of bias for quantitative studies to critique them.
- **McCauley** (2008) undertook a non-systematic literature review of leader development. The total number of studies is therefore unclear and she did not report using any assessment instruments.

A total of one non-systematic and six systematic literature reviews on leadership development for physicians were therefore located that both informed and were compared to the HEE review.

2.5.1 Analysis of the Extant Reviews

This section describes the analysis of the EMD SLRs and the original MULTI SLR. All but Frich et al. used multiple databases, which is helpful given the multi-disciplinary nature of the field of leadership development. All but Husebø and Akerjordet and MULTI employed multiple researchers working independently, an approach which strengthens the credibility of the findings by minimising bias (Liberati et al., 2009). The Husebø review employed one researcher for some sections and two for others. Every review restricted their search to English language articles and three of six plus MULTI included only peer-reviewed journals to isolate the best available evidence. Two studies, Steinert et al. and Straus, Soobiah, and Levinson, did not declare whether or not they included only peer-reviewed studies. The only review that extended to grey literature was Hartley et al., but as mentioned above, it is unclear how many studies that review included for analysis. Surprisingly, only Steinert et al. and MULTI included all the typical sample details. Hartley et al., Frich et al., and Husebø and Akerjordet left out the sample details altogether and Straus, Soobiah, and Levinson and Rosenman et al. omitted key details including age range and mean age. Not one study systematically critiqued the developmental activities or programme components such as length and location and five of the six did not critique them at all, while MULTI examined the relationship among these variables in a descriptive rather than a statistical analytical way. Rather, five of the EMD reviews simply describe these key elements along with references to studies, offering no critical comments on what is known about each and based on what evidence. In terms of forms of programme evaluation, Steinert et al., Frich et al, and MULTI critiqued them fully, three others did so only superficially, and Hartley and Hinksman did not critique them at all. Even though the EMD reviewing authors frequently commented on poor reporting in their included studies, they themselves often left out key information and neglected to analyse core aspects of leadership development in detail, limiting the value of their findings and conclusions. Pervasive gaps were therefore identified in the EMD SLRs in terms of sample details and critiques of developmental activities, programme components, and evaluation, and these issues were addressed in the design and presentation of the findings of HEE.

On deeper analysis, the shortcomings of the relevant available literature reviews are equally apparent. Four of the six reviewed plus MULTI neglected to rate and rank the credibility of the studies. Steinert et al. (2012) used a highly subjective assessment method and did not present any details of the scores for each article, merely stating that they applied an assessment tool to the included studies. The same is true of the MULTI analysis, as described earlier. Rosenman et al. (2014) used MERSQI but only provided the final score for each study, not the rating for each aspect within the study. Husebø and Akerjordet provided their complete risk of bias chart with ratings for each study, demonstrating the level of transparency needed, although, as discussed earlier, their instrument is of limited use for leadership development. These omissions, along with the absence of a rating instrument altogether in four of the six reviews, seriously undermine the transparency and credibility of the reviews and deprive the readers of a sense of how credible each study's findings and conclusions are (Liberati et al., 2009). It is for this reason that, as will be explained in chapter four, the HEE SLR provides as much information on the analysis as possible, allowing readers to judge the quality of each study for themselves (Liberati et al., 2009). This practice is in line with the PRISMA guidelines, the Cochrane Review Handbook ('Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0', 2011), and the Cook and West (2012) guide.

Only three EMD reviews and MULTI used the Kirkpatrick outcome levels (Frich et al., 2014; Rosenman et al., 2014; Steinert et al., 2012), and one of those three, Rosenman et al., did not separate subjective and objective outcomes. This is an important distinction and indeed, as will be explained in chapter five, even more such distinctions are needed. It is surprising that in the Straus et al. review, not only are organisational level outcomes (Level 4a) and benefit to patients outcomes (Level 4b) not mentioned, their instrument did not even allow for them. Steinert et al.'s review similarly neglected the important Level 4b outcomes. A major flaw in all the reviews is a failure to critique their included studies' reported outcomes; the researchers often simply describe the outcomes without offering any sense of the calibre of evidence reinforcing them. Three of the reviews (Hartley & Hinksman, 2003; Rosenman et al., 2014; Steinert et al., 2012) provide only highlights of reported outcomes with no indication of the credibility that is attached to each claim. Although Rosenman et al. offer a MERSQI score for each study, this rating is not factored into the description of the findings. The Straus, Soobiah, and Levinson and the Frich et al. reviews both provide details of the evidence to reinforce each of the studies' reported outcomes, but do not critique that evidence per se. Only Husebø and Akerjordet and MULTI carefully critiqued certain studies' reported outcomes. Therefore, the failure of the EMD SLRs and MULTI to provide clear and transparent rankings of their studies' credibility and apply them to critique the studies' reported outcomes in order to separate better evidence from weaker is a concern that informed the design and presentation of the findings of HEE.

Interestingly, only two EMD SLR reviews and MULTI reported negative or nuanced programme outcomes, information that is very useful to better understand the complex phenomenon of leadership development (Liberati et al., 2009; 'Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0', 2011). Two reviews reported only positive outcomes, which is particularly unusual in the case of the Frich et al. review; with a sample size of 45 studies the absence of negative outcomes leads one to question whether this aspect of their analysis was overlooked. Not one review attempted to draw correlations among variables, such as between the length of the programme and the level of seniority of the participants. Given the importance of how various aspects of leadership development research and practice relate to each other, the HEE SLR included this analysis in Stage 5, as will be described in chapter four. Lastly, none of the reviews offered conclusions tiered according to the credibility of the included studies in order to isolate what is known and based on what evidence. For the most part, the review authors simply presented superficial raw data or general syntheses and then made overarching observations to formulate their conclusions. To address these issues, as one of its major defining features, the HEE SLR involves structured analysis and conclusions based on the credibility of the studies.

An analysis of the extant SLRs and, to some extent MULTI, unveiled multiple pervasive omissions that significantly compromised the transparency, credibility, and usefulness of the reviews' conclusions.

It is slightly ironic that many of the omissions in the extant SLRs of which the reviewing authors are guilty are things that they themselves criticise in their included studies. This situation reflects an observation in the PRISMA guidelines: key information is often poorly reported in systematic reviews, diminishing their potential usefulness (Liberati et al., 2009). The critiques outlined in this chapter helped inform the design of this study, one of which was intended to provide the kind of transparent and credible analysis and conclusions that are needed in the field (Liberati et al., 2009).

The tables below outline the analysis of the EMD SLRs and MULTI. In these tables, green denotes aspects of the review design that enhanced the credibility of the analysis, findings, and conclusions and minimised the risk of bias; yellow denotes aspects that somewhat lessened the credibility and increased the risk of bias; and red denotes aspects that diminished the credibility and increased the risk of bias.

Table 2.7

Critiques of the EMD SLRs, MULTI, and the HEE SLR (1/3)

#	Author and Publication Year	# Researchers	# Databases	Year Range	English Only?	Peer Reviewed Only?	Target Population	Included Articles
1	Hartley (2003)	**2**	**6**	1997 - 2003	Yes	No	Medical leadership development	**Unclear**
2	Steinert (2012)	**3**	**6**	1980 - 2009	Yes	N/A	Physicians	48
3	Straus (2013)	**2**	**4**	1948 - 2011	Yes	N/A	Physicians in Academic Medical Centres (AMC's)	10
4	Frich (2014)	**2**	**1**	1950 - 2014	Yes	Yes	Physicians	45
5	Rosenman (2014)	**2**	**6**	1990 - 2012	Yes	Yes	Interdisciplinary Health Care Action (HCA) teams	45
6	Husebø (2016)	**1 for parts and 2 for others**	**7**	2000 - 2009	Yes	Yes	Multi professionals in acute hospital settings	12
Non-Systematic Review								
7	McCauley (2008)	**1**	**N/A**	N/A	N/A	N/A	N/A	**Unclear**
Original MULTI SLR								
8	Geerts (2015 - forthcoming)	**1**	**4**	2005 - 2015	Yes	Yes	Professionals in multiple domains	56
HEE SLR								
9	Geerts (2017 - forthcoming)	**2**	**7**	2007 - 2016	Yes	Yes	Physicians	25

🟩 **Enhanced credibility and minimised risk of bias**
🟨 **Somewhat lessened credibility and increased risk of bias**
🟥 **Diminished credibility and increased risk of bias**

Table 2.8

Critiques of the EMD SLRs, MULTI, and the HEE SLR (2/3)

#	Author and Publication Year	Sample Details Provided	Critiques of Developmental Activities?	Critiques of Programme Components?	Critiques of Programme and Participant Evaluation?	Use of Kirkpatrick Outcome Levels?	Credibility Assessment Instrument	Rate/ Rank Credibility of Studies?	Chart of Critiques / Ratings
1	Hartley (2003)	NR	No, only descriptions are provided	No	No	No, only a description of the model is presented	N/A	No	No
2	Steinert (2012)	Yes	Non-systematically	Non-systematically	Yes	Yes	N/A	Yes, 5-point scale: 1/5 no clear conclusions can be drawn; 3/5 conclusions can probably be based on results; and 5/5 results are unequivocal	No
3	Straus (2013)	Absent: age range, mean age, level of seniority	No	No	Somewhat: scattered descriptions are mentioned	No	Newcastle-Ottawa (NOS) and the Critical Appraisal Skills Program worksheet for qualitative articles	No	No
4	Frich (2014)	NR	No	No	Yes	Yes	N/A	No	No
5	Rosenman (2014)	Absent: age range, mean age, gender	No, only the most common are listed	No, only the most common are listed	Somewhat	Yes, but no separation: subjective/ objective	MERSQI, Cook and Beckman's elements of validity	Yes	Yes, but only the final scores
6	Husebø (2016)	NR	No	No	Very basically	No	The Cochrane Collaboration's tool for assessing risk of bias for quantitative studies	No	No
Non-Systematic Review									
7	McCauley (2008)	No	Yes	No	No	No	N/A	No	No
Original MULTI SLR									
8	Geerts (2015 - forthcoming)	Yes	Yes	Yes	Yes	Yes	N/A	N/A	N/A
HEE SLR									
9	Geerts (2017 - forthcoming)	Yes	Yes	Yes	Yes	Yes	MERSQI	Yes	Yes

Colour coding:	
🟩	Enhanced credibility and minimised risk of bias
🟨	Somewhat lessened credibility and increased risk of bias
🟥	Diminished credibility and increased risk of bias

Table 2.9

Critiques of the EMD SLRs, MULTI, and the HEE SLR (3/3)

#	Author and Publication Year	Critiques of Reported Outcomes	Analysed Correlations Among Variables?	Only Positive Outcomes Reported?	Descriptions of Negative/ Nuanced Reports?	Tiered Conclusions?
1	Hartley (2003)	No	No	N/A	No	No
2	Steinert (2012)	No	No	No	Yes	No
3	Straus (2013)	Somewhat: they provide details of the evidence supporting outcomes, but do not critique them	No	No	Yes	No
4	Frich (2014)	Somewhat: their chart provides details of the evidence supporting outcomes, but does not critique them	No	Yes	Yes	No
5	Rosenman (2014)	No, not individually	No	N/A	No	No
6	Husebø (2016)	Yes	No	No	Yes	No
Non-Systematic Review						
7	McCauley (2008)	N/A	No	Yes	Yes	No
Original MULTI SLR						
8	Geerts (2015 - forthcoming)	Yes	Yes, but not statistical analysis	No	Yes	No
HEE SLR						
9	Geerts (2017 - forthcoming)	Yes	Yes	No	Yes	Yes

Colour coding:	Enhanced credibility and minimised risk of bias
	Somewhat lessened credibility and increased risk of bias
	Diminished credibility and increased risk of bias

The figures above represent the colour-coded analysis of the seven extant reviews, MULTI, and HEE.

Stage 3 of the PhD methodology involved identifying the strengths and weaknesses of the designs of EMD and MULTI before proceeding to the HEE SLR, which is depicted in Figure 2.4 below.

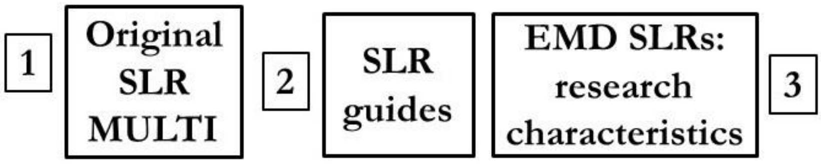


Figure 2.4 PhD Methodology, Stage 3.

2.5.2 Methodological Strengths of Extant Reviews

Several methodological strengths emerged from consultation of the resource guides and review of the EMD SLRs that informed the design of the HEE SLR. These methodologies included using multiple researchers working independently (D. A. Cook & West, 2012), searching for articles in multiple databases, and reporting all sample and participant details from the included studies. Another strength that was identified was critiquing the effectiveness of the developmental activities, programme components, and the studies' evaluation. The Kirkpatrick model was determined to be helpful for categorising the studies' reported outcomes. Using an assessment instrument to evaluate studies' credibility and rate and rank the quality of studies, and publishing a chart that outlines the details of this step were identified as measures that reviewers can take to significantly enhance the transparency and credibility of their work (Liberati et al., 2009). Likewise, in the interests of determining what is known and based on what credibility, critiquing studies' reported outcomes according to their methodologies was also identified as a priority. Analysing relationships among the variables as extensively as possible and providing negative or nuanced reports also emerged as a way to add another level to the results. Finally, one step that not one of the reviews analysed at this stage was providing tiered conclusions based on the credibility of each study. This step was identified as crucial because the seeming ubiquity in the field of people making claims without specifying the level of credibility of supporting evidence. The MULTI SLR identified a vast range among the credibility of the included studies' findings, which, given the lack of a formalised method of ranking the evidence, frustrated the conclusions, an issue that was determined to be common to other reviews as well. This finding led to the unique and defining feature of the HEE SLR: grouping the included studies by credibility into categories of good, moderate, limited, and anecdotal evidence.

A list of the methodological strengths of SLRs is presented in Table 2.10 below:

Table 2.10

Methodological Strengths of Effective SLRs

✓	Multiple researchers working independently
✓	Searching multiple databases
✓	Reporting all the sample and studies' details
✓	Critiquing the effectiveness of the developmental activities
✓	Critiquing the effectiveness of the programme components
✓	Critiquing the effectiveness of the programme evaluation
✓	Use Kirkpatrick's model of outcomes categorisation
✓	Applying an assessment instrument to evaluate the included studies' credibilit
✓	Rating and ranking the studies' quality
✓	Publishing the full results of the assessment, rating, and ranking
✓	Critiquing the studies' reported outcomes based on their methodologies
✓	Analysing the relationships among the variables
✓	Investigating negative and nuanced reports
✓	Producing tiered conclusions according to the studies' credibility

The above research measures strengthen the credibility and enhance the usefulness of the findings and conclusions of reviews.

2.6 A Unique and Defining Feature of This Study

Given the points raised above, one of the main priorities for the HEE review was to generate conclusions that were tiered according to the credibility of studies in order to meet the needs of the research and practitioner communities alike. Husebø and Akerjordet (2016), Rosenman et al. (2014), Frich et al. (2014), and others echo the need for increased scientific rigour in terms of reliability and validity in the field. The tiered conclusions feature is one of the defining features of this study and this characteristic is thought not to exist elsewhere in leadership development literature. The PRISMA guidelines mention that other systematic reviewers formally rated or assessed the overall body of evidence addressed in their reviews and described the strength of their study recommendations as related to their assessments of the quality of evidence (Liberati et al., 2009), but no such occurrence was identified in the EMD SLR. The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system is one example of this type of rating (Guyatt et al., 2008). As will be described in the findings section, there is a massive range amongst the HEE included studies' credibility, from the lowest possible MERSQI score (4.5/18) to 15/18, and the majority are of compromised credibility, despite the restriction of the literature search to peer-reviewed journals.

When articles' conclusions are not tiered in a review according to the studies' credibility, the danger is that findings reinforced by strong empirical evidence can be grouped together with reports that are entirely anecdotal. For example, if study A, featuring a randomised controlled trial, concluded X, while article B, which appeared in a peer-reviewed journal but was seriously flawed, reported conclusion Y, "what is known" in the field of leadership development until now would often suggest, without qualification, that X and Y are both true. This situation is at best confusing, certainly misleading, and potentially harmful. Steinert et al. (2012)'s review, for example, offers conclusions for leadership interventions, but gives no indication of whether each point is drawn from a single study, multiple studies, the majority of studies, or the most credible studies. An even more significant example is seen in MacLeod (2012), who describes the "Yale Goal Study." This study allegedly surveyed graduates in the Yale Class of 1953, asking if they had specific written goals for the future, and apparently found that only three per cent had such goals. A 20-year follow up subsequently reported that those three per cent accumulated more personal financial wealth than the other 97 per cent of the class combined (MacLeod, 2012). Following a review, the Yale University Library stated, "It has been determined that no "goals study" of the Class of 1953 actually occurred" (Sider, 2017). This situation gives rise to multiple challenges: not only was the original study fabricated, but its results can be mistaken for or represented as truth by referencing MacLeod's article even though the study never happened.

While there may be valuable information in anecdotal studies based on the experience of those who designed and implemented a programme, the risk of bias is high in such studies and there is little to no concrete data to reinforce it the conclusions. To use an everyday example, if one asked a mother for parenting advice, she would no doubt have valuable tips, but one does not create a national health policy based on one person's experience. In fact, not only is it unwise to base the design of a costly programme on anecdotal information, it can be argued that studies that appear in academic journals but whose claims are not substantiated by credible evidence can be harmful. These unsubstantiated conclusions, given an air of authority by inclusion in scholarly journals, could be used to justify significant decisions such as resource allocation, programme design choices, or research grant funding. Tim Judge, Associate Dean for Faculty & Research at the University of Notre Dame, in a lecture at Cambridge University, stated, “We must guard against the unwarranted influence of the leadership book publishing complex” (Judge, 2016). The same caution can be extended to all information that claims to be “evidence” of leadership development effectiveness, but is not clear and transparent about the studies’ methodological credibility. Trouble arises when authors blur the lines between “findings” (and imply that these are reinforced by solid evidence) and “reports”, which are appropriately qualified, in their own studies and when citing others’.

The purpose of the HEE SLR is to test the characteristics of good reviews by putting them into practice in as transparent and objective way possible and to contrast the findings and conclusions to those of other studies. This approach emerged in response to a clear need in the field and was refined through the background stages described in this chapter, which highlighted the importance of elucidating what is known based on what evidence. How the preliminary steps contributed to the design of the HEE SLR is outlined in Figure 2.5 below.

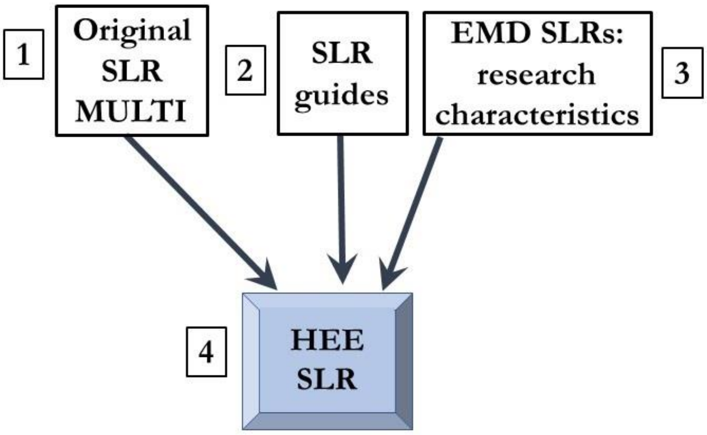


Figure 2.5 PhD Methodology, Stage 4.

The figure above depicts how the analysis of the reference guides and extant and MULTI reviews contributed to the design of the HEE methodology. This included using a validated instrument to critique the included studies’ credibility, basing the findings and conclusions on these ratings, and analysing the relationship among variables.

2.7 Outstanding Questions at This Point

At this point in the research, the outstanding questions were:

- What can be said about the calibre of the current state of the literature? This became the first research sub-question.
- What are effective ways to group the studies according to their credibility so as to tier a review's analysis and conclusions?
- What does the best evidence say makes for optimal leadership development? This became sub-question two.
- What is known about the relationships among variables of developmental activities, programme components, evaluation, and outcomes?
- To what extent are principles of optimal leadership development universal versus contextual? This became sub-question four.
- What are the implications for practice based on the answers to the previous three questions?
- What information is there in the literature regarding optimal ways of measuring leadership, particularly after development interventions. This became sub-question three.
- What are the implications for research, both in terms of conducting systematic reviews and individual leadership development studies?

These questions, which the HEE SLR intended to address, contributed to forming this thesis's research questions.

2.8 PhD Methodology: Further Stages

Having now described the background to the design of the HEE SLR, the remainder of the PhD methodology is outlined in order to give the reader a sense of the overall project and analytical development. It should be reiterated that the *content* of the literature review of extant reviews has been added to the conclusions of the best available evidence in chapter six and the conclusions explored, rather than here. Stage 5 involved combining the findings of MULTI and the EMD SLR with those of the HEE SLR, as depicted in Figure 2.6 PhD Methodology Stage 5 below.

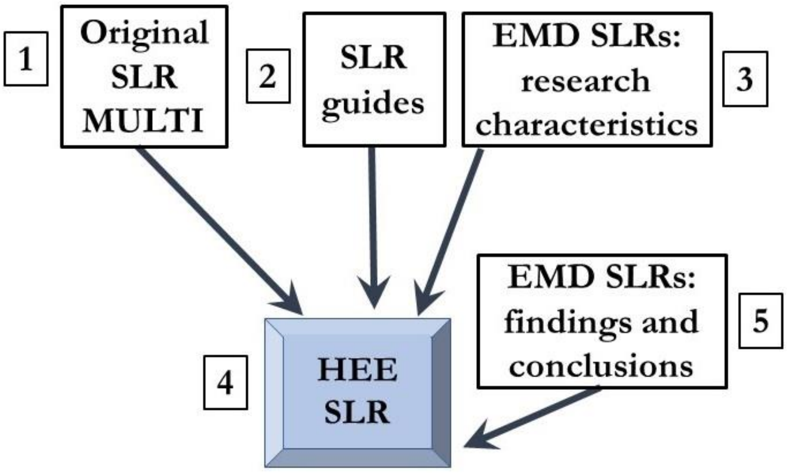

1
Original SLR MULTI
2
SLR guides
EMD SLRs: research characteristics
3
EMD SLRs: findings and conclusions
5
4
HEE SLR

Figure 2.6 PhD Methodology Stage 5

Above is a depiction of how the findings and conclusions of the HEE were compared to those of MULTI and EMD.

The next stage involved describing the conclusions of the best available evidence, which are based on the good and moderate evidence studies from the HEE SLR. As mentioned previously, when those conclusions from MULTI and EMD SLR agreed with those from the HEE SLR, citations were added, and when there were interesting nuances or noticeable contrasts, these were mentioned as well. This stage is depicted in Figure 2.7 below.

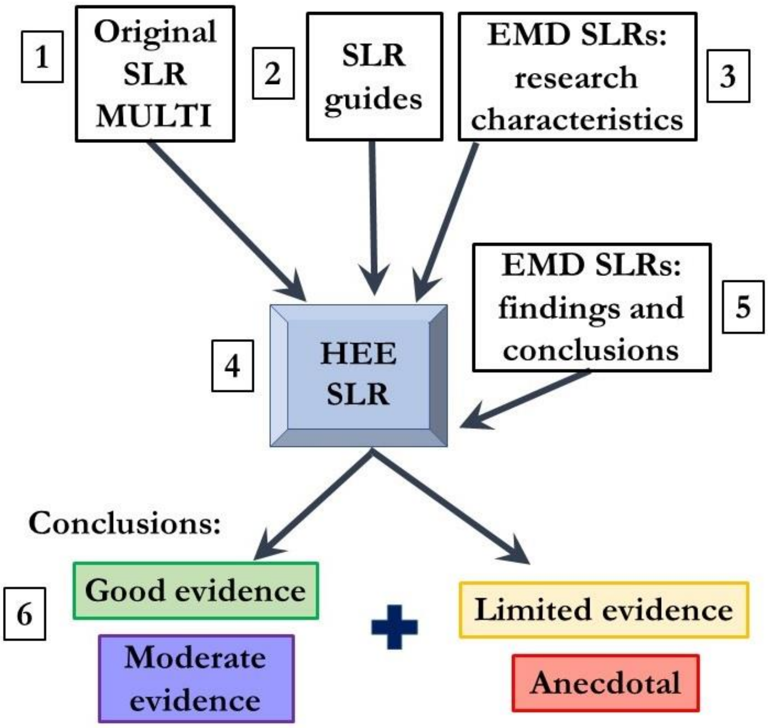


Figure 2.7 PhD Methodology, Stage 6.

A depiction of how the conclusions from the best available evidence were devised and reinforced by points from limited and anecdotal evidence HEE studies, along with MULTI and EMD, is presented above.

Stage seven involved the formation of the two conclusions explored. As mentioned previously, this stage involved selecting two conclusions from the best available evidence and investigating them in more detail by reviewing the included studies again from those perspectives. The first of these conclusions is how Knowles's (1984) principles of adult learning apply to leadership development. Several studies cited these principles, but there was no one comprehensive explanation of their application to leadership development even though it seemed that such a resource would be valuable for researchers and those designing programmes alike. As this further stage of research progressed, it became clear that two principles should be added and that Dale's (1969) educational model the Cone of Experience could also help illustrate the key functions of developmental activities.

The second conclusion explored describes a set of factors before, during, and after interventions that are said to facilitate the transfer of leadership learning, particularly following programmes. This topic was identified as crucial in discussion between the two researchers because of the number of studies that claimed to have failed for these reasons. Given the commonly recognised importance of leadership, as described in chapter one, and the cost of programmes, the researchers felt this situation must be addressed. This set of factors was compiled from the included studies based on three similar sets of factors, as well as best practice points, including from studies that claimed to have failed. The factors cover key features of programme design, delivery, and evaluation, as well as aspects of organisational culture that appear to have a significant impact on the application of leadership. This stage is depicted in Figure 2.8 below.

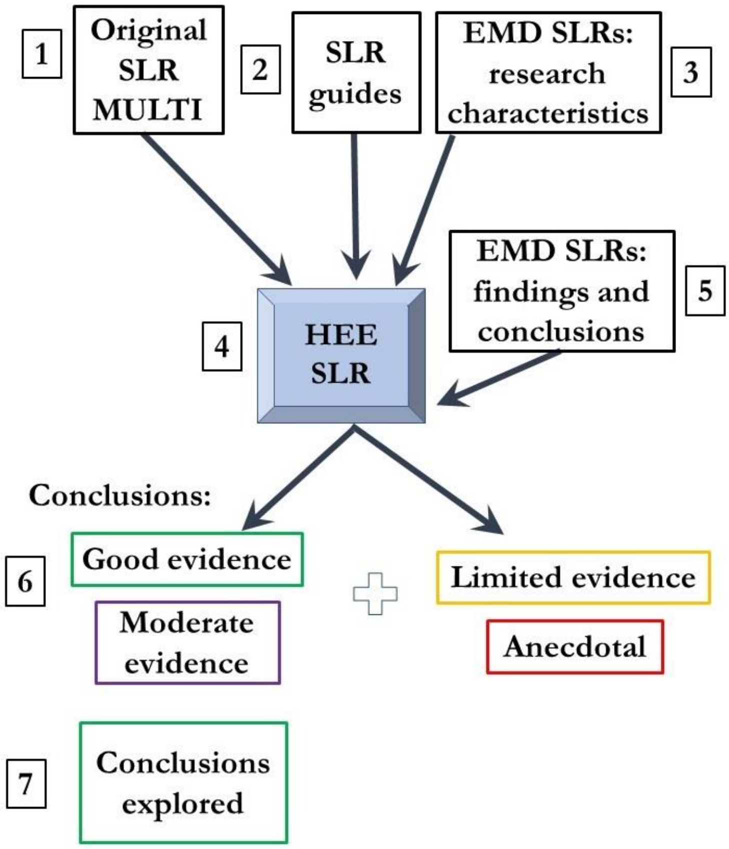


Figure 2.8 PhD Methodology Stage 7.

Above is a portrayal of the seventh stage of the PhD methodology, connecting the two conclusions explored with the sources that were consulted to prepare them: the conclusions from the best available evidence from HEE and the included studies from MULTI and EMD. These two conclusions will be fully explored in chapter six.

Stage eight was another unplanned development that arose by virtue of the systematic evidence analysis methodology. As the list of principles of optimal leadership development and measurement emerged, along with points from the conclusions explored, they seemed to fit into a sequence of design, delivery, and evaluation characteristics that could comprise a complete prototype of a theoretical model. Thus, Stage 8 is the creation of this model, as depicted Figure 2.9 below.

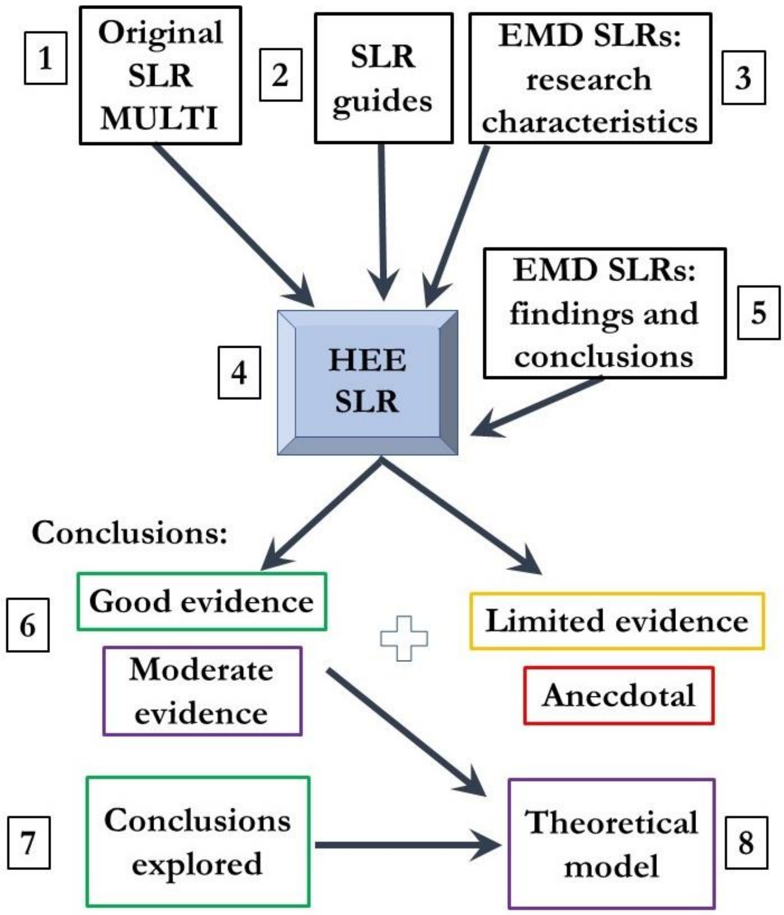


Figure 2.9 PhD Methodology, Stage 8.

Above is a graphic outlining how the conclusions and conclusions explored led to the formation of the outcomes-based prototype theoretical model.

The final stage of the PhD methodology was determining the implications for research and practice. The former arose from Stages 3 and 4, which isolated key research characteristics of the included studies from the three original reviews and their application to HEE. The latter, implications for practice, represent a summary of the conclusions from the best available evidence, the conclusions explored, and the theoretical model for the purposes of real world application. Stage 9 is depicted in Figure 2.10 below.

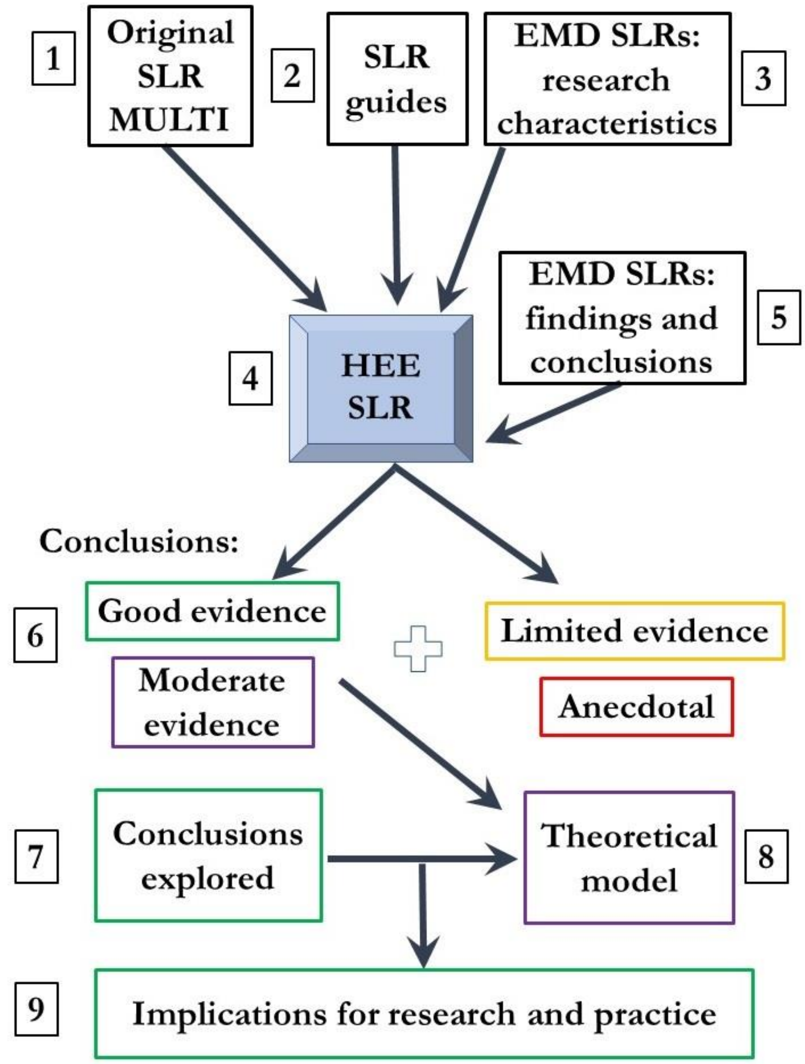


Figure 2.10 PhD Methodology, Stage 9.

Above is the visual outline of the how the previous stages in the PhD methodology informed the formulation of the implications for research and practice. A diagram of the full PhD methodology is included in the appendix on page 347.

2.9 Definitions

The final section of this chapter is the set of definitions that will be used for the remainder of this study. The lack of consensus definitions of the terms that follow makes it especially important for researchers and programme providers to articulate definitions, a process that Hartley and Benington (2010) assert is a key prerequisite for effective leadership in any given setting.

Leadership is the process of leaders and team members collaborating meaningfully to realise a shared vision (J. Geerts, 2009).[2] This definition echoes Rost (1993): "An influence relationship among leaders and followers who intend real changes that reflect their mutual purposes" (p. 124). In this definition, leadership is not restricted to a specific person, role, or position; it is a process in which participation is intentional and voluntary; it necessarily includes "team members" (or "followers," "subordinates", or "supervisees"), each of whom has a meaningful role; and it is directed toward a shared future reality, rather than being simply a relationship. This concept is distinct from management, which strives to make current operations more efficient (Northouse, 2006). Finally, this definition avoids the common mistake, seen in Bohmer (2012), for example, of equating "leader" with "leadership" (Van Aerde, 2013), as if "teacher" and "education" are synonymous.

A **leader** is anyone who takes responsibility of, or is ultimately accountable for, the process of realising a shared vision in a given situation (J. Geerts, 2009). This usually involves setting and communicating the strategy and ensuring that the organisational culture facilitates this process. This definition also means that as the leadership process advances, people can shift roles from leaders to team members and vice versa at different times, but the role of the leader is the one who is accountable for the process's success. This is similar to a football team captain who is substituted off during a match: s/he passes the captain's armband to another player who becomes the leader as part of the same overall mission. This definition also implies that being a leader does not depend on a titular role or position.

A **team member** is anyone who collaborates voluntarily in the process of realising a shared vision (J. Geerts, 2009). This is preferable to "follower" or "subordinate", since it values the person and the contribution of each team member. This term also implies an intentionality about their involvement in the process, rather than simply following, possibly blindly, or positionally being "below" the leader, as in the case of subordinate.

Capabilities: Robinson (2010) defines these as "what people need to be able to do and to be to perform a particular function," which, she claims, "involves a seamless and dynamic integration of knowledge, skills, and personal qualities" (p. 3). This term is often used

[2] The following three definitions emerged from an unpublished Master's thesis by the author whose methodology involved analysing dozens of definitions of leadership and the definitions' component parts before arriving at the ones included in this section.

interchangeably with "competencies," "capacities," and "skills" (Bartram, 2005; Hartley & Benington, 2010).

Clinical leadership is the process of clinical healthcare staff "setting, inspiring, and promoting values and vision," including ensuring that the patient is "the central focus in the organisation's aims and delivery" (Jones et al., 2011, p. 1).

Leadership development is "a continuous, systematic process designed to expand the capacities and awareness of individuals, groups, and organisations in an effort to meet shared goals and objectives" (Dugan, 2011, pp. 79–80), as well as "to participate effectively in leadership roles and processes" (D. V. Day, 2004, p. 841). It is therefore not restricted to individual or formal interventions, nor to individuals. Day's (2000) article makes an important distinction between leader and leadership development. This definition also suggests that leadership development has an agenda and a direction: meeting goals and preparation to lead in a specific context, which implies *application*.

Professionals: those who have specialised knowledge and capabilities based on specific experience, education, and/or training, in a corporately-organised occupation, which forms part of their identity. This draws largely on Freidson (1983), who describes professionals by stating that they "gain their distinction and position in the marketplace … from their training and identity as particular, corporately-organised occupations to which specialised knowledge, ethicality, and importance to society are imputed, and for which privilege is claimed" (p. 25).[3] The definition used for this study is intended to accommodate professionals in the traditional sense, others to whom the term is commonly applied, such as architects and military officers, and senior business leaders whose expertise can come from experience rather than post-graduate education or formal training.

"Developmental activities" refers to the tools and vehicles that are intended to facilitate leadership development and meet a programme's or organisation's objectives (Allen & Hartman, 2008). This term, adopted by McCauley (2008), is used synonymously with "pedagogical components," "development practices" (Day, 2000), "learning activities," and "sources of learning" (Allen & Hartman, 2008). These activities include didactic teaching

[3] Resolving the question of what specific training, knowledge, and occupations qualify and how to determine the importance to society and ethicality exceeds the scope of this study.

methods, resources such as assigned readings, experiential components such as action learning, support methods such as coaching and mentoring, and assessments, such as tests or presentations (Allen & Hartman, 2008). Though naturally interconnected with curricular topics and components, they are treated separately in this study: the “what” (curriculum) versus the “how” (developmental activities).

“Discernible outcomes” refers to tangible improvements at the individual, team, or organisational level, usually singular, that can be taken as valid but are not statistics. Examples would include having received a promotion, having opened a new office branch or site, having implemented one’s action learning project, or having had one’s entire hospital adopt a new policy or protocol. Even when offered by participants without external confirmation, given their verifiable nature, they seem reasonably dependable and yet do not naturally qualify as statistics. This term is useful for outcomes that do not appropriately fit in the categories of self-ratings, since discernible outcomes are more credible, external ratings, which are unnecessary for discernible outcomes, and statistics, since discernible outcomes are often singular.

“Organisational culture,” following Hatch’s (1993) understanding, consists of “a set of assumptions, values, norms, symbols and artifacts within the organization, which convey meaning to employees regarding what is expected and shape individual and group behavior” (p. 657). In addition to the explicit aspects of the culture, implicit indicators of what is valued are what is prioritised, what is supported with resource allocation, what is trained, measured, and celebrated, and what are determinants for selection, promotion, and rewards.

“Self-efficacy” refers to one’s confidence in one’s capacity to execute behaviours necessary to produce specific performance attainments and exercise control over one’s motivation, behaviour, and social environment (Bandura, 1997)

3 Chapter Three: Philosophical Framework

This chapter outlines the philosophical framework, including the ontological and epistemological assumptions, that underlies the current study. It also includes a discussion of randomised controlled trials, which is the MERSQI instrument's highest-rated design.

3.1 Ontological and Epistemological Assumptions and Theoretical Framework

Denscombe (2010) asserts that the philosophical foundations of a study shape its methods and questions; they determine what qualifies as credible evidence according to the researcher; and they give an indication of what kind of conclusions might or might not be able to be drawn from an analysis of the data collected. Many scholars, however, suggest that declaring a definite adherence to an individual position or framework at a study's outset can prove to be limiting and unhelpful. Greene (2008) explains that this is partly because a single approach to research only yields a partial understanding of the phenomenon. Similarly, many, including Ercikan and Roth (2006), suggest that polarising research into either quantitative or qualitative approaches is unproductive, since, particularly in the social sciences, both types of data are needed. The shortcomings of this kind of polarisation are discussed in detail in chapter five. Furthermore, Crotty (2003) points out that in addition to not being watertight compartments, the terms "theoretical framework," "ontology," and "epistemology" are often used interchangeably, or in different or contradictory ways.

A mixed methods approach, on the other hand, allows researchers the freedom to select the philosophical framework and research design that best addresses their research questions, regardless of whether they align with any particular paradigm (Denscombe, 2010). When researchers have the full gamut of instruments at their disposal, they are able to overcome the shortcomings and biases of single approaches, which is particularly advantageous in the social sciences. Denscombe (2010) further contends that a mixed methods approach enhances the accuracy and breadth of data through triangulation and allows for a more thorough analysis of multiple perspectives. The benefits of combining qualitative and quantitative data, along with using multiple sources and times of data collection, are asserted by Dunning (2012) and made evident in the analysis of the included studies in the HEE SLR. For these reasons, quantitative, qualitative, and mixed methods studies were included in MULTI and the HEE SLR.

3.2 Ontology

Ontology is the study of the nature of things, of reality. In social science, Blaikie (1993) asserts that ontology concerns social reality: what exists, what it looks like, what units and structures make it up, and how those units interact with each other. The implications of different ontological stances for leadership development are meaningful when one considers how the three common positions apply to leadership and its development.

3.2.1 Objectivism – Generic

Objectivism, as Denscombe (2010) describes it, with its belief that reality is ordered and includes causes and effects that are external to humans' perceptions, would suppose that leadership principles are generic or universal and transferrable to all situations and contexts. From this perspective, the role of leadership researchers would be to identify and analyse the core principles and capabilities, as well as how to best situate and support great leaders to maximise their impact. Leadership development would serve to simply communicate these principles and enable participants to practice them, regardless of their industry or context. This will be explained in more detail in the following paragraphs. There are many, including Yukl (2010) and Getha-Taylor and Morse (2013), who take the position that leadership generically spans organisations and sectors. The most extreme version of an objective ontology would be similar to Carlyle's Great Man Theory, which essentially supposed that certain people possess innate and superior leadership abilities and do not need training (Carlyle, 2007). One critique of this stance is that scholars have been unable to agree on a consensus definition of leadership or set of capabilities (Gronn, 2004). A second point of contention, as will be described in chapter five, is that studies have shown that the context is important in leadership and that development programmes are likely to fail if they do not take this into account, such as that described in McGurk (2010).

3.2.2 Subjectivism – Contextual

The inverse ontological position of objectivism, is subjectivism, whose premise is that reality is entirely constructed in the minds of people and is reinforced through their interactions with others (Denscombe, 2010), would suggest that leadership is wholly contextual. This relates not only to professional domains, but to organisations, levels of seniority, situations (such as a time of crisis or restructuring) and, in an extreme sense, to each individual leader. Leadership development providers would not be able to say with any certainty whether any programme's content would translate to participants' workplaces, thus rendering the traditional forms of leadership development useless. As a counter to this position, McAlearney et al. (2005), Taylor (2010), and others have argued that many leadership principles apply effectively from one sector to another and that there are core capabilities, such as having a shared vision, that seem to be at least to some extent generic. Thus, strict objectivist or subjectivist ontological positions do not seem appropriate for leadership or its development.

3.2.3 Relevance of Ontology to Leadership Development and Research

The question of ontology has several implications for leadership development. First, as suggested earlier, it determines to what extent findings from one study or in one context are relevant to others. Second, it guides the choice of several aspects of programme design, including selection and the target audience for programmes in terms of open programmes versus those that are highly specific to individual professional domains, organisations, or roles (eg middle managers). For example, Burnes and O'Donnell (2011) contend that insights into how to lead effectively are translatable and "as relevant to [amateurs] as they are to those at the top" (p. 24); whereas others, such as Van Aerde (2013), suggest that different training is needed at different stages of one's career. Third, it affects how much providers emphasise content (more objectivist) or its application and tailor interventions to the individual (more subjectivist). Fourth, it influences how providers and researchers measure the impact of interventions, such as by relying on universal outcome metrics for all participants (objectivist), or by devising different metrics for each participant (subjectivist).

Fifth, a divergence in leadership ontologies between providers/facilitators and participants can adversely affect programme outcomes (Hartley & Benington, 2010), partially due to the latter's expectations. For example, in a programme where the underlying understanding of leadership was largely subjective, a participant who had a strongly objective leadership ontology might judge that the intervention lacked credible content. The inverse situation might leave a participant feeling that the programme was too abstract and wanting more catering to her or his own learning style and applicability to her or his own context. Von Krogh and Roos (1995) state that those designing training and development programmes need to be clear about their understanding of the nature of leadership, as do managers when leading, and human resources personnel when recruiting and rewarding staff. Since there is no consensus definition of leadership or set of capabilities, it cannot be taken for granted that these understandings are shared among providers and participants. Ontology also affects researchers, particularly in terms of what qualifies as "evidence." For example, an objectivist researcher evaluating a subjectivist programme could collect content-based assessments from participants who have focused on the application aspect of the knowledge or on personal development, such as self-awareness. Similarly, in the inverse situation, a subjective researcher could collect individual reports of post-programme benefits and miss an opportunity to use a common metric to identify the representative nature of outcomes. Therefore, if ontological understandings are not aligned among providers, participants, and researchers, programme and research outcomes can suffer.

3.2.4 Ontology Applied to Leadership Development: Specifics

What follows below are descriptions of how the two aforementioned ontological positions would apply to different aspects of leadership development. First, the programme **design overall** for an objectivist would be based exclusively on generic principles of optimal leadership development. A subjectivist would adapt every aspect of every programme to the individual participant's needs and preferences.

The **sample** for a leadership intervention from an objectivist perspective could be of any size, but it would be ideal if all the participants were considered great leaders or at least have obvious potential to become them. In terms of selection criteria, participants would preferably be nominated by their supervisors based on their superior leadership ability, not self-selecting. A control group could be included to measure post-programme performance. Since the principles of great leadership are considered generic, there is no need to restrict the samples to the same domain, profession, organisation, or level of seniority. From a subjectivist point of view, the ideal sample size would be one individual or one team from the same organisation, since leadership is considered contextual. For this reason, there would be no control group, since their contexts are too different to be useful anyway, and the sample participants would volunteer for the programme, which would be as specialised as possible to the participant(s)' domain, profession, organisation, and level of seniority.

The **programme**, designed by an objectivist, would only use a needs assessment to determine in what ways the participants' superior leadership skills are needed in the organisation. The intervention would centre on a universally-authoritative capability framework. The goals would be content-heavy and would convey knowledge of generic leadership principles to identified great leaders. The location is less relevant, but the programmes could be replicated anywhere without changing the content or format. The faculty would be established technical experts and leaders in their field, that is, great leaders themselves. Finally, the developmental activities would be lectures, case studies and reading materials on key leadership principles and accounts of great leaders, guest speakers, and stretch assignments and action learning projects that would enable great leaders to put their superior skills to further use. As mentioned previously, in the purest form of leadership objectivism, leaders would not need any development whatsoever; they would just need opportunities and for everyone else to stay out of their way while they are left to do what they do best.

For a subjectivist, a needs assessment is essential for determining what would best suit the individual or team given her/his/their context. The only value a capability framework could offer unless designed specifically for the participant(s) would be to facilitate discussion and reflection on what the participant feels is best for her/him. The programme goals would be person- and application-centred and focus on self or team-development and self-awareness. The topics would be totally adapted and flexible, often in discussion with the participant. The location would likely be internal, but regardless, it would need to be highly specialised. The faculty would not need to be great leaders, but rather, skilled at coaching and facilitating so as to enable participants to apply their knowledge and skills and develop their self-awareness. Finally, the developmental activities that a subjectivist would include are coaching, 360s, action learning projects to be implemented in the participants' own workplace, reflection, and Personal Development Plans (PDPs).

Finally, in terms of **measurements**, an objectivist would focus on Kirkpatrick Level 2b (objective knowledge and skills), 3b (objective behaviour change), 4a (objective organisational impact), and 4b (objective benefit to clients or patients). They would rely on statistics and external raters to provide this data. A subjectivist would lean towards Level 1 (participant satisfaction), Level 2a (change in attitude and perceptions), 2b (subjective increase in knowledge and skills), 3a (subjective behaviour change), 3b (objective behaviour change), 4a (subjective organisational impact), and 4b (subjective benefit to clients and patients). They would collect subjective numbers and descriptions of outcomes. For a summary of the ways in which ontology affects leadership development, see Table 3.1 below.

While these polarised views may seem extreme, most programmes tend to lean toward one side of the spectrum or the other and, as mentioned previously, it is useful for providers, participants, researchers, and readers of studies to be clear about the programme's ontology.

Table 3.1

Ontology Applied to Leadership Development

Programme Details		Ontological Position		
Category	**Variable**	**Objectivism**	**Middle**	**Subjectivism**
Sample	Size	Large, as long as they exhibit great leadership skills or have the obvious potential to excel		One individual or team is preferable
	Selection criteria	Nomination by supervisors based on exceptional leadership competence		Application
	Control group	Yes		No
	Profession/specialty	Any		Highly specialised
	Level of seniority	Any		Highly specialised
	Interdisciplinary	Yes		No
Programme	Needs assessment?	Only needed to see in what ways superior leaders can impact the organisation		Yes, tailored to the individual and her context
	Capability framework?	Yes		No, or only to facilitate self-reflection/awareness
	Programme goals	Content-heavy: convey knowledge and prepare great leaders		Application-centred and focused on self or team development and awareness
	Topics	Universal: principles of great leadership		Totally adapted and flexible, in discussion with the participant
	Location	External or a replicable programme		Internal or highly specialised
	Faculty	Established technical experts and leaders in their field		Experts in coaching and facilitating
	Developmental activities	Lectures, case study analysis, guest speakers, reading assignments on great leaders, stretch assignments, and action learning projects		Coaching, 360's, action-learning, reflection, PDP's
Measurements	Kirkpatrick outcome levels	2b (obj), 3b, 4a (obj), 4b (obj)		1, 2a (subj), 3a, 3b, 4a (subj), 4b (subj)
	Raters	Statistics, external		Self-reports
	Type of data collected	Objective		Subjective numbers and descriptions

Above is a depiction of the application of the two extreme ontologies to various aspects of leadership development and research.

3.2.5 The Third Option

In light of this discussion, this study's ontological position is that there are leadership principles and principles of optimal leadership development that are to some extent universally or widely applicable, a premise which gave rise to the title of this thesis. Moving forward, if indications that some aspects of leadership are generic appear, then four questions ensue: 1) Which of these leadership principles or capabilities translate effectively and which appear to be contextual? Many, including Gronn (2010), argue that research into this question is needed because of the differing contextual dynamics of leadership. 2) Of those that appear to be generic, how directly do they transfer and how much adaptation to each specific context is required? The second pair of questions are the same as the first, but pertain to principles of optimal leadership development. This relates to the goals, content, developmental activities, programme components, and evaluation. This study focuses primarily on the third and fourth questions.

Along with the statements above regarding possible generic aspects of leadership and leadership development, recent work by Goodall and others highlights a contextual facet of leadership. Goodall's (2011) Theory of Expert Leadership asserts that being a technical expert in terms of knowledge and experience of, and skills and performance in, the core business of the organisation produces the best organisational performance. The "core business" is described as the most important or primary endeavour of an organisation in terms of its success and profits (Zook & Allen, 2001). An "expert" is one who demonstrates exceptional performance in this specific domain of activity (P. Johnson, Zualkernan, & Garber, 1987), in addition to acquiring the domain-specific knowledge and technical abilities and skills (Alavi & Leidner, 2001; Nonaka, 1994). This exceptional status is accomplished by working one's way up, often over a long time, and as a consequence is also able to do the job of one's subordinates to a high standard (Artz, Goodall, & Oswald, 2016; Goodall, 2009; Goodall, Kahn, & Oswald, 2011; Goodall & Pogrebna, 2015). Examples include an outstanding scholar running a university, rather than a non-academic business manager or a less prominent scholar. Goodall and Baker (2015) add a final point: that being an expert leader includes having management and leadership skills, whether through training, innate ability, or experience.

This suggests that although many generic leadership principles may apply commonly across professional domains, a significant determinant in leaders' effectiveness in the role is this notion of expert leadership (Goodall, 2011). This is why, to take one example, there is no middle entry in the military: senior officers have all proceeded up the ranks and proven themselves from the most junior level. This piece will be further discussed in the summary of the answer to the fourth research sub-question relating to the generic versus generic nature of leadership, which is located in the discussion in chapter seven.

This thesis also recognises that the cultural and organisational context in which one leads is important for providers and researchers to consider, as are several factors before, during, and after leadership development interventions that significantly affect the application of learning. These are described in detail in the second conclusion explored. This is in line with Denscombe's (2010) contention that this type of ontology imagines that the reality of the social world as varying between cultures and groups, rather than there being a single, objective reality. Mason (1997) adds that this position values various people's knowledge and experiences, which is a key principle of adult learning (Knowles, 1984). This will be explained in the first conclusion explored in chapter six as particularly important when studying professionals and leaders who have accrued a wealth of experience.

Taylor (2010), a surgeon by training, makes an interesting claim that is worthy of further investigation: "for all intents and purposes, therefore, the leaders in business, the leaders in medicine, their behaviours are the same, but the atmosphere, administrative structure, and culture differ in many ways" (p. 49). This contention holds both aforementioned extreme ontologies together. From a leadership development provider's perspective, combining aspects of both ontologies means presenting empirically-supported content and enabling participants to determine whether or how they can be applied to their individual context. Another way this can manifest itself in the programme design stage is by performing a needs assessment to determine what leadership outcomes are needed for a particular intervention, given the participants and their organisational context. The role of research is to investigate the extent to which leadership and leadership development for professionals is generic versus contextual and to identify optimal ways to enable participants to apply leadership principles to their professional situation.

3.3 Epistemology and Theoretical Perspective

Epistemology is the study of how truth about reality can be acquired, which for researchers is one's theory of knowledge. Blaikie (1993) explains that in the social sciences, this relates to possible ways of gathering knowledge about the social reality.

3.3.1 Positivism

A positivist researcher, normally associated with an objectivist ontology, would gather knowledge of the social reality by electing for a confirmatory design, scientific methods, a randomised controlled trial (RCT) experiment methodology with a control group, data collection methods including statistical analysis, and purely quantitative data (Denscombe, 2010). These are excellent for establishing causation and adding credibility to findings and reported outcomes. Failing to use dependable methods of data collection or relying purely on qualitative data diminishes studies' quality and the strength of their findings and conclusions, as was evident in the HEE SLR, which will be described in chapter six. The drawback of the positivist approach, according to Cohen et al. (2010) and Alvesson and Spicer (2012), is that this type of research disregards the social and contextual factors that have already been explained as important to consider in leadership development.

3.3.2 Constructionism and Interpretivism

A main advantage of the constructionist epistemology and corresponding interpretivist theoretical framework is engaging complex phenomena by gathering data that reflects an appreciation for each person's experiences, which are likely to differ (Denzin & Lincoln, 2003; Schwandt, 1998). As suggested previously, it is very useful for researchers to pay attention to and account for every participant's experience of leadership development, especially outliers', to fully understand the nuances of the process and its outcomes. This breadth of analysis is not possible when only quantitative data is gathered, nor when only majority opinions are highlighted (Alvesson & Spicer, 2012). Alvesson and Spicer (2012) assert that the rich data collected by in-depth qualitative inquiry of all perspectives allows for a more complete understanding of the phenomenon, which is why including outlying and negative perspectives was a variable of investigation of the reviews included in this study. Likewise, including manifold perspectives from multiple raters adds further scope and credibility to the understanding of to what extent, how, in what ways, for whom, and in which circumstances programmes are effective or not. When literature review and single study authors provide thorough information about the feature organisation(s) and their contexts, it enhances the potential for generalisability to other contexts (Yin, 2003). This is further augmented when multiple sites are analysed and presented. Denscombe (2010) suggests that the common challenges to the constructivist epistemological position are that their findings lack rigour and are anecdotal and incomplete. This is because of its denial that universal truths can be discovered, favour for only qualitative data, and lack or programme-wide outcome metrics.

When used to reinforce, explain, or nuance quantitative data; however, as it was used in this study's conclusions and conclusions explored, this approach can be very valuable.

3.3.3 How Epistemology Relates to Leadership Development Research: Specifics

Epistemology affects leadership development research in several ways. As suggested earlier, in terms of **study design**, a positivist researcher's purpose would be to identify or test generic leadership or leadership development principles, or measure great leaders' performance and impact. The theoretical framework would be confirmatory and there would be only objective data collected, which would be done by way of experiments, quantitative surveys, or action learning projects where researchers trial how organisations can enable great leaders to maximise their impact. The data collection methods would be experiments, questionnaires (closed-ended questions with quantifiable data), statistical analysis, participant observation, and video analysis of great leaders in action. Researchers approaching leadership development from a constructivist or interpretivist epistemology would, in the purest form, merely encourage participants to reflect on their own experience, since no findings would be thought to apply properly to other contexts. Otherwise, the purpose of research would be to discover the details and nuances of how participants experience, benefit from, and apply learning from, interventions to their own context. The theoretical framework would be exploratory, the data collected would be subjective, the methodologies would be grounded theory, ethnography, or case studies of individuals or teams. The methods would be interviews, questionnaires (open-ended), document analysis of participants' journals, participant/programme observation, and video analysis.

In terms of **measurements**, a positivist would utilise some variation of the following outcome metrics: Level 2a: increased motivation, Level 2b (obj): increased knowledge, Level 3b (obj): promotions, taking on a greater leadership role, awards, meeting personal goals, Level 4a (obj): developing and implementing a new programme, meeting organisational goals, statistically improved organisational performance, Level 4b (obj): implementing an action learning project, improved clinical outcomes. They would rely on statistics and external raters and would collect data pre, post, and post-post to measure the performance of great leaders and teams. For a constructivist or interpretivist, the outcome metrics would be: Level 1: PPEs, Level 2a: increased motivation, confidence, aspirations to lead, engagement, self-awareness, Level 2b: self-reported increased knowledge and skills, Level 3a (subj): increased leadership behaviours, meeting personal goals, Level 3b: taking on a leadership role, Level 4a (subj): meeting organisational goals, general organisational benefits, Level 4b (subj): implementing an action learning project, benefit to clients or patients. The data would rely on self-reports and would be collected ideally at all points: pre, during, post, and post-post. For a summary of this, please see Table 3.2 below.

Table 3.2

Epistemology Applied to Leadership Development Research

Programme Details		Epistemological Position		
Category	**Variable**	**Positivism**	**Middle**	**Constructivism and Interpretivism**
Study design	Purpose	To identify or test universal leadership and leadership development principles, or measure the performance of great leaders or teams		Purest form: just to encourage participants to reflect on their own experience. Otherwise, to discover the details and nuances of how participants experience, benefit from, and apply learning from, interventions in their own context.
	Theoretical framework	Confirmatory		Exploratory
	Type of data collected	Objective		Subjective
	Methodology	Experiments, survey, action learning		Ethnography, grounded theory, case study
	Methods	Experiments, questionnaires (closed-ended questions), statistical analysis, participant observation, video analysis		Interviews, questionnaires (open-ended questions), document analysis (participants' journals), participant observation, video analysis
Measurements	Kirkpatrick outcome levels	2a, 2b (obj), 3b, 4a (obj), 4b (obj)		1, 2a (subj), 3a, 3b, 4a (subj), 4b (subj)
	Outcome measures	**2a**: Increased motivation, **2b**: Increased knowledge, **3b**: Promotions, taking on a greater leadership role, awards, meeting personal goals (obj), **4a**: Developing and implementing a new programme, meeting organisational goals (obj), statistically improved organisational performance, **4b**: Implementing an action learning project, improved clinical outcomes		**1**: PPE's, **2a**: Increased motivation, confidence, aspirations to lead, engagement, self-awareness, **2b**: Self-reported increased knowledge and skills, **3a**: Increased leadership behaviours, meeting personal goals, **3b**: Taking on a leadership role, **4a**: Meeting organisational goals (subj), general organisational benefits, **4b**: Implementing an action learning project, benefit to clients or patients (subj)
	Raters	Statistics, external		Self-reports
	Times of data collection	Pre, post, post-post		Pre, during, post, post-post

The table above depicts how the two opposing epistemologies would apply aspects of leadership development research.

3.3.4 Randomised Controlled Trials (RCTS)

Randomised controlled trials (RCTs) are widely considered the gold standard in research in terms of demonstrating causality (Mark, 2008; Sullivan, 2011), particularly when they control for alternative explanations that would otherwise be plausible (T. D. Cook & Campbell, 1979). This explains why the MERSQI instrument reserves the highest points for this design. This methodology involves controlled, quantitative, comparative experiments that attempt to draw a causal relationship between outcomes and interventions (Sullivan, 2011).

Key principles and advantages

Freedman (2012) cites three key features of RCTs:

1) The outcomes of experimental subjects assigned to receive a treatment is compared to the outcomes of subjects assigned to a control group. The control condition is often defined as the absence of treatment, but it need not be. This feature is not exclusive to RCTs.
2) The assignment of participants to treatment and control groups is done at random, through a randomising device such as a coin flip
3) The manipulation of the treatment – also known as the intervention – is under the control of an experimental researcher

As a contrast, in observational studies, treatment assignment is not usually random; participants typically self-select into the treatment group. Observational studies also have no experimental manipulation, which is what makes them observational (Dunning, 2012).

There are many advantages associated with RCTs. The first is that they are designed to make the relationship between cause and effect clear and to obviate confounding, both in terms of possible causes and effects (Dunning, 2012). A confounder or confounding variable is an unobserved variable that is a potential cause of an outcome being studied that leads to distortions or false associations (Dunning, 2012). The second advantage is that the effect sizes of the treatment group in RCTs are seen as being determined with less bias and more internal validity than observational studies (Sullivan, 2011). Another advantage is the relative simplicity and transparency of the data analysis (Dunning, 2012). The straightforward comparison of the difference in mean outcomes between the two groups often suffices to estimate a causal effect.

There are also two common features of RCTs: randomisation, mentioned above, and counterfactual conditionals.

Randomisation

The RCT methodology randomises those who receive the treatment and intervention and, in some cases, who does not. It is common to have an RCT without a placebo or a group that receives no treatment, as is the case when two groups are compared who experience different versions of the same treatment, with no group receiving no treatment. Comparing the outcomes of the two groups is what leads to the impact of the intervention. In practice, a randomisation approach involves the researcher identifying a target population and the randomising which participants receive the treatment within the population. In this case, each participant has an equal probability of being selected from the population, ensuring that the sample will be representative (Cresswell, 2003).

Randomisation generally depends on equal variables between groups, which increases the confidence that outcome differences can be attributed to the intervention (Sullivan, 2011).

Randomisation removes confounding because it establishes an pre-intervention symmetry between the groups (Dunning, 2012) and that any unobserved explanations of outcomes and confounding factors will be symmetric across the treatment and control groups. This symmetry ensures that sizable differences between the two groups provide reliable evidence for the causal effect of the treatment (Dunning, 2012). Thus, randomisation produces *statistical independence* between these confounders and treatment assignment (Dunning, 2012). To summarise, randomisation ensures that any differences in outcomes between the groups are due either to chance error or to the causal effect (Dunning, 2012).

Advocates argue that without random assignment, selection differences are likely to occur (Mark, 2008). This means that even without the intervention making a difference, the individuals in the treatment group would probably differ initially from those in a comparison group on average. Again, random assignment precludes any systematic bias due to initial selection bias (Mark, 2008) and is said to increase internal and external validity (Dunning, 2012).

Cluster randomisation

Instead of randomising individual participants, randomisation can also be done at the cluster level, where the primary sampling unit is a bunch or cluster of individuals, such as a city or healthcare centre (Maxim, 1999). In this case, the unit of randomisation is the group which will randomly receive the treatment. The unit of analysis at which data is collected and outcomes are compared is often the individual, such as students' tests scores, which are then analysed by comparing intra and inter-cluster results.

Clusters are particularly effective when the random selection of individual elements is ineffective (Maxim, 1999) or logistically impractical. Second, cluster control trials can address contamination or instability where individuals share or discuss treatment with individuals in the control group, which could potentially affect the impact. Third, randomising at the cluster levels can mirror the level at which the intervention would actually be implemented, such as healthcare teams.

Counterfactual conditionals

A second underlying principle of RCTs is counterfactual conditionals (Dunning, 2012), which are largely addressed through the lack of a systematic selection bias for the control group (Mark, 2008). The term counterfactual conditionals refers to possibilities of what would have happened to the same individuals at the same time if they had not received the treatment or intervention and vice versa, how others would have fared if they had received the treatment (Dunning, 2012). These can be challenging to tackle, since comparing the same individual at different stages over time will not usually give an accurate estimate of a treatment's impact, given the various other factors that affect internal validity. The way that RCTs achieve this is by way of strong estimates in contrast to the control group, which mimic these conditionals.

Blindness

A third common feature of RCTs is blindness of the part of the researcher to treatment-assignment status, which is said to minimise bias and increase internal validity (Dunning, 2012).

Validity

Internal validity is the extent to which one can demonstrate that one's treatment produced the changes and is thereby having an impact on a given outcome (the dependant variable) and that other sources of influence have been controlled (Jackson & Verberg, 2007). The experimental procedures, treatments, or experiences of the participants can strengthen or threaten the internal validity and thereby, the researcher's ability to draw correct inferences from the data (Cresswell, 2003). Examples include using inadequate procedures, such as changing the instrument during the experiment, issues with the intervention, such as experiment and control group members discussing with each other, or challenges with the participants, such as if they were attending an additional development programme at the same time.

Dunning (2012) states that RCTs posses a high level of internal validity because confounding is accounted for, the data collection methods are credible, and the data analysis is simple and transparent by comparing the effects of the control group to that of the experiment group.

External validity refers to the extent to which one can make extrapolations from the effects of a particular study to other groups in general and to substantive and theoretical questions (Dunning, 2012; Jackson & Verberg, 2007). This is sacrificed when researchers draw inappropriate inferences from the data to other populations, contexts, or past or future situations (Cresswell, 2003).

Dunning (2012) suggests that in RCTs in social sciences, the sample group is usually a convenience sample selected from a non-random process, such as those participating in an intervention. This means that they are often not representative of the population as a whole. The randomisation of participants to experiment and control groups ensures that the effects of the treatment *for the study group* are unbiased, but that does not necessarily mean the results are generalisable to other populations (Dunning, 2012). Thus, Dunning (2012) cautions against making strong claims of generalisability from any one RCT without replicating the experiment, though he adds that observational studies face the same challenges and often do not even feature a control group.

Three points should be made in reference to the external validity of RCTs. The first is that some experiments *are* conducted in multiple sites, such as that by Ten Have et al. (2013). Second, other experiments are amenable to replication, a point that is included in this thesis's recommendations for further research. Third, Cresswell (2003) suggests that robust reporting of the findings, sample, and context can enhance external validity, another point that is echoed in this thesis. Furthermore, like case studies, when the context, sample, and intervention are described in detail, readers can decide for themselves the extent to which the study's findings apply to their own situation (Boaden, 2006). Dunning (2012) adds that the external validity is enhanced when the intervention attempts to inform interesting, wider substantive and theoretical questions.

Application to Leadership Development

Traditional forms of RCTs are exceedingly rare in leadership development studies, as demonstrated in this study. This is because the key hallmarks of RCTs mentioned above can be difficult to implement.

Randomisation is rare and a significant challenge since few organisations would allow their leaders, especially at higher levels, to risk being randomly allocated to the control group (Collins & Holton III, 2004). Including a control group through comparisons of equal seniority leaders within the same organisation may lead to threats to internal validity, such as interaction of selection and experimental treatment (Campbell, 1969). There are two alternatives. The first is to compare two separate interventions with no placebo group, which allows researchers to assess the various perceived merits and shortfalls of each treatment with every participant potentially benefitting. The second alternative to be discussed below is a cluster randomised controlled trial.

It can also be challenging for researchers to have direct control over the treatment, particularly in a way that isolates individual variables. Naturally, an action learning RCT is an option, however, this study has shown that such a practice was not found in the literature.

It can also be difficult at times to separate the effect of an individual intervention from other external influences, such as one's natural career progression or participants benefitting from additional developmental opportunities. Similar issues can arise when attempting to separate an individual's development when working in complex teams and gathering data at the team or organisational level. Lastly, it can also be challenging to handle confounding conditionals and makes analysing the effects of individual elements challenging, given the reality that most leadership programmes are delivered as packages (Galli & Muller-Stewens, 2012).

Perhaps the most effective solution is the cluster randomised controlled trial, as was the case with the Jeon et al. (2013) study. This is why it was made the feature article, as described in chapter seven. Randomising at the cluster level in terms of teams or departments at different healthcare centres can potentially prevent the interaction of selection and experimental treatment threat (Campbell, 1969) successfully, among other advantages. This approach also reflects the team level at which the intervention is actually implement (Maxim, 1999). In practice, the control group in a cluster RCT can continue its business as usual, providing the usual care, whereas the experiment group will do likewise with the added effect of the intervention. One can reasonably expect the confounding conditions to be accounted for when the control/experiment group matching is balanced and equal (such as by location, size, and type of hospital etc), particularly when regularly collected metrics are used, such as workplace satisfaction, and especially when clinical outcomes are the or part of the focus. As mentioned previously, randomising is much more feasible at the cluster level and blindness can be possible as well. Lastly, although it can still be challenging to separate which aspects of the treatment are attributable for any changes that present, the impact could reasonably point back to the intervention if such changes occur.

Limitations of RCTs

Despite the value of RCTs, there are limitations as well. Even respected advocates acknowledge the challenges in successfully implementing experiments outside of carefully controlled laboratory settings (Mark, 2008). Conversely, many causes of interest to social scientists are difficult to manipulate experimentally (Dunning, 2012). Third, as discussed previously, random assignment may not be feasible for practical or ethical reasons (Mark, 2008). Fourth, random assignment is further obstructed when the researcher is not also the developer of the intervention being tested (Mark, 2008). Fifth, the focus of RCTs can be considered too narrow and too local to isolate what is optimal in development, to design policy, or to advance knowledge about developmental processes (Deaton, 2009). Sixth, another limitation is that RCTs can be considered valid at the micro level but not in terms of aggregation upwards to broader knowledge or expand to why (Francis Fukuyama quoted in Dunning, 2012). Lastly, the quality of quantitative impact evaluation is dependent to some extent on sample size and post-post tests.

Alternatives

Before moving on, it is important to consider similar alternatives to RCTs. Three are discussed below: natural experiments, quasi-experiments, and regression analysis. Rather than exhausted treatments of each, a brief description will be provided followed by the relevance to leadership development research and the limitations of the approach.

Natural experiments are similar to true experiments in that they compare outcomes of control and experiment groups and often feature random or as-if random allocation to investigate the effect of causes (Dunning, 2012). These aspects distinguish natural experiments from typical observational studies where the intervention assignment is *not* as-if random. Other methodologies attempt to control confounders statistically after data is collected, rather than by comparing just the effects of the two groups, having accounted for confounders at the design stage (Dunning, 2012). In the case of natural experiments, data is collected from "naturally" occurring phenomena, rather than researchers having direct control over the treatment variables as in a laboratory experiment (Dunning, 2012). As with true experiments, natural experiments often feature simple and transparent data analysis and are grounded in credible hypotheses about the data-gathering process (Dunning, 2012).

Unlike true experiments, natural experiments typically involve mixed methods, collecting both quantitative and qualitative data, since the natural phenomena are often similar to case studies (Dunning, 2012). Dunning (2012) suggests that this approach can illuminate causes and effects in a superior way to typical true experiments and observational studies. The reason is that they can establish causality more convincingly because of the as-if randomisation that limits self-selecting confounding (Dunning, 2012). Further, natural experiments enable researchers to investigate the effects of variables that are impossible or challenging to control in true experiments (Dunning, 2012). As mentioned previously, in leadership development studies, both are crucial. Lastly, Dunning (2012) suggests that robust qualitative data can validate the causal and statistical models employed in quantitative analysis.

A limitation of natural experiments is that since the researcher does not have direct control over the treatment variables, they cannot be isolated to prove causation between an individual element of the intervention and outcomes. This is a central challenge in leadership development research, which is common to observational studies as well.

Quasi-experiments: Quasi-experiments involve the study of the casual effects of a treatment between apparently equivalent control and experiment groups (Jackson & Verberg, 2007; Maxim, 1999). Like natural experiments, quasi-experiments tend to lack the randomisation of participants; however, the distinguishing features are that quasi-experiments are planned treatments, not naturally occurring events, and that researchers often have some degree of control over the intervention (Maxim, 1999). Although Maxim (1999) says quasi-experiments should only be used when true experiments are not possible, describing the former as "inevitably a second-best approach to testing hypotheses" (p. 176), for reasons described previously such as instances when randomisation is not possible, they can be effective.

The limitations of quasi-experiments are similar to those of true and natural experiments in terms of challenges isolating specific variables, as well as accounting for confounding conditionals.

Regression analysis: Another common approach is regression analysis, which is a group of statistical processes designed to analyse the relationship among the dependent variable and predictors, or independent variables, one at a time (Dunning, 2012). While there is value to this approach, Dunning (2012) lists several limitations. The first is that establishing conditional independence, which involves accurately identifying and measuring confounding variables, is challenging. A second limitation arises when influential variables are overlooked or irrelevant or poorly measured variables are included, providing false conclusions. Finally, the direction and extent of confounding is often unverifiable (Dunning, 2012).

The strengths of RCTs: a review

Despite the aforementioned methodological limitations, RCTs have a valuable role in leadership development research, particularly cluster RCTs. Dunning (2012), Reed et al. (2007) who designed the MERSQI instrument, and others contend that RCTs are the most reliable way to provide credible evidence of causal effects of interventions in dynamic social science contexts. Without random or as-if random assignment to balance the variables, unobserved or unmeasured confounders may threaten valid causal inferences (Dunning, 2012). Dunning (2012) adds that RCTs provide opportunities to learn about the direction and size of the causal effects, which, he suggests, alternative methodologies typically cannot.

There are two essential points to highlight. The first is that including as many of the principles of optimal research, including those described above, is more important than rigid adherence to one methodology or another, especially given that different designs can address different questions. The second key point is that combining highly credible quantitative data with illustrative and nuancing qualitative data is the type of information most needed in the field.

3.3.5 A Combined Approach

As with the ontological position mentioned above, the epistemological stance and theoretical framework that seem to be most appropriate to study leadership development is somewhere in the middle, drawing key elements from both opposing positions. Thus, for reasons stated in the previous paragraphs, it is preferable to gather both quantitative and qualitative data using a range of methods and sources of data. The systematic evidence analysis methodology was deemed most suitable for the kind of multi-layered investigation that this thesis sought to undertake. The intention was to collect robust, rich detail of the phenomenon that respects each participant's experience, while still presenting a variety of quantitative and qualitative outcomes. Schwandt (2000) argues that this type of approach can appreciate subjectivity without dismissing the objectivity of knowledge, which holds both aspects of social science in a harmonious tension resulting in a richer study. Finally, as discussed in the introduction, the ultimate measure of leadership development is the *application* of leadership learning to one's workplace (Raelin, 2011). Thus, this study's aim is both to investigate claims of cause and effect between leadership development and performance outcomes, as well as discussing how, why, for whom, and in which circumstances these are most effective or not. In order to do this accurately and robustly, qualitative, quantitative, and mixed methods studies are required as part of the systematic evidence analysis methodology.

4 Chapter Four: Justification of the Sample and HEE SLR Methodology

This chapter presents the justification of the sample choice of professionals, as well as doctors, and the unique features of medical leadership. The second half outlines the methodology of the HEE SLR in detail.

4.1 Sample: Justification of the Focus on Professionals and Leaders at All Levels

Professionals were selected as the focus of this study for several reasons. The term "professional" is not restricted in this study to the traditional four professions mentioned in chapter one, but includes those whose careers involve occupying a leadership role in a corporate organisation, such as CEOs or military officers. The first reason for choosing this sample is that professionals are a group of generally respected, often highly educated people and the collective term is widely understood, despite the lack of a consensus definition. This makes professionals a commonly intelligible and credible sample, which increases the likelihood that findings could transfer to other similar domains. This is especially true of those involving other professionals even if no studies from their domain were identified in the study, such as lawyers. Bryson, Stokes, and Wilkinson (2017) suggest that findings from studies of physicians, for example, would be considered representative of other knowledge-workers, such as teachers, scientists, and employees working in public or non-profit sector organisations. An interesting feature of leaders in the professions is that they must lead a range of others, including many who are as educated and experienced as they, such as doctors leading other doctors. Taylor (2010) asserts that this can be challenging, since he says that as a result of their education and professional status, there is invariably a sense of entitlement among doctors and surgeons. This dynamic renders more authoritarian styles of leadership ineffective and increases the need for support and development for leaders, especially when leaders and team members are at relatively similar levels of experience.

Choosing professional leaders at all levels, including senior and executive leaders, represents *purposeful* sampling, which entails selecting participants who are most appropriate for researchers to answer the research questions based on their knowledge, experience, and different perspectives (Creswell, 1998). Studying professional leaders makes the findings more likely to be accepted and incorporated widely, given the amount of experience they have accrued and the high level of responsibility they bear in at times stressful circumstances. This is true because they frequently operate in high-pressure, constantly changing environments and in large, complex organisations with often enormous numbers of staff and clients (Edler, Adamshick, Fanning, & Piro, 2010). Thus, their realm of influence can be wide and their decisions can affect many people, for some, the whole organisation, which heightens the importance of good leadership and thereby leadership development. For these reasons, according to Miles and Huberman's (1994) sampling typology, doctor leadership training especially at the higher levels qualifies as a *critical* case because of the intensity of the environment and the factors listed above which make its potential application to other contexts high. Similarly, in the case of business, leaders' decisions can affect their employees' and clients' livelihoods and in the military and healthcare, for example, soldiers', civilians', and patients' lives are on the line (Edler et al., 2010). Professional leaders tend to work in environments with clear organisational structures and hierarchies, which make them easily intelligible to others, along with comparisons to their own organisational structure. Finally, many including Edler et al. (2010) and Taylor (2010), suggest that there are commonalities among the skills required across professional domains, such as proactive decision making, collaboration, and cooperation. For these reasons, leadership and development understandings for professionals at all levels appear to have the potential for wide-spread generalisability.

Equally, the likely potential applicability to other contexts is far greater with professionals than amateurs, such as a professional Navy captain who has commanded in battle versus a teenage, reservist Navy cadet. Although more discussion of leadership development for different levels of seniority will follow in the next chapter, the findings of programmes for senior leaders are not presumed to be restricted to that specific level, since Burnes and O'Donnell (2011) assert that leadership principles apply to leaders at all levels. Most importantly, professionals and senior leaders are an effective sample for leadership development research for two reasons. The first is that their knowledge and experience is likely extensive and higher than those of students or very junior leaders. This experience represents a frame of reference that enables them to make seasoned judgments of an intervention and its impact. Second, in the case of senior leaders, since they are at the top levels of their organisations, their ability to influence co-workers, organisational outcomes, protocols, and policy is much higher than that of junior leaders, which heightens the need for empirical data to inform programmes at that level (Gilpin-Jackson & Bushe, 2007). As mentioned earlier, it is also interesting to compare the findings of studies of leadership development programmes for senior leaders with professionals at middle and junior levels to investigate the extent to which the principles are common or distinct at different stages.

Therefore, studying leadership development for professionals and leaders at all levels is important for this thesis mainly because of the potential for findings to be considered relevant to other contexts and also to examine the extent to which leadership development translates across levels of seniority. This sample features people who are often highly educated, have direct experience leading, usually in clearly identified roles (eg managers or CEOs), and operate in environments where the pressure and stakes are often high.

4.1.1 Justification of the Medical Profession and Doctors

There are several reasons for selecting physicians as the sample for the investigation of leadership development for professionals for the HEE SLR. The first is that, as explained in the first chapter, although the profession is well-established, the formal training of doctors as leaders is relatively new, emerging largely in the past two decades (Lee, 2010). Schwartz et al. (2000) declare the need to establish an empirically-based curricular foundation, which is a process they describe as being still in its infancy and still evolving. The number of leadership development programmes for doctors is growing and it is now seen by many as a priority; yet, there are significant gaps in the research. Second, as demonstrated by the HEE SLR, despite its relative newness of the phenomenon, there is a body of available literature, though it has not yet been explicitly outlined what precisely is known and with what calibre evidence. Third, the enormity of the healthcare field necessitates effective leadership. Jones, McCay, and Keogh (2011) assert that, in reference to healthcare, "Every person in the world needs it, high proportions of gross domestic product (GDP) are spent on it, governments are judged on it, populations are determined by it, and almost everyone has an interest in how it is delivered" (p. 1). In the UK, for example, eight per cent of the GDP is spent on healthcare; and the NHS employs 1.4 million people, which makes it the third largest civilian organisation in the world (Jones et al., 2011). Consequently, Jones, McCay, and Keogh (2011) state that for healthcare organisations to deliver high-quality care, effective leadership is needed at every level.

A fourth reason for selecting physician leadership development is that medicine is similar to other domains in many ways, which makes it appropriate for the sake of comparison and translatability. Physician leaders operate as decision-makers in large, complex, high-intensity environments with constrained budgets, changing team leadership and membership roles, and where people's lives are at stake, which are the same circumstances faced by leaders in the military, for example (Edler, Adamshick, Fanning, & Piro, 2010). As mentioned previously, Edler et al. (2010), like Taylor (2010), argue that the core leadership skills of proactive decision making, collaboration and cooperation, and planning and programme design are common among medical leaders and those in other domains. Many hospitals have similar identifiable roles to other organisations, are comprised of a wide range of employees who have different specialities, and are accountable to some extent to their clients.

Fifth, the nature of the medical profession makes it a good choice, since McKinn and Swanwick (2011) state that "healthcare professionals – doctors in particular – are among the most trusted members of society" (p. 182). This further increases the study's generalisability. Sixth, the NHS Institute for Innovation and Improvement and Academy of Royal Colleges (2010) claims that because of their legal duty as responsible for the care of patients, all physicians have an intrinsic leadership role within healthcare services. This duty of care extends even to medical doctor (MD) administrators, which is not removed when they give up their clinical practice in favour of administrative roles in hospitals. Although there is a good deal of literature and development programmes for nurses, doctors were selected for this study since doctors tend to be the final decision makers on healthcare teams involving doctors and other medical professionals, which is because physicians are the ones ultimately accountable for patient outcomes (McAlearney et al., 2005). Doctors are also more likely than nurses to be in senior leadership roles in hospitals, though there are exceptions.

Seventh, one of the strongest reasons for choosing medical leadership development is that the ultimate outcome of leadership development for physicians is clear: benefit to the patients (Kirkpatrick Level 4b). The justification for this claim was mentioned in the introduction. In some other professional domains, such as business, there is not a universally-accepted ultimate outcome. One example of this is that in the MULTI review, of 56 included studies, only three included Level 4b outcomes and two of them were healthcare leadership development programmes. Furthermore, if it can be shown that clinical outcomes (such as preventable patient deaths) can be positively impacted by development programmes, this would contribute to assuaging doubts about the yield of such programmes, in addition to adding knowledge to the field. Outcomes of this nature are likely to be considered intelligible to and, worthwhile by, many, especially when tied to economic benefits as well, as in the Jeon et al. (2013) study. As will be described further on, these clinical outcomes are data that is routinely collected by hospitals, therefore it is easy to incorporate them into studies and use the clinical outcomes of the intervention group as a comparison to those of other sites or national averages. Therefore, having physicians as the sample for the HEE SLR component of the PhD is beneficial because of the relative newness of leadership development programmes, the scale, importance, and relatability of the healthcare industry, and the central leadership role that doctors play.

4.1.2 Unique Features of Medical Leadership

In addition to the important commonalities, there are several differences between medical leadership and that in other domains, which make it interesting to study. The first, as mentioned previously, is that although doctors are highly educated, many physician leaders have never had any previous formal leadership training or education (B. Taylor, 2010). This is in stark contrast to the military, for example, where leadership development is a continual part of every officer's career. In fact, as stated earlier, in some ways, it is said that the medical training doctors receive is actually detrimental to the demands of leadership, particularly at the senior levels, as the former is partly characterised by inculcating autonomous decision-making and personal achievement (Stoller, 2009). Second, as mentioned previously, many doctor leaders are promoted or hired into leadership roles based largely on achievements other than leadership competence, such as research, teaching, or clinical performance (B. Taylor, 2010). Taken together, the implications of these factors mean that development programmes need to respect the education and professional experience that doctors have, while taking into account the possible massive range in participants' leadership experience, especially in terms of formal training. Third, senior doctor leaders do not typically get time off for leadership development, unlike military officers, for example, who are afforded time away from their professional roles for training. This puts extra emphasis on the medical leadership programmes to ensure that they are as efficient and effective as possible. Fourth, many doctor leaders even at the most senior administrative levels have pressing demands in non-leadership areas because of their role as clinicians, teachers, and researchers (B. Taylor, 2010).

Fifth, healthcare organisations are often very complex environments. Stoller (2009) describes them as characterised by various professional work forces and silos or fiefdoms, which Mintzberg (1998) describes as professional bureaucracies. Taylor (2010) suggests that doctors are dependent on hospital resources but are essentially private businesspeople responsible for the care of patients, which is largely administered by unionised, salaried healthcare professionals. Mintzberg (1998) says that this means that even when salaried by a hospital, doctors do not perceive themselves as reporting to hospital leadership. As a result, Bohmer (2012) suggests that hospital senior management often lack the positional power enjoyed by leaders in other settings, which restricts their ability to reprimand or fire a doctor for poor performance or offer financial rewards for excellent performance. This is in contrast to organisations where senior management, such as business executives, has veritable direct control over all their employees (Mintzberg, 1998). Similar to the seventh point above, a final defining feature of healthcare is that despite its challenging complexity, it has the luxury of a clear moral imperative and unifying purpose: patient well-being (Lee, 2010). This is unlike other industries where the agendas of various stakeholders may conflict and confuse an organisation's overall focus and moral compass (Bohmer, 2012). Taylor (2010) again reiterates that although leader behaviours may be similar to other domains, the environment of healthcare organisations differs in several ways.

The effect of these dynamics of the healthcare field on leadership is fourfold: it enhances the complexity of the leadership environment and results in situations of doctors leading other doctors and healthcare professionals without the relative simplicity of top-down management relationships. Second, as stated in background piece in the appendix on page 338, more doctors are moving into senior administrative roles, since the credibility that goes along with their clinical expertise helps avoid the "we/they" mentality, among other effects. Third, Mintzberg (1998) argues that in medicine, the structures and processes of performance control and improvement largely come from doctors, not the non-clinical organisational administrators. When doctors are at the most senior levels of an organisation, the effect can be symbiotic rather than oppositional, which is demonstrated in the introduction by the example of the correlation between physician CEOs and patient outcomes. In light of the nature of the healthcare field, effective and efficient leadership development is needed for physician leaders to function optimally. Finally, as stated throughout, both in terms of programme desired outcomes and post-programme evaluation, the Kirkpatrick Level 4b (benefit to patients) is of the utmost importance.

Having explained why physician leadership development was chosen as the focus for the HEE SLR component of this study, the focus now turns to its methodology. Again, the reason why this is located here in place of the overall PhD methodology is because this review is a culmination of the various background steps explained in chapter two and because it is the HEE SLR that generated the conclusions of the best available evidence, which is the core of this thesis's conclusions.

4.2 HEE SLR Research Design

As explained in chapter two, the design of the HEE SLR began with the decision to conduct a systematic literature review to collate the best available evidence in the academic literature and, through the process's transparent, reproducible, and scientific nature, minimise bias and strengthen the credibility of the findings and conclusions (S. Green et al., 2011; Husebø & Akerjordet, 2016; Liberati et al., 2009). Another initial step to provide rigour and guidance and minimise bias in the review process was to develop a research protocol (D. A. Cook & West, 2012; Lefebvre, Manheimer, & Glanville, 2011; Liberati et al., 2009; Steinert et al., 2012) by consensus among a team of 11 that was led by the author of this thesis. Assembling a professionally competent team with a diversity of perspectives is said to be one of the most important decisions in the review process, since it enhances the quality and generalisability of the review (D. A. Cook & West, 2012). For a full list of the team members, please see the appendix on page 348. At each stage of the research from this point until the conclusion of the HEE SLR, an HEE fellow for medical education served as a collaborating researcher. As mentioned in chapter two, one non-systematic and six extant systematic literature reviews (SLRs) on leadership development for doctors, along with MULTI on professionals in multiple domains, were analysed in depth to inform the design of the HEE study.

4.2.1 HEE SLR Research Question

As explained earlier, the first step in conducting a systematic review is deciding on a focused question (Cook & West, 2012; 'Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0', 2011).

The research question was: *what evidence exists in the academic literature of effective leadership development for doctors?*

The goal was to isolate the best available evidence in a way that was not done in any of the other reviews mentioned above.

"**Effective**" in this case was defined by the research team as studies that provided evidence of causation or correlation between interventions and improved performance outcomes at the individual, organisational, or clinical levels. The calibre of the evidence was naturally factored into the weight of the study's findings and conclusions, as will be described shortly.

"**Effective leadership development**" was understood as a complex concept that refers to a combination of programme elements, including the structure, length, and format, objectives/goals, content, developmental activities, and forms of measuring the reported impact of the interventions. These elements related to the question of optimal leadership development, as well as to measurement, particularly post-programme.

Once the research question was established, the Participants, Interventions, Comparison, Outcomes, and Study Design (PICOS) framework (Liberati et al., 2009; O'Connor, Green, & Higgins, 2008; Cook & West, 2012; 'Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0', 2011), as outlined in Table 4.1 below, was used to describe the study's component parts.

- The target **population** was physicians
- The **intervention** was leadership development interventions
- The **comparison** group was outcomes of intervention participants versus doctors or cluster sites who/that did not participate in the leadership development initiative in question, when included in studies' control groups

- The **outcomes** were analysed at the individual, organisational, and clinical levels according to the Kirkpatrick model that was described in chapter two
- Finally, all **study designs** were included, provided that the interventions were evaluated. Although the PRISMA guidelines suggest that some reviewers decide to exclude studies of high risk of bias (Liberati et al., 2009), given Husebø and Akerjordet's (2016) conclusion that all studies they identified were at a high risk of bias, it was felt that restricting the study to randomised trials or quasi experiments would have limited the sample size and omitted a large amount of useful data. Also, the absence of RCTs in the HEE review and the small number of experiments made including many designs a helpful decision. Including studies of low quality or high risk of bias can still be useful as long as their credibility and shortcomings are made clear. This was accomplished by publishing details of each study and the credibility evaluation instrument (MERSQI) scores for each aspect of each study, which will be described in more detail below. The findings and conclusions of these types of studies were used to reinforce or nuance the findings and conclusions of the best available evidence, as explained earlier.

Table 4.1

PICOS Framework

PICOS (Participants, Interventions, Comparison, Outcomes, Study Design)
P – Physicians
I – Leadership development programmes or interventions
C – When possible, compare outcomes to those of physicians who did not participate in leadership development
O – Impact on outcomes at the individual, organisational, and clinical levels
S – Qualitative, quantitative, and mixed methods designs were included

Above is a depiction of the HEE SLRs component parts according to the PICOS framework.

4.2.2 Literature Search

As with MULTI, the search strategy of scholarly literature was guided by two specialist librarians from the University of Cambridge, one from the Faculty of Education and the other from the Faculty of Medicine, a decision reinforced by Cook and West (2012). The search was conducted in the following electronic databases: Business Source Complete, ABI, ERIC, Pubmed/Medline, Embase, Scopus, and Web of Science, as well as the Cochrane Central Registry. For a full description of the search strategy, please see the appendix on page 349. Utilising multiple databases is necessary because the overlap between them is incomplete (D. A. Cook & West, 2012), which is especially likely in this study given the interdisciplinary nature of leadership development. Articles were limited to those published in English, as was the case with all six EMD SLRs, and in peer-reviewed academic journals in the period from 2007 to 2016.

The keywords used in all the searches were: “lead*” AND (“educat*” OR “develop*” OR “teach*” OR “taught” OR “train*”), each allowing for variations (eg “educating”). When it was possible to limit, the filter was set to adult human populations. The population was not specified beyond that because of the multitude of variations of synonyms of “doctor” (eg physician, resident, consultant, medical director, oncologist) used in article titles and key words.

Given the scope of this study, unpublished studies and the copious quantities of popular leadership literature were not included. Although common strategies (D. A. Cook & West, 2012), contacting individual researchers was not done, nor was including unpublished studies, since this was felt to detract from the replicability and transparency of the database search results.

The initial search yielded a provisional sample of 18,999 records, which was predictably large. As with MULTI, this enormous initial sample was predictable. Identical homograph issues arose, such as articles relating to lead, as in lead poisoning, and the colloquial use of the term. In this review, the number of non-relevant hits was increased because of the complications associated with specifying the population, as mentioned above.

To enhance objectivity and avoid mistakes, in line with the PRISMA guidelines, two researchers worked independently at each step of the research process (Liberati et al., 2009; Cook & West, 2012; ‘Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0’, 2011).

After the author of this thesis (JG) performed the search explained above, the second researcher (SA) performed the same search independently in a representative sample of databases and got identical results as the first researcher. Verifying by way of a representative sample is a measure of “good book keeping”, as recommended by Liberati et al. (2009) in the PRISMA guidelines. It is also said to enhance objectivity and avoid mistakes (Liberati et al., 2009).

4.2.3 Inclusion Criteria

The first step was to consult the titles for potentially relevant articles and identifying and removing duplicates (D. A. Cook & West, 2012). The PRISMA (2009) guidelines state that outlining the eligibility criteria is essential for appraising the validity, applicability, and comprehensiveness of a review. As with MULTI, given the number of hits, it was not feasible to record justification for each of the excluded articles. 599 articles appeared to be relevant and the second step involved reading abstracts or full texts to evaluate whether they met the inclusion criteria listed below (D. A. Cook & West, 2012). Cook and West (2012) reinforce the importance of clearly defining the inclusion and exclusion criteria, both conceptually and operationally. Verified by consensus of the initial research committee, studies between 2007 and 2016 inclusive were analysed as part of the review of empirical studies provided that their:

- Designs focused on leadership development programmes or **interventions** (eg coaching)
- Designs involved **evaluating** the effectiveness of the intervention or participants' leadership following a programme, rather than simply presenting a model or theory. For example, Ackerly et al. (2011) described a programme but had not yet collected any evaluative data, thus it was excluded. This is similar to the eligibility stipulation made in the Steinert et al. (2012) review, but this was not true surprisingly in the Rosenman et al. (2014) review. Studies with qualitative and/or quantitative methods were included, provided they met the other inclusion requirements, as explained in the previous chapter
- Sample group included **physicians** (although they need not have been exclusively physicians)
- Study focus was not on one individual task or capability, such as the paper by Gurrera et al. (2014), which featured a workshop to teach residents to make a business plan

Thus, studies focused on medical students were excluded on the basis of not being directly relevant to the current study's focus on qualified doctors. This is similar to the exclusion criteria that Frich et al. (2014) employed in their literature review. Studies were also excluded if leadership was only one of many learning outcomes. For example, Stergiopoulos et al. (2009) described a programme that taught eight different topics, of which leadership was only one, thus it was excluded. In the Rosenman et al. (2014) review, only ten per cent of their included studies identified leadership as the primary focus, which renders many of their studies not directly relevant to the purpose of this study.

Once again, the second researcher (SA) applied the above inclusion criteria to a representative sample of initial articles and the interrater agreement of the results was 100 per cent, which Cook and West (2012) suggest is a required measure in all cases at this stage.

Some studies not identified in the initial search were added after reviewing the bibliographies of relevant articles. The next step involved separating the empirical studies from those which included useful background information but would not be analysed in the same way as their empirical counterparts, including other systematic reviews (Liberati et al., 2009).

25 unique empirical studies met the inclusion criteria, after 206 relevant articles were excluded, and seven relevant reviews were identified, which formed the collection for the EMD SLR. No identical interventions or data sets were described in more than one study. To make this step as transparent as possible, Figure 4.1 below outlines the review process and results in a flow diagram ('Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0', 2011). This is a level of detail which Liberati et al. (2009) describe as rare, though optimal for readers to assess its comprehensives and completeness.

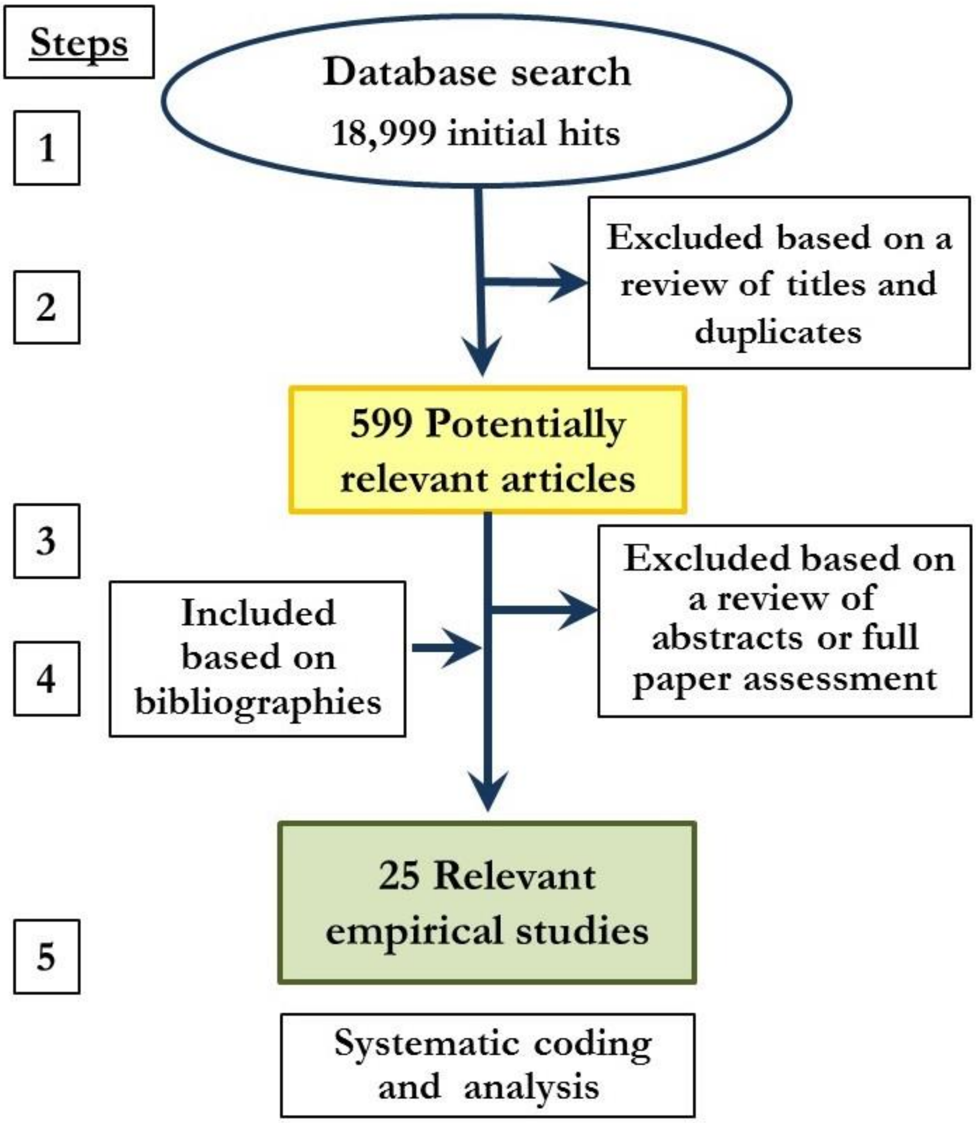

Steps
1
2
3
4
5
Database search
18,999 initial hits
Excluded based on a review of titles and duplicates
599 Potentially relevant articles
Included based on bibliographies
Excluded based on a review of abstracts or full paper assessment
25 Relevant empirical studies
Systematic coding and analysis

Figure 4.1 HEE SLR Systematic Article Search Process.

The figure above shows the process of the HEE literature search from the initial hits to the final three groupings of empirical studies, non-empirical background articles, and relevant literature reviews.

4.2.4 Coding

The 25 empirical studies that met the inclusion criteria mentioned above were then analysed extensively. The details of each study were recorded using structured data entry according to the codes displayed in Table 4.2 below, which are key features of the terms outlined in the PICOS framework, as well as study designs (D. A. Cook & West, 2012). This collection had been pilot-tested in the MULTI SLR and was approved by the initial review team before being applied to the included studies, while ensuring that there were no ambiguous definitions or other complications (D. A. Cook & West, 2012). There were four broad categories: study details, sample, programme, and measurements.

Table 4.2

Coding Structure

Coding	
Study details	Author name
	Publication year
	Purpose
	Whether they tested hypotheses
	Research questions
	Data collected: quantitative, qualitative, or both
	Methodology
	Methods and their details
Sample	Size
	Control group size (if applicable)
	Gender split percentage
	Mean age
	Profession/specialty (eg respirologist)
	Level of seniority (eg department head)
	Physicians-only or interdisciplinary
Programme	Selection criteria (eg nominated, applied and were selected)
	Location
	Faculty: internal, external, or mixed
	Number of sites
	Name of the programme
	Whether a needs assessment was undertaken
	Programme goals
	Whether they used a capability framework
	In-house or external
	Length and structure (eg six months with one day-long session every month)
	Topics addressed
	Developmental activities (eg coaching)
	Cost
Measurements	Outcome measures (eg Post-Programme Evaluations, promotions received following the programme)
	Response rate
	Reported outcomes
	Kirkpatrick measurement levels (1 – 4b)
	Outcome types (individual, organisational, economic, and patient safety/care)
	Raters (self, supervisor, peer, supervisees, facilitator, statistics)
	Type of data collected (subjective descriptions, self-reported numbers, and objective statistics)
	Times of data collection (pre, baseline, during, post, post-post)

Above is the collection of codes that were applied to the HEE included empirical studies as the initial data collection stage.

4.2.5 Data Analysis

The process of data analysis, the formation of the conclusions of the best available evidence, the conclusions explored, and implications for research and practice involved nine stages, which are presented in Figure 4.2 – Figure 4.9 below.

Before beginning the data analysis, all the articles were coded according to the items listed above and the frequency of each variable was tabulated (eg how many studies employed action learning as their methodology). For the reader's convenience, the codes collected for each study are presented in the appendix on pages 369 – 380.

Following this initial step, MERSQI, the instrument to evaluate the credibility of studies that was described in chapter two, was applied to each of the included studies, as depicted in Figure 4.2 below. The results are presented in the following chapter.

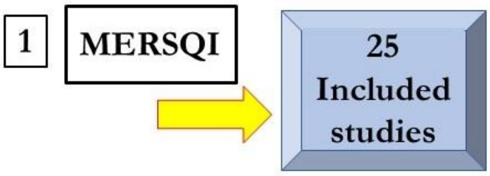


Figure 4.2 Data Analysis, Stage 1.

The figure above depicts the validated quality evaluation instrument, MERSQI, being applied to each of the included studies as the first measure of data analysis.

In addition to the overall MERSQI ratings, identifiable sets of characteristics emerged from the analysis at this stage that either strengthened the credibility of studies and the usefulness of their conclusions for the reader or had the opposite effect. The two researchers discussed both sets of characteristics and agreed on the final pairs. Thus, Stage 2 of the analysis involved synthesising these two sets of characteristics for the sake of the implications for research (see Figure 4.3 below).

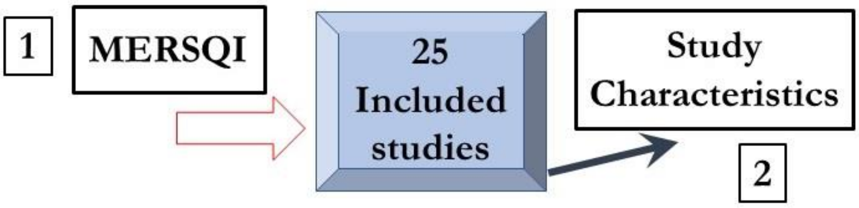


Figure 4.3 Data Analysis, Stage 2.

Above is a graphic outlining how the application of MERSQI to the HEE studies resulted in two sets of study characteristics for the sake of the implications for research: those that strengthened the credibility of studies and usefulness for the readers and those that lessened them.

4.2.6 A Unique and Defining Feature of This Study

As mentioned throughout, one of the main priorities for this review as part of the overall thesis was to produce tiered conclusions based on the credibility of evidence to potentially benefit research and practitioner communities alike. This was also intended to address the many calls for enhanced scientific rigour and reliability in the field (Frich et al., 2014; Husebø & Akerjordet, 2016; Rosenman et al., 2014). As will be described in the following chapter, the span of MERSQI scores from the lowest possible of 4.5/18 to 15 and the reality that the bulk of them were of low ratings is further reinforcement for the need for clarification regarding the evidence behind claims of what is known (D. A. Cook & West, 2012).

With this as the goal, data analysis Stage 3 involved devising five major categories into which to group the studies' credibility (see Figure 4.4 below). These categories are strong, good, moderate, limited, and anecdotal evidence. As will be explained below, no studies qualified as strong evidence.

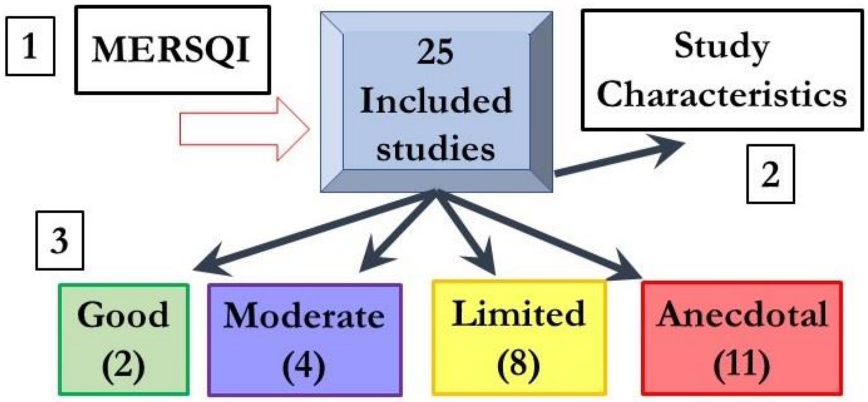


Figure 4.4 Data Analysis, Stage 3.

The diagram above depicts how the HEE included studies were grouped into four categories based on commonalities among their MERSQI quality scores for each aspect. The number of articles in each group is in parentheses below the title and they follow the same colour coding system that is used throughout this dissertation.

The debate between researchers in terms of how to define the groupings of the included studies was most earnest for this stage. The final decision was forged after careful consideration of the methodological characteristics and numerical score limits for each. The most degree of discussion related to the precise specifications of the most credible three categories (strong, good, and moderate evidence), including whether to specify a "strong" category even though no included studies met its criteria. In the end, it was decided that in the interests of promoting MERSQI's use in future studies and to detail the specifications required for the highest calibre research and consensus for all categories was eventually reached. The second researcher then tested them with a representative sample of articles to ensure the interrater consistency was 100 per cent. The description of the defining characteristics of each category is located in the following chapter.

4.2.7 Relationships Among Variables

Another key difference between the HEE review and the others included in this thesis is the statistical analysis of the relationship among variables. Husebø and Akerjordet (2016) explain that they felt it was impossible to conduct a meta-analysis due to the heterogeneity of the study designs and outcome measures, thus they elected to provide a narrative summary. They cite the Centre for Reviews and Dissemination (Centre for Reviews and Dissemination, 2016) to justify this choice and others similarly restricted themselves to descriptive analyses. While the small sample size of the HEE SLR was a challenge and precluded a meta-analysis, a recognition of the importance of analysing the relationships among the variables and making the analysis as transparent and credible as possible led to the decision to undertake the aforementioned next stage of analysis. It was anticipated that there would be higher margins and non-significant results; however, this step served the purposes of testing the usefulness of this approach to data analysis, as well as ensuring that the investigation of the relationship among variables was attempted in a credible way.

March, Sproull, and Tamuz (2003) defend the use of small samples and of organisations learning from them, particularly when obtaining large samples of identical occurrences is challenging, which is the case with leadership development interventions. The authors suggest that valuable learning can occur by aggregating similar incidents and analysing common features and implications (March et al., 2003). Likewise, Stevens (2012) asserts that using a small sample size is quite reasonable, as long as making a type I error will not have serious substantive consequences. The example he uses of this kind of consequence is concluding that a drug is safe when it might potentially be unsafe, which is a different kind of risk than conclusions regarding aspects of leadership development.

With these points in mind and to analyse the relationships among variables in a more credible way than had been done before, a series of linear regression analyses were performed to assess the bivariate correlation between all pairings of the following variables:

The **explanatory** variables (x axis) were:

- **MERSQI grouping** (good (n = 2), moderate (n = 4), limited (n = 8), anecdotal (n = 11) evidence)
- **Programme length:**
  (A: 1 week or shorter (n = 5), B: 1 month to 10 months (n = 9), C: 1 year (n = 6), D: >year (n = 4))
- **Kirkpatrick levels:**
  (Levels 1 – 3a only (not 3b, 4a, or 4b) (Y/N) (n = 7), level 3b (Y/N) (n = 15), level 4a (Y/N) (n = 5), level 4b (Y/N) (n = 6))
- **Developmental activities:**
  (Simulations (Y/N) (n = 9), 360s (Y/N) (n = 9), lectures (Y/N) (n = 8), action learning (Y/N) (n = 8), case study analysis (Y/N) (n = 7), coaching (Y/N) (n = 6))

The **dependent** variables (y axis) that we analysed were:

- The **type of data collected:**
  (Qualitative only (n = 8), quantitative only (n = 5), both (n = 12))
- **Methodology:** case study (n = 14)
- **Methods:**
  (Questionnaire (Y/N) (n = 20), interviews (Y/N) (n = 5))
- **Sample size** (number of participants, n = 22)
- **Physicians-only** sample (Y/N) (yes (n = 15); no (n = 10))
- **Selection criteria:**
  (Applied and were selected (n = 5), nominated (n = 7), volunteered (n = 6))
- **Faculty:**
  (Internal (n = 7), mixed (n = 8))
- **Location:**
  (In-house (n = 18), external (n = 7))
- **Programme length:**
  (A: 1 week or shorter (n = 5), B: 1 month to 10 months (n = 9), C: 1 year (n = 6), D: >year (n = 4))
- **Developmental activities:**
  (Simulations (Y/N) (n = 9), 360s (Y/N) (n = 9), lectures (Y/N) (n = 8), action learning (Y/N) (n = 8), case study analysis (Y/N) (n = 7), coaching (Y/N) (n = 6))
- **Kirkpatrick Outcome Levels:**
  (Levels 1 – 3a only (not 3b, 4a, or 4b) (Y/N) (n = 7), level 3b (Y/N) (n = 15), level 4a (Y/N) (n = 5), level 4b (Y/N) (n = 6))

For each pairing, the p value, R-Squared value, and whether the correlation is statistically significant at p = .05 are presented in Table 5.24, Table 5.25, and Table 5.26 in chapter five.

For methodologies, only case study was analysed since the sample sizes of the others were too small to make for an effective comparison. Sample sizes were also deemed too small for all the data collection methods other than questionnaires and interviews, mandatory selection criteria (n = 2), and external faculty (n = 2).

The level of seniority of the participants was considered as a variable to analyse, but they were either unspecified or too heterogeneous to make this useful, as outlined in the findings section. The topics used as the content for interventions were also considered, but they were far too numerous and diverse to make this feasible. This stage also unveiled variables that were not correlated with MERSQI ratings, beyond the actual instrument assessment criteria, which was already presented. Once again, both researchers discussed the choices and arrived at total agreement. Again, there was debate, especially in light of the small sample size, but the decision was made to move forward. The first author performed the linear regression calculations using the digital programme GraphPad and then both researchers discussed the results and their implications to ensure there was absolute agreement.

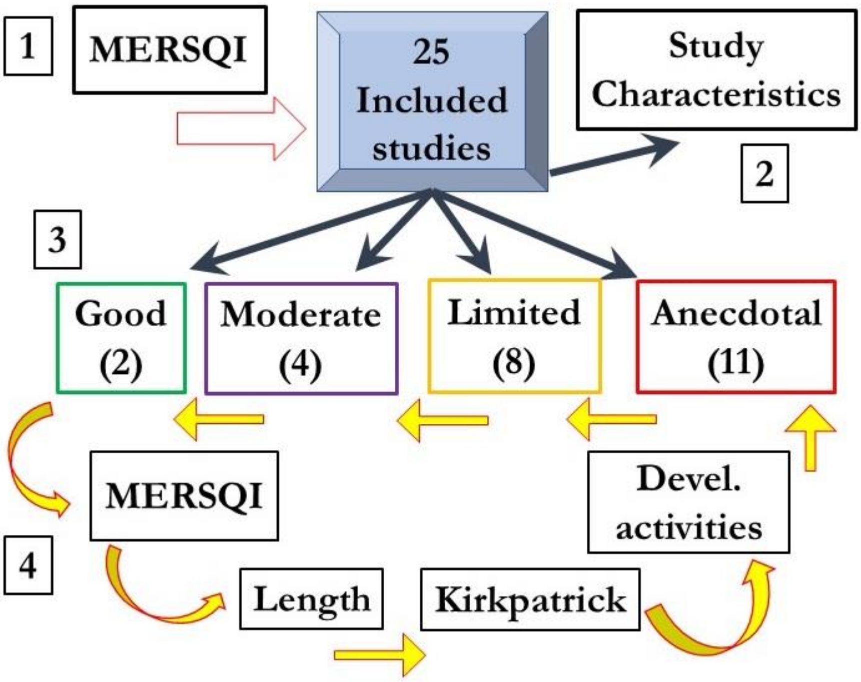


Figure 4.5 Data Analysis, Stage 4.

As seen above, aspects of the articles and the programmes they studied, along with the MERSQI groupings, were used as variables in bivariate linear regression analyses to investigate the relationship among variables.

Stage 5 involved synthesising the conclusions of the good and moderate evidence studies to clearly isolate the best available evidence. Points from the limited and anecdotal studies and the statistical analysis were added when they reinforced or nuanced conclusions in the more credible studies. Unless otherwise specified, a conclusion from the lower calibre studies or statistical analysis was not presented among the conclusions unless a better calibre study had reported the same thing. Both researchers discussed the conclusions meticulously until there was complete agreement.

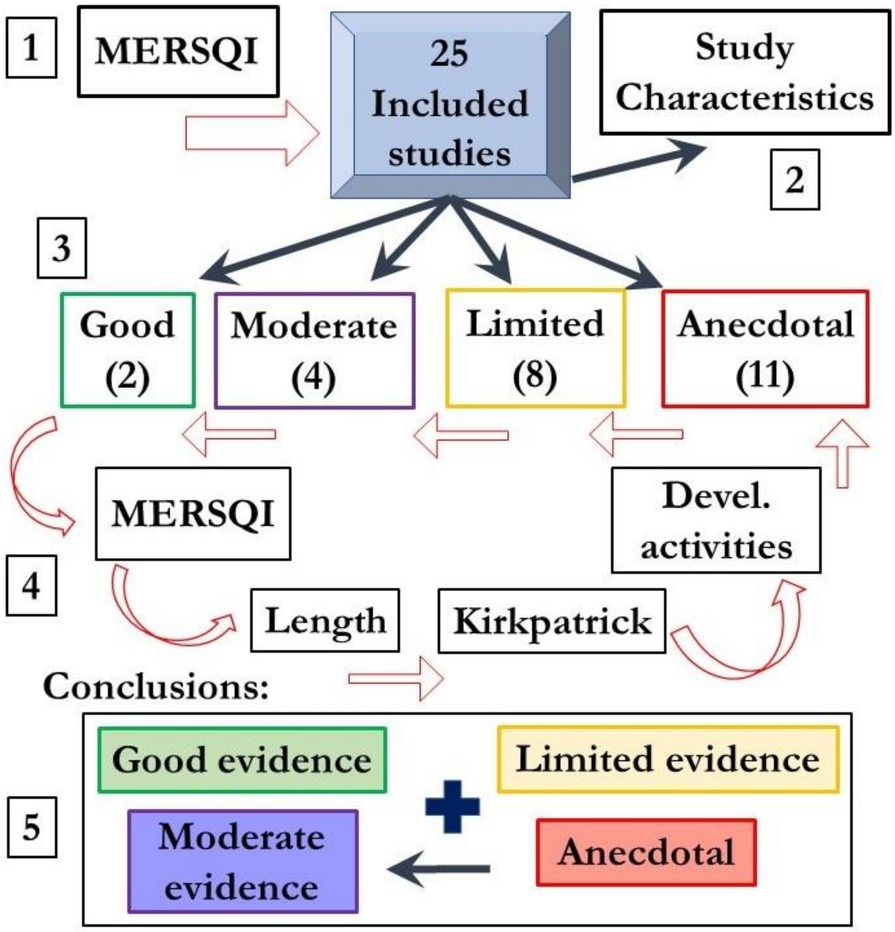


Figure 4.6 Data Analysis, Stage 5.

Above is a graphic outlining the basis for the conclusions of the best available evidence and how the lower calibre studies and statistical analysis findings were added to reinforce and nuance them.

As explained earlier and illustrated in Figure 4.7 below, Stage 6 involved exploring two conclusions that emerged from Stage 5 as worthy of further investigation. The included studies were then reviewed again through the lens of these two conclusions, a process which was later expanded to include the MULTI and EMD included articles. These have been explained in detail in chapters one and two and will not be repeated here.

As mentioned previously, the included studies were then reviewed again from these two perspectives and discussed until the interrater agreement was 100 per cent.

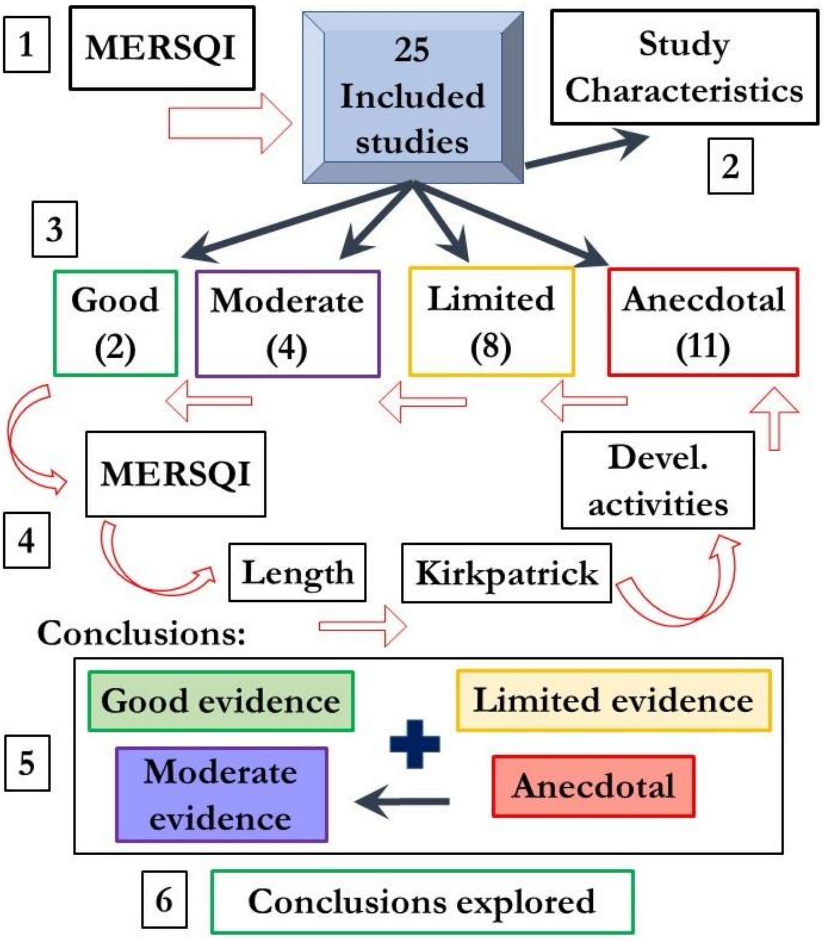


Figure 4.7 Data Analysis, Stage 6.

The figure above is a depiction of how two of the conclusions from the best available evidence were used as lenses to perform a deeper investigation of the included studies.

Stage 7 involved comparing the findings and conclusions from MULTI and the EMD SLR to the findings identified during each stage of the HEE SLR analysis. When the extant reviews reinforced the HEE conclusions, citations were added, and when there were notable differences between them, they were mentioned as well. The major theme from MULTI was also combined with the second conclusion explored from HEE, since they were both the same topic.

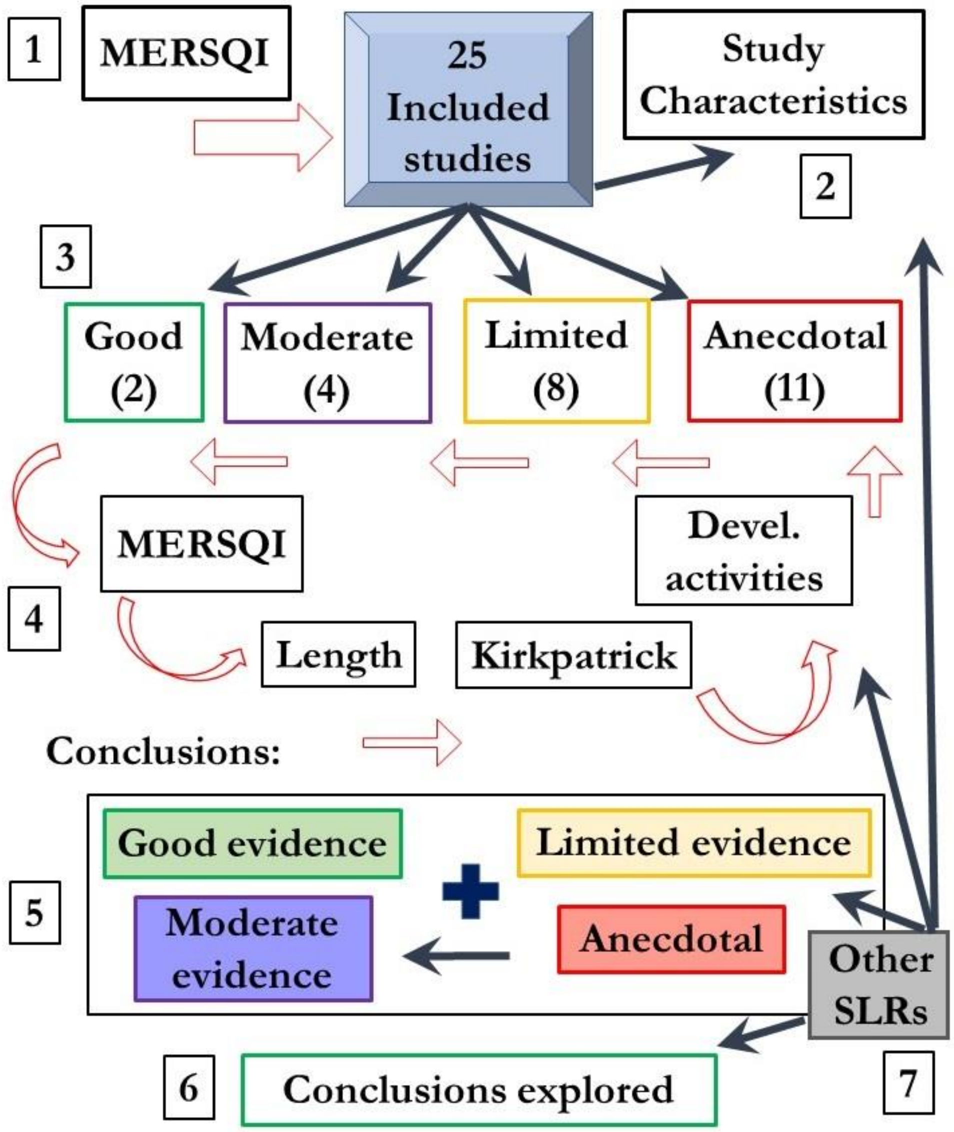


Figure 4.8 Data Analysis, Stage 7.

The figure above offers a visual of how the other data sources (MULTI, EMD, and the statistical analysis) were compared and contrasted to the conclusions from the best available evidence, statistical findings, and points for conducting effective research.

The final stage of the analysis was to combine all the previous steps to produce revised implications for research and practice and the discussion. A full depiction of the data analysis is included in the appendix on page 368. They are naturally based on the tiered rankings of the best available evidence by making it clear what is known and with what evidence. In their review, Steinert et al. (2012) echo the demand for this, stating that providers of leadership development should incorporate elements of programme design that have been said to be associated with positive outcomes into future programmes. Unfortunately, the authors give no indication of which of their own findings are based on more credible evidence than others.

This reinforces the need for the content and methodology of this study, while at the same time demonstrating the problem with the current state of literature and the need for the type of systematic and transparent approach featured in this study.

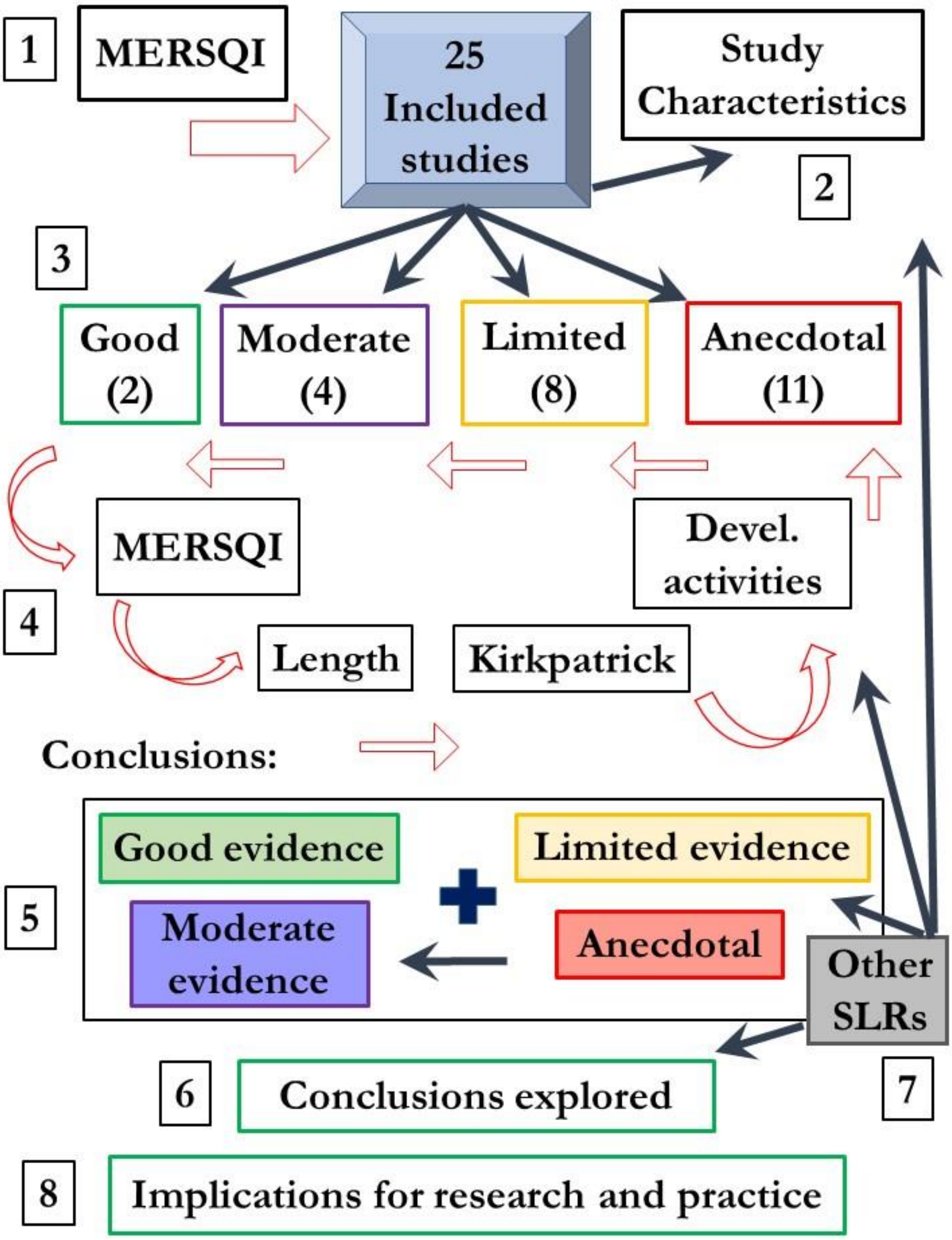


Figure 4.9 Data Analysis, Stage 8.

As depicted above, the final stage of the HEE methodology involved revising the implications for research in practice in a way that maintained the conclusions of the best available evidence at their core and included points from the other data sources as well.

4.2.8 Overall Evaluation Framework

The background reading of the non-empirical studies, MULTI, and EMD led to an initial evaluation framework for leadership development interventions. This involves certain considerations related to the design, data collection, and analysis, which all have the potential to influence the content and quality of the results of evaluation. This will be revisited in the discussion chapter as part of the theoretical model.

Design stage considerations

- Will organisational culture, including potential barriers to the application of leadership, be addressed (pre, during, and post)? If so, how?

Desired outcomes

- Which **stakeholders'** input will be factored in to the design of the programme?
- What overall outcomes do stakeholders want following the intervention (at the organisational, clinical, and economic levels), as well as those from participants (individual outcomes)?
- What are the **programme/developmental objectives**? (if different from the previous point)
- Will participants be allowed to personalise their goals and how they are evaluated? If so, how? Will examples of outcomes be provided or will just the categories be listed (eg clinical outcomes)? Examples can clarify what is meant for each category to avoid inappropriate or blank responses, but can also restrict responses to the set provided.
- In terms of **outcomes and impact**, which **Kirkpatrick levels** are being targeted? These can focus on assessing the quality of the programme (Level 1), on the individual-level in terms of knowledge and skills improvement (Levels 2a and 2b), individual-level behaviour change (Levels 3a and 3b), organisational level change (Level 4a), and benefit to clients or patients (Level 4b).
- Will **economic** outcomes be considered? These can include direct economic outcomes such as decreased spending in one's department and indirect outcomes such as the money saved by lowering the staff turnover rate compared to the cost of hiring new employees.

What is being evaluated

- How will the effectiveness of the programme overall be measured in terms of meeting its goals? For example, PPEs, attendance rates, graduation rates, discernible outcomes (Level 3b, 4a, and 4b).
- How will the **programme components** (such as length) and **developmental activities** (such as lectures) be assessed? Options include open-ended questions, such as asking participants to list any outstanding components, Likert-scale ratings, providing a list of all the components and asking for comments or ratings, specific questions for every component, and objective outcomes comparing two groups or at two different times.
- How will **developmental activities** be assessed? Will it be simply for quality control and perceived effectiveness, will it be according to their function, as described in chapter six of this thesis, or will it be tied to specific goals and outcomes? Further considerations can also relate to specific aspects of activities, such as (in reference to coaching), how many sessions or what lengths were considered optional?

Data collection instruments and methods

- Which data collection **methods** will be used and what weight will be given to each? For example, questionnaires, interviews, focus groups, observation etc
- Will they be formal, informal? Will they be structured, semi-structured, or open-ended?
- If questionnaires are being used, will they be personalised, standardised, or validated?
- At what **point(s)** will data be collected? Options include pre-programme, baseline, during the intervention, post, and post-post.
- To what will the data be compared to assess **relative** improvement? This can involve data sets collected at different times, such as baseline to post and post-post, or data contrasted to data collected on a control group, last year's statistical performance data, other sites' performance, or national averages.
- What will count as evidence of improvements being **sustained** through post-post measures?
- How will individual improvement (if there is any) be isolated and identified amidst team performance and other factors?
- How will causal relationships or correlations between the intervention and the outcomes be drawn that account for confounders, including other influences, such as other concomitant professional development?

Types of data collected

- What **evidence** and reports are considered indicative of successfully achieving the developmental objectives? These can involve Post-Programme Evaluations (PPEs), self-reports by descriptions and ratings, objective data (such as facilitator, supervisor, peer, or supervisee ratings), individual performance data, discernible outcomes (such as being promoted, opening a new office branch, or implementing one's action learning project), or statistics (such as lowering post-operative patient mortality by ten per cent over a six-month period).
- Will participants have an opportunity to describe alternative benefits or outcomes through an open-ended question?
- Will there be an opportunity for participants to offer open-ended feedback on the intervention more generally? Will they be encouraged to add further insights based on other experiences of leadership development programmes?
- How will constructive, critical, and outlying perspectives be solicited?

Evaluators

- Will the evaluators be internal or external?
- Will the **facilitators** be evaluating aspects of the programme as well? If so, how?

Analysis

- Will demographics such as age, gender, role, profession, and specialty be considered?
- How will the data be analysed? Will it statistical or descriptive?
- To what other sources will the data be compared? For example, the best available evidence from the published literature.

Practical/Logistics

- Will the evaluations be anonymous or not? Will they involve digital or paper copies? How will response rate be addressed?
- Is the evaluation feasible?
- Has survey or evaluation fatigue been considered?

Use of Findings

- How will the data be used after? Will it be published in academic journals or other written media, published online, disseminated internally, used only to refine programmes?

4.2.9 Ethical Considerations

The study was designed and conducted in accordance with the *BERA Ethical Guidelines for Educational Research* (2011).

The research did not involve primary research involving participants, but was restricted to existing published material. For this reason, no informed consent was necessary, as the information was already in the public domain. To minimise bias and maintain the ethic of respect for other authors, a validated assessment tool was used to evaluate the calibre of evidence in each study, along with publishing the raw data findings and a transparent analysis. This was also enhanced by adhering to the PRISMA statement (Liberati et al., 2009) and the Cochrane Handbook for Systematic Reviews (Version 5.1.0', 2011). The transparent inclusion criteria described earlier in the chapter was applied to all included studies, leaving the work free from prejudice based on any author of sample participant demographics such as age, race, or gender. Since the author of this thesis, nor the second researcher in the HEE SLR, did not include an empirical study of their own in the sample, there is no risk of compromise or bias in favour of their own programme(s), as can be the case with action research. In fact, the author's original literature review (MULTI) was critiqued alongside the others in a transparent manner. Furthermore, the thesis was not biased in favour of larger programmes or samples, which could possibly limit the included set in terms of location or institutional financial situation. The range of samples in the findings reflects this diversity. Although the research was conducted in collaboration with an HEE fellow, there was no pressure or influence from HEE that would in any way compromise the objectivity of the research at any stage.

The methods used for the study were selected not to produce favourable results, but after careful consideration of alternatives. This process and the justification of the final choices are described above. Several steps were taken to ensure that there is no suspicion of falsifying or distorting the results and to make the study amenable to external scrutiny, as the guidelines recommend (British Educational Research Association (BERA), 2011). These include presenting all the raw data in the body of the text or the appendix, making the analysis and connections among the findings, analysis, and conclusions very clear and transparent, and including a second, independent researcher. These measures were also intended to enhance the reliability, validity, and generalisability of the findings, analysis, and conclusions, as well as.

There are two final ethical considerations relevant to this study. The first is that the results of this study could appear to discredit an otherwise valuable leadership development programme if the article that described it received a low MERSQI rating. An attempt has been made to differentiate between the reported success of the interventions and the quality of the studies themselves. Furthermore, it has been echoed that the quality of research in the field, both in terms of by providers evaluating their own programmes and academics studying interventions, needs to improve. The consequences of not evaluating effectively have already been outlined. Also, as will be described in later chapters, it is believed that clear and transparent reporting of the calibre of published studies is a potentially valuable way to contribute to this improvement effort. Finally, by reinforcing the importance of isolating clinical outcomes as goals and metrics for leadership development programmes, it is possible that clinical options not included in this way could be overlooked. The evidence seems clear that outcomes-based programmes are more effective and it is entrusted to the professional discretion of the healthcare professionals to ensure that no important clinical priorities suffer as a result of striving to improve others.

Therefore, this study has followed the BERA guidelines, chiefly through its transparent methodology, reporting, and analysis.

5 Chapter Five: Findings: MERSQI, Raw Data, and Statistical Analysis

This chapter examines the findings of the analysis of all three reviews (MULTI, EMD, and HEE). It begins with the application of MERSQI instrument to the HEE included studies and then explains the hierarchical groupings according to the studies' calibre, which were used to guide this study's analysis and conclusions. The raw data from the literature reviews are then presented section-by-section in terms of the study designs and the sample, programme, measurements, and outcome details. For the HEE SLR, the analysis of the data is further separated based on the calibre of the studies. Within each section of the raw data findings, the data from the different reviews are compared and discussed, particularly in reference to the good and moderate HEE studies. The final section of the chapter describes the results of the statistical analysis applied to the variables in the HEE SLR. With this, the attention turns to the HEE MERSQI ratings.

5.1 HEE MERSQI Score Ratings

Before discussing the MERSQI ratings for the HEE included studies, it is helpful to outline the features of a study that would receive a MERSQI perfect rating of 18, which are included in Table 5.1 below.

Table 5.1

Features of Earning a MERSQI 18 Score

Features of MERSQI 18 Rating Studies	
✓	A randomised control trial at ≥2 institutions
✓	A response rate of ≥75%
✓	Objective data collected (not only self-assessments)
✓	The internal structure validity, content validity, and criterion validity are reported for evaluation instruments
✓	Appropriate data analysis beyond just descriptive analysis
✓	Outcome measures include Level 4b benefit to patients

Above is a summary of the elements of study designs that would earn a study a perfect MERSQI score. Table 5.2 and Table 5.3 below present the colour-coded MERSQI tabulations for the 25 HEE SLR included studies. The background colour for each total score reflects the four categories of evidence that are described below. The range of scores is from 4.5 to 15 with a mean of 9.94 and a standard deviation of 2.74. When not rounded up, this places the *mean* in the anecdotal calibre evidence category, which is the lowest of the five groupings. This is a good indication that the aforementioned authors' lamentations about the poor calibre of work in the field is accurate. The range of scores is similar to those in the Rosenman et al. (2014) review, which, in terms of the studies focusing primarily on leadership, ranged from 6.5 – 14.5 with a mean of 11.4 and a standard deviation of 2.9. As in the HEE study, the highest score in the Rosenman et al. review (at 16.5) was a short team-based intervention.

The findings that follow are listed in the order in which they were presented in chapter two as part of the explanation of the MERSQI instrument.

In terms of **study design**, not one study in HEE used a randomised controlled trial (RCT). This contrast to some of the extant reviews' findings may be partially attributable to the frequent number of RCTs in the EMD studies that were short, task-oriented interventions aimed at improving teamwork, rather than longer programmes or ones whose goals targeted broader, softer leadership skills. Abrell et al. (2011) suggest that more complex skills take longer to develop. This might indicate that designing RCTs for leadership interventions can be challenging, though as will be discussed further on, it is nevertheless possible. Four HEE studies featured nonrandomised, two-group (NR2GP) designs including a control group; seven employed single group pre and posttest (SGPP); and the remaining 14 were single group cross-sectional or posttest only (SGCS). This is reminiscent of Hartley and Benington's (2010) claim that a good deal of leadership development research employs cross-sectional designs, which precludes establishing correlated or causal links between interventions and outcomes or ruling out alternative explanations. This imbalance in favour of less credible designs reflects an overall weakness in the bulk of the field's literature and confirms the need for better calibre research.

In terms of **sampling**, 24 of 25 studies featured single-institution interventions, with only that by Ten Have et al. (2013) studying the same intervention at more than one centre (four, in this case). Nine of the studies had a response rate of 75 per cent or higher; eight were in the 50 – 75 per cent range; and eight did not report their response rate.

For **data collection**, slightly more than half the studies (n = 17) collected objective data, with eight relying on subjective data, which will be discussed in more detail further on.

In terms of **validity of evaluation instruments' scores**, nine studies reported the validity of their evaluation instruments and 16 did not. Surprisingly, only 11 studies reported the content of their instruments in full, while 14 did not. This omission makes it challenging to ascertain what exactly was asked and whether all the data are presented or merely highlights. Other studies were biased in favour of the programme/intervention by asking for only positive benefits, such as that by Butler, Forbes, and Johnson (2008). Similarly, only seven studies analysed the relationships among variables and 18 did not. As mentioned in the EMD section, given that leadership development is a complex phenomenon, it is important to investigate aspects in connection to each other, not in isolation. This further reinforces the choice in this thesis to investigate the relationship among variables in a comprehensive and transparent way.

In terms of **data analysis**, only six studies used appropriate analysis to adequately answer their research questions and defend their conclusions, whereas 19 were considered inappropriate. The high degree of subjective-only data is partly attributable to this assessment. Only 13 studies (52%) went beyond a purely descriptive analysis, whereas 12 studies relied entirely on a non-scientific, descriptive analysis.

In terms of **outcomes**, studies were rated according to the highest Kirkpatrick level outcomes that they reported. For example, a study whose participants claimed to have achieved outcomes at Level 1, 2a, 2b, 3a, and 4b would be given the highest score (3) and the same score as a study that reported only 4b outcomes. Also, the fact that the instrument does not allow for 4a outcomes will be discussed in chapter seven. Of the included studies, only six reported outcomes at the 4b level and, as will be mentioned further on, of those, many were not reinforced by objective data. Given the definition of leadership used in this study and the application focus of leadership development, it is important to get beyond just individual development, though few studies did, unfortunately. 13 studies reported up to, and including, Level 3 behaviour outcomes, but the instrument does not distinguish between subjective and objective data (Level 3a from 3b). Five studies were restricted to Level 2b knowledge and skills acquisition outcomes, and one included only Level 1 and 2a outcomes. The breakdown of reported outcomes categorised by the Kirkpatrick model will be described in more detail further on.

The tally of the final scores will also be discussed in the following section concerning the MERSQI score groupings.

Table 5.2

MERSQI Applied to the 25 Included Studies (1/2)

Author	Study design	Sampling			Validity of Evaluation Instruments			Data Analysis			
		No. of Institutions	Response rate	Type of data	Internal structure	Content	Relationships to other variables	Appropriateness of analysis	Sophistication of analysis	Outcome level	Total score
Hemmer (2007)	SGPP	1	NR	Objective	NR	NR	NR	Inappropriate	Descriptive only	2b	7.5
Korschun (2007)	SGCS/P	1	79%	Self-reported	NR	R	NR	Inappropriate	Beyond descriptive	4b	10
Miller (2007)	SGCS/P	1	66%	Self-reported	R	R	R	Appropriate	Beyond descriptive	3	11.5
Daniels (2008)	NR2GP	1	71%	Objective	R	R	R	Appropriate	Beyond descriptive	3	14.5
Bergman (2009)	SGPP	1	74%	Self-reported	R	R	R	Inappropriate	Descriptive only	3	10
Edmonstone (2009)	SGCS/P	1	57%	Self-reported	NR	NR	NR	Inappropriate	Descriptive only	3	7
Malling (2009)	NR2GP	1	77%	Objective	R	NR	NR	Inappropriate	Beyond descriptive	3	12
Murdock (2009)	SGPP	1	NR	Objective	NR	NR	NR	Inappropriate	Descriptive only	3	8
Cherry (2010)	SGCS/P	1	NR	Objective	NR	NR	NR	Inappropriate	Descriptive only	2b	7
Day (2010)	NR2GP	1	53%	Objective	R	R	NR	Inappropriate	Beyond descriptive	3	12.5
Kuo (2010)	SGCS/P	1	94%	Objective	R	R	NR	Inappropriate	Beyond descriptive	3	12
Edmonstone (2011)	SGPP	1	NR	Objective	R	NR	NR	Inappropriate	Beyond descriptive	4b	11
Sanfey (2011)	SGPP	1	50%	Objective	NR	NR	R	Inappropriate	Beyond descriptive	3	11
Bearman (2012)	SGCS/P	1	92%	Self-reported	NR	R	NR	Inappropriate	Descriptive only	2b	7.5
Shah (2013)	SGCS/P	1	NR	Self-reported	NR	NR	NR	Inappropriate	Beyond descriptive	2b	6

Note. RCT = Randomised controlled trial. NR2GP = Non-randomised, two groups. SGPP = Single group, pre and post test. SGCS/P = Single group, cross-sectional or posttest only. R = Reported. NR = Not reported. Green = most credible. Purple = second most credible. Yellow = third most credible. Red = least credible result, including if relevant variables were not reported.

Table 5.3

MERSQI Applied to the 25 Included Studies (2/2)

Author	Study design	Sampling			Validity of Evaluation Instruments			Data Analysis			
		No. of Institutions	Response rate	Type of data	Internal structure	Content	Relationships to other variables	Appropriateness of analysis	Sophistication of analysis	Outcome level	Total score
Ten Have (2013)	NR2GP	4	100%	Objective	R	R	NR	Appropriate	Beyond descriptive	3	15
Vimr (2013)	SGCS/P	1	NR	Objective	R	NR	NR	Inappropriate	Descriptive only	4b	9.5
Blumenthal (2014)	SGCS/P	1	100%	Self-reported	NR	R	NR	Appropriate	Descriptive only	2b	8.5
Dickey (2014)	SGCS/P	1	NR	Self-reported	NR	NR	NR	Inappropriate	Descriptive only	1/2a	4.5
MacPhail (2014)	SGCS/P	1	70%	Objective	NR	R	R	Inappropriate	Beyond descriptive	3	11.5
Satiani (2014)	SGCS/P	1	NR	Objective	NR	NR	NR	Inappropriate	Descriptive only	3	7.5
Nakanjako (2015)	SGCS/P	1	100%	Self-reported	NR	R	R	Inappropriate	Beyond descriptive	4b	11
Patel (2015)	SGPP	1	77%	Objective	NR	R	R	Appropriate	Descriptive only	4b	13.5
Fernandez (2016)	SGCS/P	1	60%	Objective	NR	NR	NR	Appropriate	Beyond descriptive	4b	11.5
Pradarelli (2016)	SGPP	1	100%	Self-reported	NR	R	R	Inappropriate	Descriptive only	3	9.5

Note. RCT = Randomised controlled trial. NR2GP = Non-randomised, two groups. SGPP = Single group, pre and post test. SGCS/P = Single group, cross-sectional or posttest only. R = Reported. NR = Not reported. Green = most credible. Purple = second most credible. Yellow = third most credible. Red = least credible result, including if relevant variables were not reported.

Above is depicted the colour-coded MERSQI score for each aspect of each of the included studies. As mentioned previously, the order of colours according to study quality, from highest to lowest, is green, purple, yellow, red.

5.2 MERSQI Groupings: A Unique and Defining Feature of This Study

As mentioned previously, in the interests of providing clear and transparent analysis and tiered conclusions according to the credibility of studies, commonalities in the designs of the HEE studies fit appropriately in five hierarchical categories. These categories were devised and solidified after discussion and debate between the author of this dissertation and the collaborating researcher. These categories are: strong, good, moderate, limited, and anecdotal evidence.

The highest category is **strong evidence** and is characterised by randomised controlled trials within the MERSQI score range of 15.5 – 18/18. As mentioned previously, there was none of this calibre identified in the HEE review. Experiments can be challenging to orchestrate due to the direct, precise, and systematic control researchers need to have in order to conduct them (Yin, 2003). Obtaining control groups can also be difficult, especially for higher seniority level samples (Collins & Holton III, 2004). Despite these two points, leadership development experiments are nevertheless possible, as evidenced through studies described in the EMD SLR.

The next category is **good evidence** and is characterised by having established a correlation between leadership interventions and objective outcome data using pre and post-post test measures and a control group. Good evidence studies required a MERSQI score of 14 – 15.5; they are represented by the colour green; and there were two studies of this calibre, which are Dannels et al. (2008) and Ten Have, Nap, and Tulleken (2013).

The next category is **moderate evidence** and is characterised by having established a correlation between interventions and objective outcome data but were limited by incomplete reporting or other gaps in the study. For example, in the Malling et al. (2009) study, the authors report that a leadership intervention produced no improvement in participants' leadership skills measured by a Multi-Source Feedback instrument compared to a balanced control group. In and of itself, this is not a problem; however, the details of the actual intervention are scant and the evaluation of the programme is not thorough enough to ascertain to what extent the design and delivery of the programme was responsible for the lack of improved performance, or if this can be attributed to another factor such as the organisational culture. Similarly, the study by Day et al. (2010) compared the curricula vitae (CVs) of seven years' worth of orthopaedic surgeons who had undertaken a mentorship intervention to the CVs of a control group who applied to the same programme and were not accepted. The outcomes included leadership role and research publications, among others, before the programme and using a post-post test. Although the authors report that the increase in the post-post academic rank was 48 per cent in the experiment group compared to 21 per cent in the control group, like the Malling et al. study, the details of the actual intervention are almost entirely absent, which limits the usefulness of the findings. In another example, the Kuo et al (2010) study provided useful statistics of post-programme outcomes, but left unhelpful gaps in the sample details and the longitudinal projects, which were a key part of the programme. The authors also neglected to include any mention of programme failings or outlying perspectives, which typically strengthen the discussion (Alvesson & Spicer, 2012). Finally, the Patel et al. (2015) study provided good examples of action learning projects for quality improvement and support interventions that can reinforce them; however, they omitted many details of the data collection, making it unclear what exactly was asked and with what consistency the responses were. Also, only positive outcomes are reported. Collectively, these reporting flaws detract from the completeness of the four studies' conclusions. Moderate evidence studies required a MERSQI score of 12 – 13.5; they are represented by the colour purple; and there are four studies identified of this calibre, which were all referenced above.

The next category is **limited evidence** and is characterised by being based purely on participants' perceptions and by either a lack of objective data to reinforce those perceptions, or by other major gaps in the study. For example, participants in the Sanfey et al. (2011) study reported that many leadership skills were enhanced in the short term, but this was not verified by other raters or objective data. The Korschun et al. (2010) study did not fully describe their data collection instruments and many of the data collected are not reported completely. In another example, Nakanjako et al. (2015) did not appear to investigate which programme elements were effective. MacPhail et al. (2015) used retention and promotions as outcome measures, but did not compare the intervention participants to a control group or national averages, which diminishes their usefulness. Limited evidence studies required a MERSQI score of 10 – 11.5; they are represented by the colour yellow; and there were eight studies identified of this calibre.

The next category is **anecdotal evidence**, which unfortunately included the most studies and the overall mean (9.94), and is characterised by being based purely on the authors' perceptions, or by being plagued by other major reporting issues or gaps in the study. These tended to result from omitting key details of the sample, programme, data collection instruments, or data collected. The result are snapshots of reported post-programme benefits or outcomes or elements of effective programmes from the authors' perspectives without any data to substantiate them. Anecdotal studies required a MERSQI score of 4.5 – 9.5; they are represented by the colour red; and there were 11 studies identified of this calibre.

For a summary of the MERSQI score evidence category details, see Table 5.4 below.

Table 5.4

MERSQI Groupings

Evidence	Characteristics	MERSQI scores	n	Studies and MERSQI score
Strong	Randomised controlled trial	15.5 – 18	0	N/A
Good	Correlated objective outcome data Pre and post-post measures Control group	14 - 15.5	2	Dannels 2008 (14.5) Ten Have 2013 (15)
Moderate	Correlated objective outcome data Incomplete reporting or gaps in the study	12 - 13.5	4	Malling 2009 (12) Day 2010 (12.5) Kuo 2010 (12) Patel 2015 (13.5)
Limited	Based purely on participants' perceptions Or other major gaps in the study	10 - 11.5	8	Korschun 2007 (10) Miller 2007 (11.5) Bergman 2009 (10) Edmonstone 2011 (11) Sanfey 2011 (11) MacPhail 2014 (11.5) Nakanjako 2015 (11) Fernandez 2016 (11.5)
Anecdotal	Based purely on the authors' perceptions Or other major gaps in the study	<10	11	Hemmer 2007 (7.5) Edmonstone 2009 (7) Murdock 2009 (8) Cherry 2010 (7) Bearman 2012 (7.5) Shah 2013 (5) Vimr 2013 (9.5) Dickey 2014 (4.5) Satiani 2014 (7.5) Blumenthal 2015 (8.5) Pradarelli 2016 (9.5)

Note. Green = most credible. Purple = second most credible. Yellow = third most credible Red = least credible result, including if relevant variables were not reported.

Above is a depiction of the five MERSQI calibre groupings in terms of their name, defining characteristics, MERSQI score range, colour, n value, and the studies that qualified for each.

5.3 MERSQI Category Examples

To illustrate how the MERSQI instrument was applied to the included studies, how the score weighting for each MERSQI component contributes to the overall score, and how these results led to creation of the category groupings, four examples are provided below. These are presented in Table 5.5 and then described.

Table 5.5

MERSQI Category Examples

				Sampling			Validity of Evaluation Instruments			Data Analysis			
#	**First author**	**Grouping**	**Study design**	**Institutions**	**Response rate**	**Type of data**	**Internal structure**	**Content**	**Relationships to other variables**	**Appropriateness of analysis**	**Sophistication of analysis**	**Outcome level**	**Total score**
1	Ten Have	Good	Balanced control group	4	100%	Objective	Reported and used a validated instrument	Reported	Not reported	Appropriate	Beyond descriptive analysis	3 (4b implied)	15
2	Patel	Moderate	Single group, pre and post test	1	≥77%	Objective	Reported and used a validated instrument	Not reported	Not reported	Appropriate	Beyond descriptive analysis	4b	13.5
3	Sanfey	Limited	Single group, pre and post test	1	50-74%	Objective	Not reported	Not reported	Reported	Inappropriate	Beyond descriptive analysis	3	11
4	Shah	Anecdotal	Single group, cross-sectional or post-test only	1	N/A	Self-reported data	Not reported	Not reported	Not reported	Inappropriate	Descriptive analysis only	2b	5
			Randomised Control Trial	≤2	≥75%	Objective	Reported	Reported	Reported	Appropriate	Beyond descriptive analysis	4b	≥13
			Non-randomised, two groups	1	50-74%	Self-reported data	Not reported	Not reported	Not reported	Inappropriate	Descriptive analysis only	3	12-12.5
			Single group, pre and post-test		N/A							2b	10-11.5
			Single group, cross-sectional or post-test only									1/2a	≤9.5

5.3.1 Good Evidence Study: Ten Have et al. (2013), MERSQI Rating: 15

The Ten Have et al. (2013) featured a one-day intervention for intensivist trainees centred on improving observable behaviours during medical interdisciplinary ICU rounds. This is a patient-centered communication session designed to integrate care delivered by specialists from different disciplines. Participants were video-taped during four real-life, progressively complex IDR scenarios concerning formulating a patient plan of care in conflicting situations. Participants were given peer and expert feedback after each simulation, as well as six weeks later based on a new videotaped and analyzed IDR.

The **study design** is a quasi-experiment with a balanced control group comprised of experienced intensivists from the ICUs who did not participate in the leadership intervention.

The intervention was run at four separate **institutions**, which enhances the generalisability and external validity of the results.

Data was collected on all participants, so the **response rate** was 100 per cent, which maximises the representative nature of the results for the given sample for this sample.

The authors collected two forms of **objective data**: peer and trained facilitator feedback based on participants' performance during the ICU rounds. The raters evaluated participants' plan of care and the process by which it was agreed and understood by others on the team, as well as how it was delivered. The assessment involved applying a **validated instrument** based on quality indicators of the plan of care. The authors also tested for inter-rater agreement.

The details of the **data collection** are reported in full, which enhances the validity and reliability of the study.

The **relationships to other variables** are not reported, since there is no indication that the intervention itself was evaluated to highlight which aspects of the programme were effective or not and in which ways.

The **analysis** was considered appropriate, given the specific goal of the intervention (to increase observable leadership behaviours and produce a plan of care). Combining a pre test with a post and a post-post measure is helpful to demonstrate improvement and whether it was sustained. While the analysis could have been extended further, such as by evaluating the intervention itself and adding qualitative data, the analysis was dependable for its purpose.

The **sophistication** of the analysis was beyond descriptive since there were several forms of data to reinforce the findings and conclusions.

Finally, the **outcome level** was 3 (observable behaviour changes), but considering that the intervention featured a patient-centred communication session and that the ultimate outcome was a plan of care, it would have qualified as a level 4b outcome, had it been stated.

It should be noted before continuing that although this study provides much useful information, there are areas in which it could be improved. For example, no information regarding participants' ratings of the intervention was provided to illuminate which aspects of the programme and in which ways they were effective or not. Second, behaviour was measured on a three-point scale, which does not allow for a distinction between satisfactory and exemplary behaviour. Third, no clinical outcome data was presented to explicate the relationship between improved physician behaviour and patient outcomes.

Overall, this was a well-designed study that involved dependable data collection and reliable results.

5.3.2 Moderate Evidence Study: Patel et al. (2015), MERSQI Rating: 13.5

The Patel et al. (2015) study featured a two-year leadership training intervention for residents focused on quality improvement (QI) and patient safety (PS). Expert faculty used a validated instrument to rate participants' performance during clinical simulations, which resulted in a three to four point increase mean rating on a scale of 15 following the intervention.

Participants completed post-programme questionnaires and reported that the intervention increased their ability to lead QI/PS activities in the future, as well as their motivation to pursue leadership positions. Many implemented their action learning projects, which allegedly directly benefited their patients.

The **study design** was a single group, pre and posttest with no control group, which limits the results somewhat.

The intervention was run at a single **institution**, which limits the generalisability of the results.

The **response rate** was 77 per cent, which is generally quite high, but leaves one to question the views of those who did not respond.

The authors collected two forms of **objective data**: facilitator feedback based on clinical simulations and the implementation of an action learning QI/PS project. The former assessment was done according to a **validated instrument**.

Many of the details of the **data** that was collected were omitted, which erodes the transparency of the study and the representative nature of the results.

The **relationships to other variables** are not reported, since the components of the intervention did not appear to have been evaluated to highlight which aspects of the programme were effective or not and in which ways.

The **analysis** was considered appropriate, since combining the facilitator ratings pre and post with the implementation of the action learning projects and self-ratings provides a comprehensive analysis.

The **sophistication** of the analysis was beyond descriptive since the two forms of objective data reinforced the self-ratings and, by extension, the findings and conclusions.

Finally, the **outcome level** was 4b, as demonstrated primarily by the implementation of the action learning projects, a discernible outcome.

Considering how comprehensive the intervention in this study was, it is unfortunate that its component parts were not analysed. The lack of a control group lessened the relative nature of the participants' improvement; however, there are good indications that progress was made and benefitted patients directly.

5.3.3 Limited Evidence Study: Sanfey et al. (2011), MERSQI Rating: 11

The Sanfey et al. (2011) study featured a ten-week intervention for doctors, academics, and medical staff at an academic medical centre.

The evaluation featured pre-programme self-ratings of participants' leadership skills and aspirations to lead followed by a posttest that asked if their skills had improved, if their leadership behaviours had increased, and whether they had received promotions following the programme. The post-post test also asked participants to rate the most important leadership skills and attributes of a leader, whether they had made changes in their leadership behaviours and professional lives after the programme, and whether there were any additional benefits from having participated.

The **study design** was a single group, pre and posttest with no control group.

The intervention was run at a single **institution**.

The **response rate** was 50 per cent, which barely merited a MERSQI point (no points are awarded for less than 50 per cent).

The only **objective data** that was collected was self-reports of promotions, which is a discernible outcome.

All the responses were self-reports and no **validated instruments** were used.

Many details of the **data** that was collected were omitted.

The **relationships to other variables** are reported somewhat through the PPEs, since participants were asked to comment on the programme in the post-post test. This could have been more specific and in-depth to add further insights into effective components.

The **analysis** was considered not appropriate, since the study relied sheerly on self-ratings without external raters to reinforce the non-discernible outcomes or a control group to demarcate the relative improvement.

The **sophistication** of the analysis was beyond descriptive since the discernible outcomes reinforced the self-ratings.

Finally, the **outcome level** was 3, restricted to the individual-level.

This study features a programme that was extremely well-rated by participants; however, since the data collection relies exclusively on self-reports and there are details missing in the reporting, the value of the findings is limited.

5.3.4 Anecdotal Evidence Study: Shah et al. (2013), MERSQI Rating: 5

Similar to the Patel et al (2015) study, the intervention in the Shah et al. (2013) study sought to improve patient safety, in this case through a two-day intervention for ophthalmologists. Researchers analysed participants' insights compared to the themes of the programme's conceptual framework. Participants also completed some form of a PPE, but it was unclear whether it was by questionnaire or interview. From these, the authors offer instructional effects of the programme.

The **study design** was a single group, cross-sectional with no control group.

The intervention was run at a single **institution**.

The **response rate** was not reported and therefore no points were awarded, since it is unclear to what extent the responses are representative.

No **objective data** was reported having been collected.

Not only were the **instruments** not described in any detail, but it was unclear which data collection methods were used.

Details of the **data** that was collected were sparse and only summarised by the authors.

The **relationships to other variables** are not reported.

The **analysis** was considered not appropriate, since so much of the data collection methods and data collected are missing. The cross-sectional design is also limiting with no baseline or post-post measures to confirm relative and sustained improvement.

The **sophistication** of the analysis was descriptive analysis only.

Finally, the **outcome level** was 2b, restricted to the individual-level only and based only on the authors' perceptions.

Although the intervention seems to have been well received and its focus on improving patient safety is valuable, the study is so eroded by flaws that the findings are of very limited value. There is also no indication of application of learning except for isolated qualitative quotes to support the authors' assertions, but their representative credibility is questionable.

5.3.5 Conclusions of the Weighting of the MERSQI Scores

Combining the assessments of the four studies together, there are several factors that significantly affected their ratings and credibility. Increasing the generalisability by including a control group, more than two institutions, and a high response rate added up to three more points than lower calibre studies. The most weighted factors in the MERSQI instrument are collecting objective data, which adds two points, and targeting benefit to patient outcomes, which is worth three points, compared to two or 1.5 for individual-level only outcomes. This is appropriate, since level 4b outcomes have been described as the ultimate outcome for healthcare leadership development. Points were added for using a validated instrument, providing the full collected data set, and exploring the relationships among variables. These measures increase the reliability and transparency of the study, as well investigating the important nuances of the phenomenon, as described earlier. The appropriateness of the analysis involves investigating whether improvements were sustained through a baseline compared to a post and post-post measure, as well as whether the design can adequately address the purpose and answer the research question(s) guiding the study. Finally, the sophistication of the analysis relates to triangulating data, including quantitative and qualitative, and addressing issues of *what* was achieved, as well as *how,* for whom, and it which ways.

As one additional representation of the state of the literature, Figure 5.1 below is a histogram depicting the distribution of the overall MERSQI scores for the 25 HEE included studies. Not only is the absence of strong evidence studies (blue) evident, but so is the fact that the majority are heavily in favour of the weaker credibility articles.

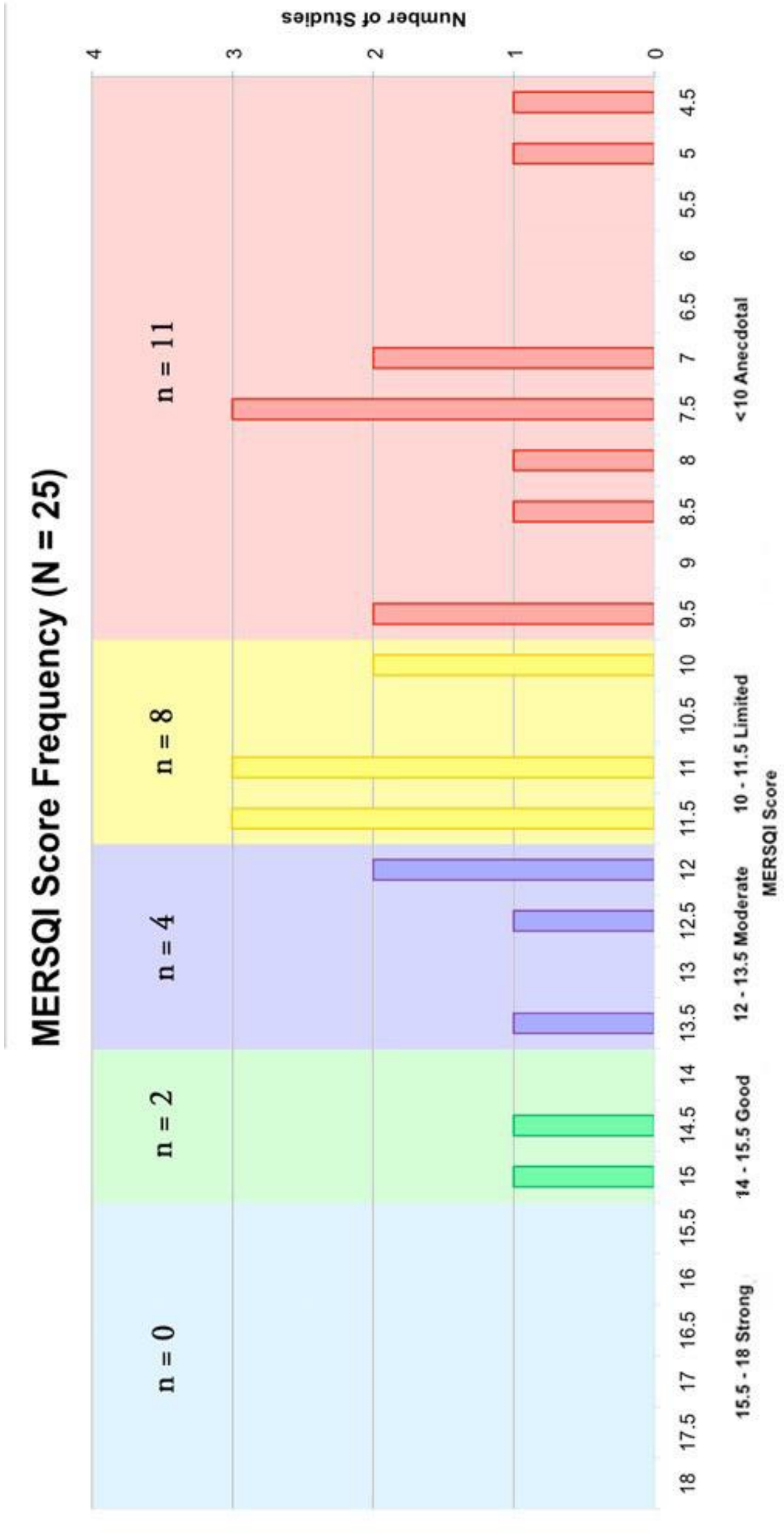


Figure 5.1 MERSQI Score Histogram.

The diagram above presents a histogram of the distribution of MERSQI scores mapped against the calibre groupings.

These calibre groupings will carry forward in the presentation of the raw data findings that follows so as to use the best available evidence as the reference point. The fact that only three studies received a score of 13 or higher out of 18 on the MERSQI scale further reinforces the claim that there is a clear lack of strong evidence and that the quality of leadership development research needs to improve. This is echoed by many, including by Rosenman et al. (2014) in their review.

5.4 Findings from the SEA Data Sources (MULTI, EMD, and HEE Included Studies)

As mentioned in the methodology section, the findings from the data collected from the various sources included in the systematic evidence analysis were compared and analysed, using the best calibre studies as the benchmark. What follows is a breakdown of the raw data findings and results of the analysis in terms of study design, sample and programme details, measurements, and outcomes of the studies. For the reader's convenience and in the interests of transparency, a full presentation of the codes for each of the included studies for both MULTI and the HEE SLR is included in the appendix on pages 350 and 369. The PRISMA guidelines state that for readers to gauge the validity and applicability of systematic reviews' results, they need enough details of the studies to determine their relevance (Liberati et al., 2009). Publishing summary data for each section of the analysis of the included studies also allows the analyses to be reproduced and examined further for patterns across studies (Liberati et al., 2009). Finally, providing the full data sets, rather than just highlights, precludes exposure to selective outcome reporting (Liberati et al., 2009).

Unless the narrative flow steered the presentation in a different direction, the general structure for each section below is as follows:

1) Overall raw data and most frequent items
2) Comparisons of the raw data among reviews (MULTI, EMD, HEE)
3) Mention of items that were noticeably absent or underrepresented
4) Reference to the raw data from the HEE good and moderate evidence studies
5) Comments on the most effective choices for each, along with examples
6) A summary table

It should be stated that the fact that nine studies were common to MULTI and HEE was factored into the analysis and discussion to avoid duplication and skewed results. Thus, the calculations of the combined MULTI (56)/HEE (25) set were out of 72, not 81. Before moving on to the data, a comment must be made about reporting issues.

5.5 Reporting Issues

As alluded to in the explanation of the MERSQI score category groupings, many of the studies featured insufficient reporting that made the analysis more challenging and detracted from the usefulness and credibility of their findings (Steinert et al., 2012). This situation is reminiscent of an earlier meta-analysis by Burke and Day (1986), which also describes a host of reporting issues. There were surprising omissions of key information regarding the studies, samples, programmes, data collection instruments, the process of data collection, measurements, analysis, and the connection between the previous items and the studies' conclusions. For example, although the outcomes of the Dannels et al (2008) study were credible, the details of the actual intervention are almost wholly absent. In four studies, such as that by Dickey (2014), it is not even clear whether the methodology is case study or action learning because it is not explained whether the authors' involvement was as independent researchers, or whether they facilitated the programme themselves. In two of the studies, it is not clear whether the authors used questionnaires or interviews to collect their data and two other articles fail to mention their methods altogether. Similarly, 13 of the 56 MULTI studies fail to stipulate at what point data was collected. A related issue is that Quaglieri, Penny, and Waldner (2007) claim that they evaluate the intervention every year, but do not mention specifically how. Likewise, Vimr and Dickens (2013) report that several of their participants' action learning projects have been implemented and are demonstrating a positive effect on quality and patient experience, but the authors do not explain what the projects are or how they are augmenting patients' experience and quality. They also mention that they used a 360-degree feedback tool, but provide no information on the results of it. Several studies did not distinguish between programme-wide outcome metrics and individually-reported outcomes or benefits, which limits them to anecdotal value.

Another challenge occurs when studies provide only highlights of their quantitative findings, rather than the full data set, or leave out the representative nature of qualitative responses among study participants. Similarly, many studies also omit outlying, particularly critical, opinions by study participants. Along the same lines, Steinert et al. (2012) note in their review that negative responses were also rarely listed, which echoes the "selective reporting bias" point made in the critique of the EMD SLRs (Liberati et al., 2009). There were exceptions to these omissions, however, and some authors, such as Malling et al. (2009), honestly volunteer the ways in which or for whom the interventions were unsuccessful. Outlying opinions are useful for nuancing key points and for understanding the phenomenon in a more complete and complex way (Alvesson & Spicer, 2012).

Elided data minimises the transparency and usefulness for readers and raises questions about whether certain details were overlooked in the analysis or left out intentionally. For example, Shah et al. (2013) mention that there were four instructional effects of their intervention, but give no indication of the representative nature of this claim. Readers might wonder how many participants exhibited these impacts or experienced these benefits. Was it a majority or merely one person? How many did not produce outcomes or benefit and why? Which aspects of the programme were considered responsible for contributing to the participants' development or the lack thereof? Which programme components can be modified to improve results? Could the differences in results correlate to differences in participants, such as level of seniority? An example of problematic reporting is that Miller et al. (2007) did not publish all the participants' responses, but list outcomes in instances when n = 2 and n = 3, despite their sample size of 210. This means that they convey findings reported by as little as less than one per cent of their population, as if these results are significant. Likewise, Sanfey et al. (2011) describe that participant reports that the advantages of networking were "frequent" (p. 356), but in fact only 12 of 110 participants identified that as a benefit. When authors fail to provide the full set of responses, given how under-representative the reports in the previous example are, it leaves one to question the credibility of the findings.

As mentioned previously, it is equally surprising is that Frich et al. (2014) allege that every single one of the 45 interventions they reviewed was successful, which was not the case in either MULTI or HEE. Burke and Day's (1986) meta-analysis of leadership development programmes, on the other hand, report that leadership development interventions were only "on the average moderately effective in improving learning and job performance" (p. 243). Likewise, Husebø and Akerjordet's (2016) review, despite their small sample (n = 12), identifies more than one study in which an intervention had no significant impact on outcomes. While it is true that often authors face space limitations when publishing, the PRISMA guidelines echo that this should not be accepted as an excuse for the omission of key aspects of the methods or results of included studies (Liberati et al., 2009). Therefore, in addition to their design strengths, the quality and completeness of the studies' reporting was found to significantly affect the credibility and usefulness of the reported findings and conclusions for the readers.

Finally, before moving to the presentation of the raw data findings, it should be repeated that there are 56 included studies in MERSQI, 25 in HEE, and 72 combined unique studies, once one accounts for the nine studies common to both reviews.

5.6 Research Designs

In terms of research designs, as depicted Table 5.6 below, the dominant approach in MULTI and HEE is case study methodology (n = 53, 74%). As mentioned earlier, it is unclear whether the methodology in four HEE studies was case study or action research, which means that the number of case studies could be nearly 80 per cent of the literature. The next most frequent overall are action research (n = 6) and survey (n = 6), though four and six of those respectively are from MULTI. Four studies are quasi-experiments, which is the same number of experiment designs in the combined sample. Of the latter group, only one, that by Jeon et al. (2013), is an RCT. The high number of case studies and case study/action research methodologies is common in leadership development literature given the complex nature of the phenomenon. As one author, Dalakoura (2010), asserts, this methodology has the potential to "generate rich insights into the mechanisms through which leadership is developed in practice" (p. 67). There are limitations associated with case studies as well. These include a typically small sample, often at one site, which can lessen the generalisability of the findings, the tendency to not include a control group, and the rarity of experiment-level connections between interventions and outcomes.

Despite the scientific value of experiments, most studies elected for an exploratory or explanatory design, with only 14 testing hypotheses (19%), only three of which derived from the HEE review. The low number of experiments and quasi-experiments and the near-lack of RCTs is problematic but not wholly unexpected. As mentioned previously, experiments require isolating one variable, which is a challenge with leadership development, and they require a high level of researcher control, which can be difficult to obtain for investigations involving leaders (Yin, 2003). More discussion of the Jeon et al. study will follow in the feature article section of chapter seven.

It is noted that not one study in the HEE SLR uses a survey methodology, which although it also has its limitations, it can be useful to audit the prevalence and perceived effectiveness of existing practices on a large scale, as well as contrasting phenomena in different contexts. Ardts, Velde, and Maurer (2010) and Mabey and Ramirez (2005) are two examples of employing this approach effectively. Of the six survey designs, only Dalakoura (2010) includes statistical analysis of business outcomes. Two of the other six surveys invited respondents to volunteer examples of business outcomes, but it is not clear whether these were supplied by all respondents. Without such information, the conclusions of these studies are based on large amounts of data derived from perceptions or descriptions of effectiveness without objective quantitative data verification, which is a limitation of Suutari and Viitala's (2008) study.

There are many drawbacks to utilising surveys for assessing the effectiveness of leadership development. The first is that doing so tends to leave key variables unspecified regarding which developmental activities or which combination of them are being described, as well as the duration of the interventions. Second, it tends to be Human Resources managers or CEOs who report on programme effectiveness, as was the case with Pinnington (2011), which can strongly bias the results, perhaps with corporate vested interests. This approach also tends to omit any nuances. Third, as suggested earlier, such persons may not be able to give accurate information on why, how, or in what ways the programmes are translating into performance outcomes, depending on how they have evaluated their programmes. Even more rarely do they volunteer information on why programmes have not been effective or what should be changed in order to improve interventions. McCauley (2008) suggests that research needs to move beyond simply whether programmes are effective or not, to investigate the specific effects of particular programme components and developmental activities, as well as combinations of them, and the role of organisational context.

Therefore, in the combined MULTI/HEE sample, the majority of studies feature a case study methodology; there are only four experiments; and there is only one RCT.

As a comparison to the extant reviews, that by Husebø and Akerjordet (2016) of 12 included studies included two randomised controlled trials and two quasi-experiments and the Rosenman et al. review (2014) of 45 studies included 12 RCTs and three nonrandomised, two-group comparisons. The higher number of RCTs identified in these reviews were very short interventions with the goal of improving observable leadership behaviour, rather broader leadership skills or organisational-level outcomes. Rosenman et al. (2014) suggest that leadership behaviours that were specified in their included studies were largely *task-centric* and *directive*. They add that time-sensitive, critical clinical situations, such as in a theatre or operating room, likely demand more directive, authoritative behaviours than more routine situations. This parallels the kind of leadership behaviours that were developed and evaluated in the experiments in the Husebø and Akerjordet (2016) review. For example, one such skill was, "Instruction to crew to red flag any significant deviation from standard operating procedure" (Husebø & Akerjordet, 2016, p. 2996). This represents a vastly different skill set compared those required as part of a year-long programme to prepare CEOs to lead an organisation, for example. As an even starker contrast to MULTI and HEE, Steinert et al. (2012) report that 15/19 studies were quasi-experiments, with only two case studies and one action research design. It is possible that this a product of the authors confusing methodologies or of excluding lower calibre studies. The lack of a published chart of study codes and critiques in the aforementioned review makes the source of differentiation between the designs of the included studies in the reviews challenging to evaluate and concomitantly detracts from the credibility of their review. Therefore, there was a higher percentage of RCTs in the extant reviews; however, this is likely attributable to very specific, short task-based interventions or a possible mislabelling of methodologies or of a restrictive inclusion criteria.

Unsurprisingly given the MERSQI weighting, of the two good evidence studies in the HEE review, one was an experiment and the other is a quasi-experiment. Of the moderate evidence studies, one was a quasi-experiment, two were case studies, and one featured an action research methodology. This variety raises an interesting question about the strengths and weaknesses of methodological approaches to study leadership development.

Table 5.6

Research Designs

		MULTI	HEE					Total
		N = 56	N = 25	Study Calibre				N = 72
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
Tested hypotheses		11 (20%)	3 (12%)	1	0	2	0	14 (19%)
Methodology	Case study	39 (70%)	14 (56%)	0	2	5	7	45 (67%)
	Survey	6 (11%)	0 (0%)	0	0	0	0	6 (8%)
	Action research	4 (7%)	2 (8%)	0	1	0	1	5 (7%)
	Action learning/Case study	0 (0%)	4 (16%)	0	0	1	3	4 (6%)
	Experiment	3 (5%)	1 (4%)	1	0	0	0	4 (6%)
	Quasi experiment	2 (4%)	3 (12%)	1	1	1	0	4 (6%)
	Grounded theory	1 (2%)	1 (4%)	0	0	1	0	1 (1%)

Above is depicted the breakdown of the various methodologies employed in the HEE and MULTI included studies.

5.7 Type of Data Collected and Collection Methods

As depicted in Table 5.7 below, in terms of the type of data collected, just more than half the studies (n = 42) collected both quantitative and qualitative data, 24 (33%) collected qualitative only, and 11 (15%) collected quantitative only. These numbers were also almost identical across the two reviews. One important clarification is that although 67 per cent of studies collected quantitative data, only 33 per cent collected *objective* data, which means that much of the quantitative data derives from self-reports. This will be discussed in more detail further on. Mixed methods are preferred for analysing the complexities of leadership development so one can use quantitative data to substantiate one's findings, draw correlations among variables, and track frequency distribution among respondents. When qualitative data is added, it allows researchers to analyse the nuances of *how*, for whom, to what extent, or in what circumstances interventions were effective or not (Kwamie, Dijk, & Agyepong, 2014; Marchal, Dedzo, & Kegels, 2010; Steinert et al., 2012). A good example of the need for both is Edmonstone (2009) who used only closed-ended multiple choice selections. The questionnaire asked respondents to select from a list of performance outcomes, but did not allow for additions or alternatives. This is helpful for generating frequency reports, but naturally limits the potential responses.

That said, it is surprising that only half the studies (51%) collected both types of data, though this is still a higher figure than the 20 per cent of studies that utilised mixed methods in the Frich et al. (2014) review and the 21 per cent in the Steinert et al. (2012) review. The restrictions of collecting only quantitative data are evidenced in Ardts, van der Velde, and Maurer (2010). In this study, participants were asked to rate the perceived outcomes and benefits of a leadership development programme, but there was no opportunity for respondents to explain the nuances of what made the programmes and its components effective or not. Similarly, both good and two of the four moderate evidence HEE studies collected *only* quantitative data, which means that they strove to show *that* something was true, but were not equipped to comment on the ever-important nuances mentioned above. Part of the reason for this is the aforementioned research designs; however, doing an experiment does not preclude researchers from adding qualitative data to form a fuller treatment of the topic. The other two moderate evidence studies used mixed methods. Therefore, in terms of research designs in MULTI and HEE, half the studies employed mixed methods, which is generally the most appropriate way to approach leadership development studies.

Table 5.7

Types of Data Collected

		MULTI	**HEE**					**Total**
		N = 56	N = 25	**Study Calibre**				N = 72
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
Data collected	Qualitative only	18 (32%)	8 (32%)	0	0	1	7	24 (33%)
	Quantitative only	8 (14%)	5 (20%)	2	2	0	1	11 (15%)
	Mixed methods	30 (54%)	12 (48%)	0	2	7	3	37 (51%)

Above is a summary of the breakdown of findings regarding data collected.

5.8 Data Collection Methods

In terms of data collection methods, the most common was questionnaires (n = 57), followed by interviews (n = 32) and document analysis (n = 20), as depicted in Table 5.8 below. More than a third of the studies (n = 25) relied on single methods alone and 21 of these were questionnaires. Relying on single measures precludes the researcher from triangulating data and clarifying responses or following up on emergent themes, an opportunity one is permitted when incorporating post-questionnaire interviews (Grotrian-Ryan, 2015; Marshall & Rossman, 2011; Roulston, 2010). Only 16 studies combined questionnaires and interviews, including only four HEE studies, and none was a good or moderate evidence study. It is unclear in three studies whether they used questionnaires or interviews and the methods in the two others were unclear altogether. The challenges associated with these errors was mentioned in the reporting issues section above. It is surprising that only nine studies (13%) used statistical analysis despite its usefulness in contrasting to participant outcomes. For example, Malling et al. (2009) performed a statistical analysis of pre and post MSF reports for 69 statements, comparing an intervention to a control group successfully.

Only the Jeon et al. (2013) study used statistical analysis to compare an intervention group's clinical outcomes to national averages or identical outcomes at another site that was not involved in the intervention, as a cluster control group. This is an unfortunate, pervasive oversight, especially since hospitals routinely collect much clinical data. Among other reasons, this inclusion is why the aforementioned study has been presented as an exemplar in chapter seven. It is also surprising that only one HEE study included programme observation, despite the usefulness of that technique in enabling researchers to get a real feel for the intervention and collect informal data from participants as the programme progresses. Both good and all four moderate evidence studies used questionnaires; one good evidence study used the only experiment; and both studies to use statistical analysis were moderate evidence ones. Therefore, the most common data collection method was questionnaires and nearly half the studies relied on single methods alone, preventing them from triangulating the data.

Table 5.8

Data Collection Methods

		MULTI	**HEE**					**Total**
		N = 56	N = 25	**Study Calibre**				N = 72
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
Methods	Questionnaires	42 (75%)	21 (84%)	2	4	7	8	57 (79%)
	Interviews	27 (48%)	7 (28%)	0	1	3	3	32 (44%)
	Document analysis	16 (29%)	6 (24%)	0	1	3	2	20 (28%)
	Statistical analysis	8 (14%)	2 (8%)	0	2	0	0	9 (13%)
	Programme observation	9 (16%)	1 (4%)	0	0	0	1	8 (11%)
	Focus group interviews	3 (5%)	1 (4%)	0	0	1	0	4 (6%)
	Unclear	1 (2%)	2 (8%)	0	0	0	2	2 (3%)
	Questionnaires or interviews	0 (0%)	3 (12%)	2	1	0	2	3 (4%)
	Conversation analysis	1 (2%)	0	0	0	0	0	1 (1%)
	Experiment	0 (0%)	1 (4%)	1	0	0	0	1 (1%)
	MSF	0 (0%)	1 (4%)	0	0	0	1	1 (1%)
	Video analysis	0 (0%)	1 (4%)	0	0	0	1	1 (1%)
	Single	19 (34%)	10 (40%)	1	2	4	3	25 (35%)
	Multiple	36 (64%)	13 (52%)	1	2	1	2	45 (63%)
	Questionnaires only	17 (30%)	7 (28%)	1	1	3	2	21 (29%)
	Questionnaires and interviews	14 (25%)	4 (16%)	0	0	3	1	16 (22%)

Above is a depiction of the methods of data collection in the included studies.

5.9 Study Samples

In terms of samples, 3,390 intervention participants and 685 control group participants were included in MULTI and the HEE SLR combined. Six studies did not include any sample information and all three of those in HEE were of anecdotal calibre. Of the 27 studies that included the sample participants' genders, 1440 were women (41%) and 2108 (59%) were men (see Table 5.9 below). Interestingly, in the HEE SLR, more than two thirds of the participants were women and although there was one programme exclusively for women, that by Dannels et al. (2008), there would have been a majority of women participants nonetheless. That percentage was lower in physician-only samples; however, with only 53 per cent women. The overall HEE numbers form a noticeable contrast to the gender split in MULTI, which featured 37 per cent women participants. Although speculations could be made about the number of women in senior professional roles, it would be premature to do so at this initial stage, although this does suggest an area of further investigation. Nearly two thirds of the studies (n = 45) did not report information on the participants' gender and even more (n = 58) did not list the mean ages. Unfortunately, in the HEE SLR, the gender and age details of the sample were not reported for all four moderate evidence studies and the mean age was omitted from both good evidence studies as well. The absence of sample details limits one extra level of analysis among variables.

Only seven studies used a control group (10%), which is similar to the 11 per cent in the Frich et al. (2014) review. This is unfortunate, since including control groups can potentially be very illuminating, such as Bowles et al.'s (2007) study of coaching and Malling et al.'s (2009) study. When analysing comparison groups, it is helpful if they are balanced in respect to size. Hassan, Fuwad, and Rauf (2010) is a good example; whereas Petriglieri, Wood, and Petriglieri (2011) had 48 in the experiment group but only seven in the control group. Taken together, these represent another indication of the insufficient calibre of research in the field. MacPhail et al. (2015) demonstrate the challenges of omitting control groups, since they report post-programme retention and promotions, but because they included no control group or national averages as a contrast, the relative nature of these figures is lost. The contrast becomes clear when compared to the Day et al. (2010) and the Dannels et al. (2008) studies, who both compared the CVs of those who completed the programme to the CVs of those who applied to the programme and were rejected. This exemplifies why including a control group is more effective than not. As expected given the MERSQI groupings, both good evidence studies used a control group, along with two moderate evidence studies; whereas, no limited or anecdotal studies did.

Only ten programmes (14%) studied multiple iterations of programmes, though including them can be useful to compare responses over time and to track the results of modifying programmes based on feedback from one iteration to the next. The credibility of the results of studies is enhanced when they involve more than one site, such as Chochard and Davoine (2011), and larger samples allow for more dependable results. For example, de Jong, Könings, and Czabanowska (2014) only involved 12 participants and many of the details of the sample are omitted; whereas Coloma, Gibson, and Packard (2012) had 166 participants over several years of the programme from eight different organisations. One good and two moderate evidence HEE studies featured multiple iterations.

In MULTI, 82 per cent of samples were single-domain, with 18 per cent being interdisciplinary. In HEE, 15 of the studies involved physician-only samples, whereas, ten (40%) were interdisciplinary (physicians and other healthcare professionals). This is similar to the 64 per cent of the interventions in the Steinert et al. (2012) review that featured only doctors. One clarification is that “interdisciplinary” is used in MULTI to mean participants from different professional domains, thus all healthcare professionals would be considered single-domain. In HEE, interdisciplinary samples indicates that physicians participated alongside other healthcare professionals. Although many authors laud the benefits of interdisciplinary programmes, such as Patel et al. (2015), only Vimr and Dickens (2013) make a case for physician-only programmes. An alternative is having profession-only syndicates or breakout sessions (such as physician-only) as part of an interdisciplinary programme to reap the advantages of both approaches. Although only one of two good evidence programmes was physician-only, all four moderate evidence studies were; thus, five of the six best evidence studies were physician-only. Therefore, many studies omitted key sample information; however, the majority of participants were women, particularly in HEE; there was a majority of physician-only programmes; and the studies tended not to use a control group or multiple iterations of the programme.

Table 5.9

Sample Details

		MULTI	**HEE**					**Total**	**Overlap**
		N = 56	N = 25		**Study Calibre**			N = 72	**MULTI/HEE**
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**	n =
Gender	Female	1136 (36%)	304 (66%)	85	NR	203	16	1380 (41%)	60
	Male	1951 (64%)	157 (34%)	12	NR	124	21	2010 (59%)	98
	NR	35 (63%)	17 (68%)	0	4	4	9	45 (63%)	7
Age	Mean	37	43.5	N/A	N/A	N/A	N/A	80.5	43.4
	NR	44 (79%)	21 (84%)	2	4	6	9	58 (81%)	7
Control group	Included	5 (9%)	4 (16%)	2	2	0	0	7 (10%)	2
Multiple iterations	Included	6 (11%)	6 (24%)	1	2	1	2	10 (14%)	2
Professional domains	MULTI: single HEE: MDs only	46 (82%)	15 (60%)	1	4	1	9	56 (78%)	5
	Interdiscipinary	10 (18%)	10 (40%)	1	0	7	2	16 (22%)	4

The table above depicts the examination of the sample details across the two reviews.

As shown in Table 5.10 below, there was a range of the level of seniority of the sample participants. The heterogeneity of levels of seniority in the HEE SLR made it difficult to group studies into traditional categories of junior, middle, and senior leaders; however, when the SLRs were combined, there was a very close distribution among the three. The most frequently studied group was mid-level professionals (n = 15); however, this is surpassed if one combines the 14 senior leaders/consultant/senior faculty studies with the nine featuring CEOs and executive leaders (26% total). The last most frequent sample was junior managers/junior physicians/residents (n = 12). There is a noticeable disparity in the studies involving midlevel leaders in favour of MULTI versus HEE, with n = 14 compared to n = 1 respectively. This is an additional contrast to Kuo et al.'s (2010) claim that the majority of leadership development for doctors is for mid-career professionals. Whether this suggests an overall lack of leadership development for physicians at the mid-career level or not is worthy of further investigation. It is surprising that of the 45 studies identified in the Frich et al. (2014) review, not one was for senior level participants. As suggested earlier, the evidence supporting the importance of leadership at the top levels of organisations is convincing enough to lend extra importance to development programmes at the highest levels. In the HEE review, one of the two good and one of the moderate evidence studies focused on this demographic.

It should be noted that more than a third the studies (n = 32) failed to specify the level of seniority or included more than one level. As mentioned in the introduction, there are calls for leadership development for leaders at all levels (Van Aerde, 2013), which makes it encouraging that there are a good number of well-reported programmes for junior leaders, including two of the four moderate evidence studies. Therefore, the most common level of participants' seniority was senior leaders, though midlevel and junior leaders were also decently represented.

Table 5.10

Level of Seniority

	MULTI	**HEE**					**Total**
	N = 56	N = 25	**Study Calibre**				N = 72
Feature	**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
Midlevel surgeons/ Middle managers	14 (56%)	1 (4%)	1	0	0	0	15 (21%)
Senior faculty/ Consultant/ Senior managers	12 (48%)	5 (20%)	1	1	1	2	14 (19%)
Junior Physicians/ Residents/ Junior managers	5 (20%)	9 (36%)	0	2	1	6	12 (17%)
CEOs/ Executives	9 (36%)	0 (0%)	0	0	0	0	9 (13%)
Other	8 (32%)	0 (0%)	0	0	0	0	8 (11%)
Physicians/ Surgeons unspecified	5 (20%)	5 (20%)	0	1	2	2	8 (11%)
Managers unspecified	6 (24%)	0 (0%)	0	0	0	0	6 (8%)
Human Resource Managers	4 (16%)	0 (0%)	0	0	0	0	4 (6%)
MBA students	4 (16%)	0 (0%)	0	0	0	0	4 (6%)
District/ Area managers	2 (8%)	0 (0%)	0	0	0	0	3 (4%)
University academics	3 (12%)	0 (0%)	0	0	0	0	3 (4%)
High potential physicians	1 (4%)	2 (8%)	0	0	1	1	2 (3%)
Middle and senior leaders	0 (0%)	1 (4%)	0	0	1	0	1 (1%)
Mixed	1 (4%)	2 (8%)	0	0	2	0	1 (1%)

The table above depicts the dissection of the sample participants' level of seniority feature in each of the two reviews.

5.10 Professional Domains: MULTI Only

As Table 5.11 illustrates, most of the research identified in MULTI is being conducted in the fields of healthcare (36%) and business (30%), though it should be repeated that nurses and school administrators were excluded from the search. Even so, only half of the MULTI healthcare studies included physician samples. It is interesting that only two studies analyse military programmes, given how well-established leadership development is in that domain, and of those, only one concentrated on operational leaders. Two studies investigated leadership in different domains, but they did so generally by comparing the private sector in relation to public, rather than specific domains such as military versus healthcare. For this reason, their analyses shed little light on the generic versus contextual question mentioned in the previous chapter. The first, McAlearney et al. (2010), relates to a transformational leadership intervention for public and private organisations. The second, Pinnington's (2011) survey, suggests that there was no difference in perceived effectiveness of leadership development practices in private versus public/not-for-profit sectors, which again does not address the aforementioned question. Thus, there is a noticeable lack of studies attempting to compare the leadership development in different professional domains.

Table 5.11

Professional Domains in MULTI

	MULTI	
Feature	**N = 56**	**n (%)**
Professional domain	Healthcare	20 (36%)
	Business	17 (30%)
	Public sector	6 (11%)
	Government	4 (7%)
	Mixed/ Unspecified	3 (5%)
	Higher education	3 (5%)
	Other	3 (5%)
	Not for profit	2 (4%)
	Financial	2 (4%)
	Military	2 (4%)

Above is an outline of the breakdown of the MULTI studies' participants' professional domains.

Before discussing the programmes themselves, the selection criteria for participants in the various programmes is worth noting, as shown in Table 5.12 below. Previous studies have reported that this factor has affected programme outcomes (Kwamie et al., 2014). The most common approaches to selection was participants who were nominated (n = 12) and those who volunteered (n = 12), followed by those who applied and were selected (n = 8). Surprisingly, nearly half the studies left the selection criteria unclear (n = 34). It is interesting to note that only two studies included programmes where participants were mandated to attend, although one of them was a good evidence HEE one. This low number is only slightly higher than that in the Steinert et al. review, in which none of the leadership interventions was mandatory (Steinert et al., 2012). This reinforces the key precursor to the principles of adult learning to be described in the conclusions explored section of the next chapter. The other good and two of the moderate HEE evidence studies featured participants who volunteered and those in the final two moderate studies applied and were selected. Two studies, though of anecdotal credibility, describe residents being involved in taking ownership of researching and designing their own leadership programme specifically for their career stage (Blumenthal et al., 2014; Dickey et al., 2014). Therefore, although the selection criteria is often unclear, it most commonly involves participants who were nominated, volunteered, or were applied and selected.

Table 5.12

Selection Criteria

		MULTI	**HEE**					**Total**
		N = 56	N = 25	**Study Calibre**				N = 72
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
Selection criteria	Unclear	32 (57%)	5 (20%)	0	0	1	4	34 (47%)
	Nominated	7 (13%)	7 (28%)	0	0	3	4	12 (17%)
	Volunteered	8 (14%)	6 (24%)	1	2	2	1	12 (17%)
	Applied and selected	4 (7%)	5 (20%)	0	2	2	1	8 (11%)
	N/A	3 (5%)	0 (0%)	0	0	0	0	3 (4%)
	Required	1 (2%)	2 (8%)	1	0	0	1	2 (3%)
	Mixed	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Randomly selected	1 (2%)	0 (0%)	0	0	0	0	1 (1%)

Above is a breakdown of the participant selection criteria employed in studies in the two reviews.

5.11 Programmes

In terms of **locations** of the programmes, as detailed in Table 5.13 below, the majority of the data comes from North America (n = 37). There were small numbers of studies from other Western countries, including the UK (n = 11), Europe generally (n = 7), Australia (n = 5), and Scandinavia (n = 3). There were only three studies from Africa, one from Asia, and none from the Middle East, or Central or South America. The extant reviews found a similar concentration (Frich et al., 2014; Hartley & Hinksman, 2003; Husebø & Akerjordet, 2016; Steinert et al., 2012; Straus, Soobiah, & Levinson, 2013). This prompts the question of the applicability of the findings arising from Western programmes to those in other continents.

Regarding **faculty**, which was only tracked in the HEE SLR, more programmes used a combination of internal and external faculty (n = 9) compared to internal only (n = 7) or external (n = 2), while nearly a third of the programmes (n = 7) left details of the faculty out. One of the two good and two of the four moderate evidence studies used internal faculty, while the other three in these groupings omitted the faculty groupings. To inform the design of the programme, only ten HEE studies (40%) reported conducting a needs assessment before launching the despite many claims that doing so improves programme outcomes (Hartley & Hinksman, 2003)**.**

Finally, in terms of **structure**, more than half of the programmes were in-house (n = 37) and 22 were external, as a contrast to the McKinsey report that suggested that most leadership programmes for clinicians were external (Mountford & Webb, 2009). This is similar to the 57 per cent of leadership programmes in the Steinert et al. (2012) review being in-house. Interestingly, nearly three quarters of the HEE studies were in-house (n = 18), compared to only 43 per cent of the MULTI programmes. 11 MULTI studies did not specify whether their programmes were in-house or external and only two featured a combination, surprisingly. Only one study compared in-house to external programmes. Suutari and Viitala's (2008) survey of perceived management development effectiveness suggests that there was no significant difference between training organised internally or by an outside provider. There is much ongoing debate on this matter, particularly since there is often a much higher cost for external programmes (MacPhail et al., 2015). More work is needed to indicate in which ways or circumstances one may be more beneficial than the other. The good and moderate evidence HEE studies were split as evenly as possible, with half of each featuring external and half featuring in-house programmes. Therefore, the majority of studies come from North America and other Western countries, many feature mixed internal and external faculty, and they tended to be in-house programmes.

Table 5.13

Programme Details

		MULTI	HEE					Total
		N = 56	N = 25	Study Calibre				N = 72
Feature		n (%)	n (%)	Good (2)	Moderate (4)	Limited (8)	Anecdotal (11)	n (%)
Location	United States	23 (42%)	15 (60%)	1	3	4	7	33 (46%)
	UK	10 (18%)	3 (12%)	0	0	1	2	11 (15%)
	Europe	6 (11%)	1 (4%)	1	0	0	0	7 (10%)
	Australia	4 (7%)	1 (4%)	0	0	1	0	5 (7%)
	Canada	4 (7%)	1 (4%)	0	0	0	1	4 (6%)
	Africa	2 (4%)	1 (4%)	0	0	1	0	3 (4%)
	Multiple (unspecified)	3 (5%)	0 (0%)	0	0	0	0	3 (4%)
	Scandinavia	2 (4%)	2 (8%)	0	1	1	0	3 (4%)
	Asia	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Australia and NZ	0 (0%)	1 (4%)	0	0	0	1	1 (1%)
Faculty	Internal	-	7 (28%)	1	2	0	4	-
	External	-	2 (8%)	0	0	1	1	-
	Mixed	-	9 (36%)	0	0	5	4	-
	Unclear	-	7 (28%)	1	2	2	2	-
Needs Assessment	Yes	-	10 (40%)	0	2	2	6	-
	No/unclear	-	15 (60%)	2	2	6	5	-
Structure	In-house	24 (43%)	18 (72%)	1	2	6	9	37 (51%)
	External	19 (34%)	7 (28%)	1	2	2	2	22 (31%)
	Both	2 (4%)	0 (0%)	0	0	0	0	2 (3%)
	Unspecified	11 (20%)	0 (0%)	0	0	0	0	11 (15%)

Above is a depiction of the programme features of the interventions described in the two reviews.

As is made clear in Table 5.14 below, the range in **length** of programmes was largely heterogenous, spanning from one or two days to four years. The most frequent lengths were 4 – 6 months (n = 10) and 8 – 11.5 months (n = 10), followed by 3 – 5 days (n = 6), year-long programmes (n = 6), programmes that were longer than a year (n = 6). The latter two are a contrast to Kuo et al.'s (2010) claim that the majority of physician leadership development is intense, short interventions. There are two noticeable differences between the lengths of programmes in MULTI and HEE: no MULTI programmes were shorter than three days; whereas there were three HEE interventions of this length. Interestingly, in the Suutari and Viitala (2008) survey of senior business leaders, 62 per cent reported having undertaken training lasting one to three days. As mentioned previously, the majority of programmes in the Husebø and Akerjordet (2016) and Rosenman et al. (2014) reviews were short interventions. Rosenman et al.'s claim that the leadership behaviours in their study were largely task-centric and directive echoes Ten Have et al.'s (2013) finding that simulations were found with good evidence to increase observable behaviour. Watkins, Lysø, and deMarrais (2011) suggest that these behaviours and what they termed "surface changes" are easier for others to recognise, whereas cognitive and softer skills are not. Thus, a preliminary, but not definitive, comment is that tasks seem to be much easier to "train" in a short period time than is developing softer or broader leadership skills, such as systems thinking or developing one's strategic perspective.

The second difference in the length of programmes between MULTI and HEE is that there is a higher percentage of HEE programmes that were a year long or longer (24% and 16% respectively) compared to MULTI (5% and 4%) and all three year-long MULTI studies are medical leadership programmes that are also included in the HEE SLR. This is not only attributable to residency programmes, since only four of ten longer programmes were for residents. This is worthy of further investigation. The second good evidence HEE study is unclear about the programme length; whereas, three of the programmes in the moderate evidence studies were a year or longer. The final moderate evidence study featured a six months-long intervention, which makes for an interesting contrast between the length of the Ten Have et al. programme along with those in the extant reviews, versus the longer, moderate evidence studies. This heterogeneity mirrors the paucity of evidence in the literature regarding the optimal length for programmes, whether generally or for specific contexts or purposes. Therefore, there was a vast range of lengths of programmes, which suggests a need for further research regarding optimal lengths for differing goals and levels of seniority.

Table 5.14

Programme Length

		MULTI	HEE					Total
		N = 56	N = 25	Study Calibre				N = 72
Feature		n (%)	n (%)	Good (2)	Moderate (4)	Limited (8)	Anecdotal (11)	n (%)
Length	1-2 days	0 (0%)	3 (12%)	1	0	0	2	3 (4%)
	3-5 days	5 (9%)	1 (4%)	0	0	1	0	6 (8%)
	1 week	0 (0%)	1 (4%)	0	0	1	0	1 (1%)
	1 month	2 (4%)	1 (4%)	0	0	0	1	2 (3%)
	1.5-3.5 months	3 (5%)	1 (4%)	0	0	1	0	3 (4%)
	4-6 months	9 (16%)	3 (12%)	0	1	1	1	10 (14%)
	8-11.5 months	8 (14%)	4 (16%)	0	0	1	3	10 (14%)
	Year-long	3 (5%)	6 (24%)	0	1	3	2	6 (8%)
	> a year	2 (4%)	4 (16%)	0	2	0	2	6 (8%)
	Unclear	18 (32%)	1 (4%)	1	0	0	0	19 (38%)

The table above outlines the distribution of programme lengths in the two reviews.

5.11.1 Developmental Activities

42 different developmental activities were included, as outlined in Table 5.15 below. The most common was workshops, which nearly than half the studies (n = 32) included, followed by 360s/MSF (n = 28), coaching (n = 27), lectures (n = 24), and action learning (n = 23). There is a heterogenous series of combinations of components and only seven studies that utilised only one activity. As one example, Miller et al. (2007) state that in their study participants commonly cited developmental activities in conjunction with one another in relation to outcomes, which reinforces a point raised in the conclusions explored section regarding the utility of combining them symbiotically. The combination of 360s and coaching is reminiscent of DeRue’s et al.’s (2012) report that structured reflection was said to enhance the outcomes of leadership development. Although Frich et al. (2014)’ and Rosenman et al. (2014) claim that the majority of physician leadership programmes are still based on the traditional lecture and seminar-format (Rowland, 2016), with only 24 programmes (33%) involving lectures in the combination of MULTI and HEE, it is possible that the trend seems to be branching out to more experiential methods such as 360s, simulations/role plays, and action learning that address actual organisational needs directly. Other authors, including Marcus (2004), Suutari & Viitala (2008), and Watkins, Lysø, & deMarrais (2011), have noted this shift as well (Blumenthal et al., 2014; Steinert et al., 2012). Getha-Taylor (2013) suggest

that this is the case because traditional models of instruction are less effective given the needs of adult learners. This will be discussed further in the conclusions explored section.

Interesting contrasts between MULTI and HEE are that coaching was included almost twice as often in MULTI (n = 24 (43%), compared to n = 6 (24%) in HEE); however, simulations, mentoring, and case study analysis were more common in HEE programmes. Facilitator feedback was absent in MULTI and peer feedback was scarce in both reviews despite the finding that one good and two of four moderate evidence HEE studies involved simulations and also included facilitator feedback. A similar contrast is that simulations featured in 81 per cent of the studies in the Rosenman et al. (2014) review compared to 24 per cent in this combined study, which seems to correlate with the former's focus of short, team-based interventions for which simulations appear to be used more regularly. It is surprising that self-reflection (n = 9), journaling (n = 5), and video-taping (n = 2) were incorporated so seldom, considering how well they draw on the principles of adult learning.

Therefore, among a multitude of different developmental activities, the most common were workshops, 360s, coaching, lectures, and action learning, with more experiential activities than in previous reviews.

Table 5.15

Developmental Activities

		MULTI	**HEE**					**Total**
		N = 56	N = 25	**Study Calibre**				N = 72
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
Developmental activities	Workshops	25 (45%)	14 (56%)	1	1	4	8	32 (44%)
	360's/ MSF	23 (41%)	9 (36%)	0	1	5	3	28 (39%)
	Coaching	24 (43%)	6 (24%)	0	0	4	2	27 (38%)
	Lectures	19 (34%)	8 (32%)	0	1	2	5	24 (33%)
	Action learning	19 (34%)	8 (32%)	0	1	4	3	23 (32%)
	Reading assignments	11 (20%)	11 (44%)	0	2	2	7	19 (26%)
	Group discussions	11 (19%)	7 (28%)	0	0	1	0	18 (25%)
	Simulations/ role play	10 (18%)	9 (36%)	1	2	2	4	17 (24%)
	PDP	13 (23%)	7 (28%)	0	1	3	3	16 (22%)
	Mentoring	8 (14%)	8 (32%)	0	3	3	2	15 (21%)
	Guest speakers	9 (16%)	7 (28%)	0	1	3	3	13 (18%)
	Small group discussion/ work	4 (7%)	11 (44%)	0	2	4	5	12 (17%)
	Psychometric/personality tests	8 (14%)	6 (24%)	0	1	5	0	11 (15%)
	Presentations	3 (5%)	6 (24%)	0	0	3	3	9 (13%)
	Self-reflection/ assessment	7 (13%)	3 (12%)	0	0	2	1	9 (13%)
	Case study analysis	4 (7%)	7 (28%)	0	1	2	4	8 (11%)
	Assignments	9 (16%)	0 (0%)	0	0	0	0	7 (10%)
	Job shadowing / assignment/ rotation	6 (11%)	2 (8%)	0	1	1	0	7 (10%)
	Networking	6 (11%)	2 (8%)	0	0	1	1	7 (10%)
	Team projects/challenges	5 (9%)	3 (12%)	0	0	0	3	7 (10%)
	Facilitator feedback	0 (0%)	6 (24%)	1	2	2	1	6 (8%)
	Online modules/ e-learning	4 (7%)	2 (8%)	0	1	1	0	6 (8%)
	Journal	4 (7%)	2 (8%)	0	0	0	2	5 (7%)
	Site visits/ observed case study	3 (5%)	2 (8%)	0	0	2	0	4 (6%)
	Peer feedback	1 (2%)	3 (12%)	0	0	0	3	3 (4%)
	Counselling/ psychotherapy	2 (4%)	1 (4%)	0	0	1	0	3 (4%)
	Internship	2 (4%)	1 (4%)	0	0	1	0	3 (4%)
	Advisory groups	0 (0%)	2 (8%)	0	1	1	0	2 (3%)
	Experiential activities	0 (0%)	2 (8%)	0	0	1	1	2 (3%)
	Examinations	1 (2%)	1 (4%)	0	0	0	1	2 (3%)
	Peer support	1 (2%)	2 (8%)	0	0	1	1	2 (3%)
	Team building	2 (4%)	0 (0%)	0	0	0	0	2 (3%)
	Video taped	1 (2%)	1 (4%)	1	0	0	0	2 (3%)
	Critical incidents	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	E-portfolio	0 (0%)	1 (4%)	0	1	0	0	1 (1%)
	Films	2 (4%)	0 (0%)	0	0	0	0	1 (1%)
	Formal qualifications	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Outdoor development	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Service learning	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Storytelling	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Yoga	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	None reported	4 (7%)	1 (4%)	1	0	0	0	5 (7%)
	Studies using only one	6 (11%)	1 (4%)	0	1	0	0	7 (13%)

Above is a table displaying the breakdown of the developmental activities across the two reviews.

5.12 Measurements: Kirkpatrick Outcome Levels

As described earlier and depicted in Table 5.16 below, the outcome metrics and reported outcomes and benefits were categorised according to Kirkpatrick's four-level model. To again clarify the distinction, outcome metrics are programme-wide measurements for all participants; whereas, reported outcomes and benefits are any items offered by respondents to open-ended questions. This section lists a combination of both. 64 per cent of the studies (n = 46) evaluated participant satisfaction (Level 1), 57 per cent (n = 41) assessed participants' change in attitude or perception outcomes (Level 2a), 71 per cent (n = 51) reported participants' increased knowledge or skills outcomes (Level 2b), 53 per cent (n = 38) described subjective participant behaviour changes (Level 3a), and 35 per cent (n = 25) claimed that there were objective participant behaviour changes (Level 3b). 18 studies (25%) reported organisational outcomes (Level 4a) and only eight (11%) reported benefit to patients or clients outcomes (Level 4b). The fact that only six HEE studies reported Level 4b outcomes is especially surprising given that it should be a priority for medical leadership development, as mentioned previously. This is even higher than the only two unique studies in MULTI that reported outcomes at this level. 15 studies (21%) relied exclusively on subjective outcomes at the Kirkpatrick Levels 1 – 3a, with six of the seven HEE studies that did so were of anecdotal calibre. Although the weighting is heavily in favour of the individual-level outcomes, the figures above refute Kellerman's critique cited earlier that most research never exceeds Level 1 evaluation. Although Level 1 outcomes on their own are insufficient, in addition to being useful for quality control, Sanfey et al. (2011) point out that they are also a valuable consideration to increase faculty motivation and future participation.

Not surprising given the MERSQI scale, but all six HEE studies in the good and moderate evidence categories included a Level 3b (objective behaviour change) measurement, but only 3/11 anecdotal studies did, which could suggest that it was a greater differentiator than its point value would indicate. Likewise, less than a third of the MULTI studies reported a 3b outcome (n = 15). One way of demonstrating a Level 3b outcome is to include supervisors' ratings of leadership behaviour after the programme, but surprisingly, only four studies did so. As a contrast, only two of the 45 studies in the Frich et al. (2014) review included 3b outcomes. It is also noted that only two studies measured multi-source feedback (MSF) pre and post intervention, which can be another effective way to measure behaviour change.

Despite the pre-eminence of measuring leadership development's impact on the organisation and as it benefits patients, unfortunately, only one good or moderate HEE study reported 4a or 4b level outcomes. Interestingly, of the studies from 2005 – 2012 (n = 56), only three studies included reports of 4b outcomes (5%); whereas, from 2013 – 2016 (n = 16), five did (31%), which could indicate that it is gradually becoming more of a priority. That said, of the aforementioned five recent studies, only two collected objective data to substantiate their claims. Striving for Level 4b outcomes in healthcare does not depend on doctor-exclusive programmes apparently, since half of the HEE studies (n = 3) that reported them involved interdisciplinary samples. Finally, only six HEE studies (25%) used 4a and 4b levels as outcome metrics, versus individual reports of outcomes or benefits at these levels. The Frich et al. (2014) review also reports a definite lack of objective outcome data at Levels 4a and 4b, with only 13 per cent of the studies including these outcomes. As mentioned previously, the Straus et al. (2013) review did not consider 4a and 4b outcomes and the Steinert et al. (2012) review failed to mention 4b outcomes. As intimated earlier, what is most curious about this is that there is a good amount of data that hospitals routinely collect, such as human resource data including workplace satisfaction reports, absenteeism, and turnover, all of which leadership skills are said to influence (Doran et al., 2004; Gagnon et al., 2006; Hayes, 2007; Jeon et al., 2013; Artz, Goodall, & Oswald, 2016). Furthermore, there are clinical outcomes that hospitals are made to gather, including for organisational performance evaluations like those mentioned in the appendix on page 338. These include survival/mortality rates, length of stay in hospital, readmission rates, and error rates, among others. This indicates that there is a wide range of available data that is not being used as it could be, and perhaps should be, as outcome metrics for leadership development programmes. Therefore, although the majority of studies claim that leadership knowledge and skills increased, there were very few reported outcomes at the organisational or clinical levels and even fewer of these claims were reinforced with objective data.

Table 5.16

Kirkpatrick Outcome Levels

	MULTI		HEE						Total	Overlap
	N = 56		N = 25	Study Calibre					N = 72	MULTI/HEE
Feature	**n (%)**		**n (%)**	Good (2)	Moderate (4)	Limited (8)	Anecdotal (11)		**n (%)**	n =
1 Satisfaction	34 (61%)		21 (84%)	0	3	7	11		46 (64%)	9
2a Attitude/perception	30 (54%)		18 (72%)	1	3	7	7		41 (57%)	7
2b Knowledge/skills	37 (66%)		22 (88%)	1	4	8	9		51 (71%)	8
3a Behaviour (subj)	28 (50%)		16 (64%)	0	3	7	6		38 (53%)	6
3b Behaviour (obj)	15 (27%)		15 (60%)	2	4	6	3		25 (35%)	5
4a Organisational	16 (29%)		5 (20%)	0	0	4	1		18 (25%)	3
4b Patients/clients	4 (7%)		6 (24%)	0	1	4	1		8 (11%)	2
>4 types	7 (13%)		14 (56%)	0	3	7	4		16 (22%)	5
Only Levels 1 + 2	14 (25%)		4 (16%)	0	0	0	4		17 (24%)	1
Only Levels 1, 2, or 3a	9 (16%)		7 (28%)	0	0	1	6		15 (21%)	1
Levels 3b and 4a or 4b	4 (7%)		6 (24%)	0	1	5	0		9 (13%)	1

Above is a table depicting the distribution of reported outcomes according to the Kirkpatrick levels across the two reviews.

5.13 Data Collected

In terms of raters, as depicted in Table 5.17 below, a variety of different sources were drawn upon to collect data. Self-ratings were the most common (n = 66), followed by supervisor (n = 23), peer (n = 18), and supervisee (n = 16). As mentioned earlier, only eight studies used statistics. More than half the studies (n = 38) relied on single raters, which prevents data triangulation and leaves studies vulnerable to response bias. Participants may be likely to under-report behaviours deemed inappropriate by others and over-report favourable ones (Solansky, 2010). They may also want to demonstrate to their supervisors that they are grateful for the opportunity and justify the participation to themselves and others by reporting positive manifestations of development (Berg & Karlsen, 2012). Finally, Edmonstone (2013) suggests that participants who applied or were nominated may feel proud to have earned a place on the programme and as a result may expect that they will perform well afterwards and have a positive impact, which, he says, is likely to influence their ratings. An example from this study is the Malling study (2009), where although participants rated themselves higher a year after the intervention compared to a baseline measure, their external raters did not change their assessments. These points make it even more surprising that only eight studies referred to statistics and only 23 used supervisors as raters.

As an interesting contrast, Solansky (2010) states that in her study, the average observer score was actually higher than the self-reports; however, she goes on to add that the exclusive use of self-reports is a major weakness of evaluations of leadership development programmes due to many kinds of response bias. With this in mind, it seems that for increased reliability and accuracy, triangulation is necessary, even though Watkins, Lysø, and deMarrais (2011) state that some changes are less readily observable to other raters. They list examples of cognitive abilities such as one's internal vision being broadened, increased confidence, listening skills, and mentoring and developing supervisees. Interestingly, the Ten Have et al. (2013) study included peer and facilitator ratings and *not* self-ratings, which is more typical of an experiment. In addition to the aforementioned study, three of the four moderate evidence HEE studies used multiple methods, whereas only 2/9 limited evidence studies did. Therefore, there is an over-reliance on self-ratings, which limits the credibility of those studies' findings (Steinert et al., 2012).

The **type of data** collected was grouped into three categories: subjective descriptions of outcomes (qualitative), subjective numbers, such as Likert scale self-ratings (quantitative), and objective data (quantitative). 58 studies (81%) included subjective descriptions, just more than half (n = 38) included subjective numbers, and only 28 (39%) presented objective data. Only six studies included all three, although Kuo et al. (2010) and Stewart (2009) are good examples; 44 (61%) were limited to subjective data only; and nearly a third (n = 23) were restricted to subjective descriptions only. As an example of the limitations of the latter group, Shah et al. (2013) report that the programme they studied was successful in improving patient safety, but other than isolated qualitative quotations to that effect, there is no credible evidence to reinforce this assertion. Straus, Soobiah, and Levinson (2013) also remark that the majority of studies used self-reported data from participants, which is echoed by Blumenthal et al. (2014) and Malling et al. (2009). Frich et al. (2014) also highlight the paucity of studies on healthcare leadership that include quantifiable outcomes. Thus, there is a significant field-wide lack of objective data to substantiate outcomes. It has already been mentioned that the Day et al. (2010) and Dannels et al. (2008) studies are two examples of using objective data effectively by comparing an intervention group who participated in a leadership programme to those who applied and were not accepted. The latter researchers compared self-assessments of leadership abilities, reinforced by promotions and time spent on administrative responsibilities. Another study that uses objective data very effectively is that by DeRue et al. (2012), which uses internship offers and starting salary figures to compare large experimental and control groups of MBA students. They report statistically significant differences between the two as a direct result of the leadership programme, which, given the comprehensive nature of their controls, leads to more credible findings. Both good and two moderate evidence HEE studies included objective data, whereas not one anecdotal study did. Therefore, there is a noticeable lack of objective data to reinforce subjective claims.

In terms of the **focus** of the evaluation, 22 studies (31%) targeted only participant outcomes and benefits, 11 (15%) critiqued only the programme, and only 40 (56%) did both. Two studies are unclear which the focus was, including one of the good evidence HEE studies. Three of the four moderate evidence studies evaluated both. As suggested earlier, to fully understand the phenomenon of leadership development, one needs to evaluate the programme in terms of what worked and what did not, to what extent, for whom, and in what circumstances. Likewise, one must also evaluate the participants' development, as measured by how they apply their learning in terms of outcomes at the individual, team, organisational, and benefit to patients/clients levels. Therefore, only slightly more than half the studies evaluated both the participants and the development programmes, despite the importance of doing so for leadership development research.

Table 5.17

Data Collection

		MULTI		HEE						Total	Overlap
		N = 56		N = 25	Study Calibre					N = 72	MULTI/HEE
Feature		**n (%)**		**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**		**n (%)**	n =
Raters	Self	52 (93%)		22 (88%)	1	4	8	9		66 (92%)	8
	Supervisor	20 (36%)		5 (20%)	0	0	2	3		23 (32%)	2
	Peer	15 (27%)		5 (20%)	1	1	1	2		18 (25%)	2
	Supervisee	15 (27%)		3 (12%)	0	1	1	1		16 (22%)	2
	Statistics	6 (11%)		2 (8%)	0	1	1	0		8 (11%)	0
	Facilitator	5 (9%)		5 (20%)	1	1	0	3		7 (10%)	3
	Single	30 (54%)		13 (52%)	1	1	6	5		38 (53%)	5
	Multiple	26 (46%)		11 (44%)	1	3	2	5		33 (46%)	4
	Self only	28 (50%)		13 (52%)	1	1	6	5		32 (44%)	5
Type of Data Collected	Subjective descriptions	46 (82%)		21 (84%)	1	3	8	9		58 (81%)	9
	Subjective numbers	32 (57%)		12 (48%)	0	4	3	5		38 (53%)	6
	Objective	16 (29%)		17 (68%)	2	3	8	5		28 (39%)	5
	All three	5 (9%)		2 (8%)	0	1	1	0		6 (8%)	1
	Subjective descriptions only	18 (32%)		7 (28%)	0	0	1	6		23 (32%)	2
	Subjective only	40 (71%)		8 (32%)	0	1	1	6		44 (61%)	4
What Was Evaluated	Participants only	19 (34%)		5 (20%)	1	4	8	7		22 (31%)	2
	Programme only	9 (16%)		3 (12%)	0	3	6	10		11 (15%)	1
	Both	31 (55%)		15 (60%)	0	3	6	6		40 (56%)	6
	Unclear	0 (0%)		2 (8%)	1	0	0	1		2 (3%)	0

The table above presents the details of the raters, type and focus of evaluation in the two reviews.

In terms of when data was collected, as shown in Table 5.18 below, the most common was a post-post measurement (n = 35). The next frequent was a post measure (n = 32), followed by pre-programme or baseline (n = 30), and only one HEE study measured during the programme, compared to 14 in MULTI. It is surprising that 37 studies (51%) elected not to include a post-post measure, which precludes assessing the application of development to the workplace. Straus, Soobiah, and Levinson (Straus et al., 2013) found a similar majority of medical leadership programme studies that were restricted to a post collection with no post-post comparison, which is echoed by Sanfey et al. (2011) and Straus et al. (2013). This is problematic since Dannels et al. (2008) suggest that the impact of any such programme is not likely to be immediate and Hirst et al. (2004) state that it is impossible to ascertain the effect without a natural lag between learning and application to one's leadership context. As an example, Abrell et al. (2011) report that the leadership development programme that they studied did not improve the perceived effectiveness of participants' leadership until six months after for supervisees and nine months for supervisors. They explain that the precise mechanisms about how leadership skills are transferred to the workplace and how much time this takes are not yet unexplored (Abrell et al., 2011). This example reflects the importance of post-post measurements. Conversely, Kwamie et al. (2014) report that four out of five teams achieved their clinical outcomes following a leadership programme in the short-term, but those results were not sustained in the medium or long-term. Thus, tracking the medium and long-term transfer of leadership learning depends on collecting post-post measures.

Action learning projects are to some extent an exception, since learning is applied as the intervention progresses; however, these projects do not account for the prevailing shortage of post-post measures. Even with action learning projects, it is still valuable to measure to what extent changes are sustained over time by adding a post-post measure. Edmonstone (2013) suggests that relying exclusively on evaluations at the conclusion of the programme exposes them to the euphoric "inevitable glow" that surrounds the end of an experience before the true test of actual application to the workplace happens (p. 149). Furthermore, post-post measures can prompt participants to reflect on how they have developed and applied their learning, which is reported as itself serving as an additional development tool (K. E. Watkins et al., 2011). A different example of why post-post measurements are important is Fernandez et al.'s (2016) finding that participants' in their study's self-reports of leadership competencies following the programme *decreased* in 50 per cent of the competencies after six months compared to immediately following the programme. The authors attribute this to challenges associated with applying their learning to the workplace. Likewise, Sanfey et al. (2011) assert that in their study, participants' perceptions of their ability to take on leadership roles and whether each saw her or himself as a leader decreased significantly from the post to the post-post test. This indicates that it can be difficult to apply leadership skills and behaviours in the long-term, especially if the workplace culture is not receptive to change. This will be discussed further in the conclusions explored section. Both good and three of the four moderate evidence HEE studies included a post-post measure, compared to only two anecdotal studies. Therefore, despite their value in leadership development research, less than half the studies included a post-post measurement to analyse the application to the workplace.

Similarly, less than half the studies (n = 30) included a pre or baseline measurement, including only 3/11 anecdotal HEE studies, compared to 5/6 of the good and moderate evidence studies. Also, only nine studies (9%) combined a pre, post, and post-post measurement. Considering the importance of tracking the long-term application of leadership development compared to a baseline measure as a benchmark, this is surprising and limiting in terms of the usefulness of the studies' findings and conclusions. Also, without a baseline measurement, it can be challenging for participants and other raters to separate the impact of the programme from prior knowledge and capabilities, how their jobs have evolved, and what they have learned from sources other than the intervention in question (K. E. Watkins et al., 2011). Finally, it is interesting that of the HEE studies, only Satiani et al. (2014) measured outcomes during the programme, although doing so can add useful data and concomitantly enable participants to reflect on their own development as the intervention progresses.

Therefore, the majority of studies included post measurements, but frequently without a post-post measurement and often without a pre or a baseline measurement. In terms of overall measurement, the most robust data collection involves multiple raters, multiple types of data, objective indicators of outcomes, and data gathered at a baseline, as the programme transpires, after the programme, and a post-post measurement.

Table 5.18

When Data Was Collected

		MULTI	**HEE**					**Total**
		N = 56	N = 25	**Study Calibre**				N = 72
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
When Data Was Collected	Pre	23 (41%)	8 (32%)	1	3	3	1	27 (38%)
	Baseline	-	4 (16%)	1	1	0	2	-
	Pre or baseline	23 (41%)	11 (44%)	2	3	3	3	30 (42%)
	During	14 (25%)	1 (4%)	0	0	0	1	15 (21%)
	Post	23 (41%)	14 (56%)	0	2	4	8	32 (44%)
	Post-post	26 (46%)	13 (52%)	2	3	6	2	35 (49%)
	Retro post	3 (5%)	1 (4%)	0	0	1	0	3 (4%)
	Retro pre	0 (0%)	1 (4%)	0	0	1	0	1 (1%)
	Post only	13 (23%)	5 (20%)	0	0	1	4	16 (22%)
	Post-post only	8 (14%)	3 (12%)	0	0	2	1	10 (14%)
	Post or PP only	21 (38%)	12 (48%)	0	1	5	6	30 (42%)
	Pre or baseline and P or PP	21 (38%)	11 (44%)	2	3	3	3	29 (40%)
	Pre or baseline and PP	2 (4%)	6 (24%)	2	2	2	0	7 (10%)
	Pre, post, and post-post	8 (14%)	2 (8%)	0	0	1	1	9 (13%)

The table above depicts a breakdown of when data was collected in the two reviews.

5.14 Outcome Metrics – HEE Only

It has already been mentioned that many authors confused programme-wide outcome metrics that are applied to assess all participants and outcomes and benefits reported individually by participants. What follows below is a treatment of both, to the best accuracy possible given the often-ambiguous reporting. This section relates only to HEE studies, whereas the following section discusses MULTI and HEE. As depicted in Table 5.19 and Table 5.20 below, nearly all the studies (n = 21) included Post-Programme Evaluations (PPEs). The other most commonly used measurements were self-reported increased skills (n = 13, Level 2b), knowledge (n = 10, Level 2b), behaviours (n = 9, Level 3a), and promotions (n = 7, Level 3b). As suggested earlier, benefit to patients (Level 4b) and organisational benefits (Level 4a) are nearly absent, with only six studies incorporating each. As depicted in Table 5.21 below, 100 different outcome metrics were utilised for Levels 1 – 3 and yet only five were Level 4b outcomes. Even with the 21 PPEs taken out, this is still a heavy weighting in favour of individual outcomes. Surprisingly, self-reported benefit to patients, which is a weak measure but is closer to what is needed than individual development exclusively, was only incorporated once. As mentioned in the introduction, there is convincing evidence that leadership interventions can positively impact on patient outcomes (Strasser et al., 2008;

'Husebø & Akerjordet, 2016; Kunzle, Kolbe, & Grote, 2010; Weaver, Dy, & Rosen, 2014). The Husebø and Akerjordet (2016) review and that by Rosenman et al. (2014) provide further evidence of clinical outcome measures that have been incorporated in studies, such as length of stay in hospital, clinical error rates, and mortality/patient survival rates. As mentioned previously, of the six good and moderate evidence HEE studies, only one included either a 4a or 4b measurement outcome. In spite of this existing evidence, few studies target outcomes at these most vital levels.

Similarly, two very credible organisational outcomes, developing and implementing a new programme and policy change, were only used once each. There were other notable absences of organisational outcomes that have been incorporated successfully in other studies, such as workplace satisfaction reports, retention of staff, staff absenteeism (Doran et al., 2004; Gagnon et al., 2006; Hayes, 2007; Jeon et al., 2013; Artz, Goodall, & Oswald, 2016), meeting or exceeding organisational goals, and economic outcomes, such as the money saved by decreased absenteeism. One study offered perceived costs saved by having an in-house versus an external programme, that by MacPhail et al. (2015), but not of effective leadership after a programme. Retention was used in three studies, but as an individual outcome in terms of intervention participants remaining at their same place of work, not an organisational outcome referring to the overall retention of staff in their department or division.

Lastly, only one study, McGurk (2010), enabled participants to set their own outcome metrics. Although there are some drawbacks to this, including possibly lessened credibility compared to a validated instrument and the fact that it would be unlikely that researchers could make comparisons of identical terms among the entire participant population, there are advantages as well. First, individually-selected outcomes can spring from personal 360-degree feedback or performance reports or needs in their specific role or organisation, which could make them more relevant and useful to participants than standardised metrics. Furthermore, in terms of clinical outcomes, participants are often from different clinical specialties or have differing realms of influence, seniority-wise, thus finding a common clinical metric that applies to all participants would not be possible. A potential progression is for organisations to collect participant-selected individual goals in one iteration of an intervention and perhaps impose common ones as programme-wide outcome metrics in succeeding instalments, while still allowing participants to select their own. Another option for the second step would be to offer future participants examples of effective outcomes from which they can choose those best suited to their context. The role of individually-selected goals is worthy of further exploration, especially given how well it fits with the principles of adult learning. Therefore, despite the numerous outcome metrics utilised in the 25 included HEE studies, the bulk of them were at Levels 1 – 3 and focused on the individual, which leaves a remarkable paucity of those at the organisational (4a) and particularly benefits to patients (4b) levels.

Table 5.19

HEE Outcome Metrics (1/2)

Kirkpatrick Outcome Levels	Feature (N = 25)	n (%)	Evidence			
			Good (2)	Moderate (4)	Limited (8)	Anecdotal (11)
Level 1	Post-Programme Evaluations	21 (84%)	0	3	7	11
Level 2a (subj.)	Increased aspirations to lead	6 (24%)	1	1	3	1
	Increased confidence	6 (24%)	0	1	4	1
	Improved self-awareness	2 (8%)	0	0	1	1
	Increased commitment	1 (4%)	0	0	1	0
	Increased engagement	1 (4%)	0	0	1	0
	Increased leadership capacity	1 (4%)	0	0	1	0
	Level total:	**17**	1	2	11	3
Level 2b (subj.)	Increased skills	13 (52%)	0	1	5	7
	Increased knowledge	10 (40%)	1	3	4	2
	Level total:	**23**	1	4	9	9
Level 2b (obj.)	Increased knowledge tests results	1 (4%)	0	0	0	1
	Level total:	**1**	0	0	0	1
Level 3a (subj.)	Increased leadership behaviours	9 (36%)	0	1	5	3
	Positive impact on their careers	2 (8%)	0	1	0	1
	Have taken on more responsibility	1 (4%)	0	0	1	0
	Level total:	**12**	0	2	6	4

Table 5.20

HEE Outcome Metrics (2/2)

Kirkpatrick Outcome Levels			Evidence			
	Feature (N = 25)	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**
Level 3b (obj.)	Promotions	7 (28%)	1	1	5	0
	Have taken on a leadership role	5 (20%)	0	3	1	1
	Improved MSF pre and post	3 (12%)	2	1	0	0
	Retention (individual)	3 (12%)	0	0	3	0
	Awards won	2 (8%)	0	1	1	0
	Research publications	2 (8%)	0	2	0	0
	Colleagues' feedback on behaviour changes	1 (4%)	0	0	1	0
	Grants earned	1 (4%)	0	1	0	0
	Increased committee involvement	1 (4%)	0	1	0	0
	Improved supervisor's rating of increased leadership behaviour	1 (4%)	0	0	1	0
	Level total:	**26**	3	10	12	1
Level 4a (subj.)	Developing and implementing a new programme	1 (4%)	0	0	1	0
	General organisational benefits	1 (4%)	0	0	1	0
	Policy changes	1 (4%)	0	0	1	0
	Strengthening organisational relationships	1 (4%)	0	0	1	0
	Level total:	**3**	0	0	3	0
Level 4b	Having implemented action learning projects	4 (16%)	0	1	2	1
	Self-report of benefits to patients	1 (4%)	0	0	1	0
	Level total:	**5**	0	1	3	1
Other	Having joined a mentoring network	1 (4%)	0	0	1	0

The tables above demonstrate the outcome metrics offered in the various HEE studies according to the Kirkpatrick levels.

Table 5.21

Outcome Metrics According to the Kirkpatrick Model Totals

Kirkpatrick Measurement Outcomes			Evidence			
		n	Good (2)	Moderate (4)	Limited (8)	Anecdotal (11)
Level 1	Participant satisfaction	21	0	3	7	11
Level 2a	Attitudes or perceptions	17	1	2	11	3
Level 2b	Knowledge and skills (subj.)	23	1	4	11	9
Level 2b	Knowledge and skills (obj.)	1	0	0	0	1
	Level 2 total:	**41**	2	6	22	13
Level 3a	Behaviour (subjective)	12	0	2	6	4
Level 3b	Behaviour (objective)	26	3	10	15	1
	Level 3 total:	**38**	3	12	21	5
	Level 1 - 3 total:	**100**	5	21	50	29
Level 4a	Organisational change	4	0	0	4	0
Level 4b	Benefits to patients	5	0	1	3	1
	Total:	**109**				

The table above provides a summary of the frequency of each Kirkpatrick level of outcome metrics in the HEE included studies.

5.15 Reported Outcomes and Benefits

As is evident in Table 5.22 and Table 5.23 below, the reported outcomes and benefits are quite similar to, but more numerous than, the HEE outcome metrics. To qualify for this category, items need only have appeared in a study based on as few as one participants' report of a benefit or outcome, as contrasted to metrics used by researchers or providers that are applied to the entire participant population.

The most common again were PPEs (n = 43, Level 1), self-reported increased behaviours (n = 26, Level 3a), knowledge (n = 25, Level 2b), and skills (n = 24, Level 2b). Other common outcomes include increased self-awareness (n = 19, Level 2a), confidence (n = 14, Level 2a), as well as increased leadership capability/ competence/ capacity/ effectiveness/ self-efficacy (n = 12, Level 3a). It is interesting that self-awareness was reported as a benefit in eight HEE studies even though it was only used as an outcome metric in two, which suggests it could be a valued benefit of programmes since participants volunteered it often. It is noted that only one study mentioned further interest in leadership development as a benefit, considering many advocates of placing interventions in the larger context of through-career progression. In HEE, there are substantial increases in reported outcomes at the 2a (attitudes and perceptions) level, from 17 outcome metrics to 39 reported outcomes and benefits and the

2b (knowledge and skills) level, from 19 outcome metrics to 40 reported outcomes. Unfortunately, the Level 4b increase from outcome metrics to reported outcomes and benefits in HEE was only from four to nine, with six of those being implementing action learning projects. Only three MULTI studies reported 4b outcomes, two of which overlapped with HEE studies. This raises the question of the ultimate outcome in non-healthcare domains, which was addressed in the justification of the sample choice of doctors. What is equally surprising and not domain-specific is that MULTI studies reported only eight 4a outcomes in 56 included studies. The only good or moderate evidence HEE study to include 4a or 4b benefits is that by Patel et al. (2015). Therefore, the recurring observation about the shortage of outcomes at the organisational and clinical level is further reinforced by the small number of 4a and 4b reported benefits.

It is surprising that in only 35 per cent of studies (n = 25) did participants report an increase in knowledge and in only 33 per cent of studies (n = 24) did they claim to have increased their skills. This does not confirm that there was no increase in these two areas in the remainder of the studies, but rather that they were not reported. It is equally surprising that only six studies included external raters' reports of increased leadership behaviours, which contributes to a previously mentioned trend of relying on self-reports. It has been mentioned that some leadership skills are less tangible and therefore more challenging for outsiders to observe and rate (K. E. Watkins et al., 2011); however, this should not preclude the effort altogether.

On a different note, research publications and grants and awards won were mentioned as Level 3b outcome, which raises the question about whether those are valuable indicators of leadership or not, which MacPhail et al. (2015) extend to promotions. It is further surprising that no studies included meeting team or organisational goals as an outcome metric or benefit and that only one, that by Bowles et al. (2007), utilised meeting individual performance goals. The only HEE study that reported a decrease in the number of outcome metrics compared to reported benefits was that by Malling et al. (2009), since although participants' self-ratings increased, their MSF ratings stayed the same. Thus, Level 3b was used as an outcome metric, but not as a reported outcome. Notably, only two studies included MSF pre and post-post ratings and only three studies mentioned meeting individual professional goals as an outcome, though these are both commonly used by human resources departments in organisations. Similarly, not one study mentioned improved performance reviews, pre and post, even though in the Suutari and Viitala (2008) study of nearly 900 senior business leaders, 39 per cent reported having regular performance evaluations, as just one example of a common organisational practice that can provide useful data for leadership outcome metrics. It is likely that many if not most organisations set and review these kinds of goals routinely; however, this data is not being included in leadership development programme outcome metrics.

Only two studies cited increasing organisational capacity as a post-programme benefit and the same number listed launching a new project or programme, which reflects again the participant-centred nature of much of the research, as well as a key advantage of action learning projects. There was only one unique study in MULTI that listed implementing action learning as an outcome, though it was used as a developmental activity in 19 studies. Surprisingly, there is only one mention of each of the following outcomes: "have improved practice in healthcare", "having used innovative approaches to improve healthcare delivery", and "self-reports of providing better healthcare to patients". This is yet another example of the paucity of 4b outcomes in healthcare leadership development, which is the ultimate outcome in this domain. This lack of outcomes beyond the individual-level is not restricted to healthcare, since in the combined MULTI and HEE set of outcomes, the following Level 4 outcomes were mentioned in only one unique study each: policy changes, increased financial performance, staff retention, and provided better healthcare to patients. Hayes (2007) is one of the few to provide an excellent combination of measures of participants' capability levels, along with key organisational performance indicators. This study provides more convincing data than Berg and Karlsen's (2012) article where participants' ratings of their own self-efficacy are put forward as an indication of organisational impact. Ardts et al. (2010) state that to fully understand the effectiveness of programmes, a wide variety of outcomes and several levels of analysis are needed. Therefore, the number of reported benefits is similarly weighted in favour of individual outcomes and there is a problematic lack of credible benefits at the 4a (organisational) and 4b (benefit to patients) levels.

Table 5.22

Reported Benefits and Outcomes (1/2)

		MULTI	HEE					Total
		N = 56	N = 25	Study Calibre				N = 72
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
Level 1	Post-Programme Evaluations	31 (55%)	21 (84%)	0	3	8	10	43 (60%)
Level 2a (subj.)	Improved self-awareness/ personal development	16 (29%)	8 (32%)	0	0	4	4	19 (26%)
	Increased confidence	8 (14%)	9 (36%)	0	2	3	4	14 (19%)
	Increased aspirations to lead	4 (7%)	7 (28%)	1	2	3	1	9 (13%)
	Increased commitment	2 (4%)	2 (8%)	0	0	2	0	4 (6%)
	Increased engagement	2 (4%)	4 (16%)	0	0	2	2	4 (6%)
	Enhanced common identity	2 (4%)	2 (8%)	0	0	0	2	3 (4%)
	Greater appreciation of others' perspectives	3 (5%)	1 (4%)	0	0	0	1	3 (4%)
	Developed a systems view/ a deeper understanding of organisational strategy	2 (4%)	1 (4%)	0	0	0	1	2 (3%)
	Developed their sense of responsibility and ethics	2 (4%)	0 (0%)	0	0	0	0	2 (3%)
	Increased interest in further training/ appreciation for the utility of training	2 (4%)	1 (4%)	0	0	0	1	2 (3%)
	Increased leadership self-identity	2 (4%)	1 (4%)	0	0	1	0	2 (3%)
	Broadening of understanding	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Developed a servant leadership attitude	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Having developed one's own personal leadership style	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Increased appreciation for the value of collaboration	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Increased awareness of leadership styles	1 (2%)	1 (4%)	0	0	0	1	1 (1%)
	Increased capacity to learn	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Increased resilience	1 (2%)	1 (4%)	0	0	1	0	1 (1%)
	Plan to change their approach to patient care	0 (0%)	1 (4%)	0	0	1	0	1 (1%)
	Level total:	**83**	**60**	1	7	25	27	**115**
Level 2b (sub.)	Increased knowledge	17 (30%)	12 (48%)	1	3	5	3	25 (35%)
	Increased skills	17 (30%)	13 (52%)	0	1	5	7	24 (33%)
	Increased leadership capability/ competence/capacity/effectiveness/ self-efficacy	9 (16%)	3 (12%)	0	1	2	0	12 (17%)
	Developed communication/ negotiation skills	4 (7%)	4 (16%)	0	0	2	2	8 (11%)
	Improved teamwork skills	4 (7%)	2 (8%)	0	0	2	0	6 (8%)
	Developed interpersonal skills	3 (5%)	3 (12%)	0	0	2	1	5 (7%)
	Developed networking skills	2 (4%)	1 (4%)	0	0	1	0	3 (4%)
	Developed a series of tools and practices	2 (4%)	0 (0%)	0	0	0	0	2 (3%)
	Developed mentoring skills	2 (4%)	0 (0%)	0	0	0	0	2 (3%)
	Developed resilience	1 (2%)	1 (4%)	0	0	1	0	2 (3%)
	Developed ideas of improving patient care	0 (0%)	1 (4%)	0	1	0	0	1 (1%)
	Greater ability to manage change	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Improved self-management	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Level total:	**63**	**40**	1	6	20	13	**92**
Level 2b (obj.)	Increased knowledge tests results	0 (0%)	1 (4%)	0	0	0	1	1 (1%)
	Level total:	**0**	**1**	0	0	0	1	**1**

Table 5.23

Reported Benefits and Outcomes (2/2)

		MULTI	**HEE**					**Total**
		N = 56	N = 25	**Study Calibre**				N = 72
Feature		**n (%)**	**n (%)**	**Good (2)**	**Moderate (4)**	**Limited (8)**	**Anecdotal (11)**	**n (%)**
Level 3a (subj.)	Increased leadership behaviours	22 (39%)	10 (40%)	0	1	4	4	26 (36%)
	Networking benefits	8 (14%)	4 (16%)	0	0	2	2	10 (14%)
	Positive impact on their careers	1 (2%)	4 (16%)	0	2	1	1	5 (7%)
	Have taken on more responsibility	0 (0%)	3 (12%)	0	0	3	0	3 (4%)
	Improved leadership effectiveness	2 (4%)	1 (4%)	0	0	1	0	3 (4%)
	Met individual goals	2 (4%)	1 (4%)	0	0	0	0	3 (4%)
	Devised PDPs	1 (2%)	1 (4%)	0	0	0	1	2 (3%)
	Gained experience as leaders	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Have improved practice in healthcare	0 (0%)	1 (4%)	0	0	1	0	1 (1%)
	Level total:	**37**	**25**	0	3	12	8	**54**
Level 3b (obj.)	Promotions	6 (11%)	7 (28%)	1	1	5	0	11 (15%)
	Have taken on a leadership role	1 (2%)	6 (24%)	0	3	2	1	6 (8%)
	Increased leadership behaviours	6 (11%)	0 (0%)	0	0	0	0	6 (8%)
	Research publications	0 (0%)	4 (16%)	0	2	1	1	3 (4%)
	Awards won	1 (2%)	2 (8%)	0	1	1	0	3 (4%)
	Retention (individual)	0 (0%)	3 (12%)	0	0	3	0	3 (4%)
	Colleagues' feedback on behaviour changes	1 (2%)	1 (4%)	0	0	1	0	2 (3%)
	Improved MSF pre and post	0 (0%)	2 (8%)	2	0	0	0	2 (3%)
	Improved supervisor's rating of increased leadership behaviour	1 (2%)	1 (4%)	0	0	1	0	2 (3%)
	Increased committee involvement	0 (0%)	2 (8%)	0	1	1	0	2 (3%)
	Grants earned	0 (0%)	1 (4%)	0	1	0	0	1 (1%)
	Level total:	**16**	**29**	3	9	15	2	**41**
Level 4a (subj.)	General organisational benefits	1 (2%)	2 (8%)	0	0	2	0	2 (3%)
	Having increased organisational capacity	1 (2%)	2 (8%)	0	0	2	0	2 (3%)
	Having launched a new project/programme	1 (2%)	2 (8%)	0	0	1	1	2 (3%)
	Strengthening organisational relationships	0 (0%)	2 (8%)	0	0	2	0	2 (3%)
	Developed a common language	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Increased financial and market performance of the organisation	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Increased retention	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Policy changes/ service improvements	1 (2%)	1 (4%)	0	0	1	0	1 (1%)
	Raised the profile of the department	1 (2%)	0 (0%)	0	0	0	0	1 (1%)
	Level total:	**8**	**9**	0	0	7	1	**13**
Level 4b (obj.)	Having implemented action learning projects	3 (5%)	6 (24%)	0	1	4	1	7 (10%)
	Having launched a new patient safety project	0 (0%)	1 (4%)	0	0	0	1	1 (1%)
	Having used innovative approaches to improve healthcare delivery	0 (0%)	1 (4%)	0	1	0	0	1 (1%)
	Self-reports of providing better healthcare to patients	0 (0%)	1 (4%)	0	0	1	0	1 (1%)
	Level total:	**3**	**9**	0	2	5	2	**10**
Other	Having joined a mentoring network	1 (2%)	1 (4%)	0	0	1	0	2 (3%)

Above are two tables depicting the breakdown of the reported outcomes and benefits according to the Kirkpatrick levels for the two reviews.

5.16 Exceptions and Reported Poor Results

Despite the high number of studies that describe positive outcomes, there are notable exceptions. To start, as mentioned earlier, it is very interesting that 100 per cent of the studies of leadership development for physicians analysed by Frich et al. (2014) reported positive outcomes, although only two of 45 studies included objective data or quality indicators. The authors propose that the former fact is the case because of publication bias; however, the findings listed below are in some ways more helpful than studies whose authors simply described the benefits of the programme that they investigated. For example, in the Hayes (2007) study, some participants reported lower follow-up scores than their initial scores. The authors speculate that this is possibly a result of participants realising that their skills were weaker than they originally thought. Edmonstone's (2009) evaluation of a clinical leadership programme concedes that it only partially met its objectives, due, it is suggested, to unclear expectations, an inadequate selection process, and a lack of follow up measures to ensure the application of learning. Leslie et al. (2005) noted some decreases in self-reported skill change from baseline to a post-measure, which the authors ascribe to respondents being most cognisant of skills that they wanted to develop further. D'Netto, Bakas, and Bordia (2008) conclude the discussion of their survey of top Australian organisations in 18 different industries by asserting that "management development effectiveness in Australia is mediocre" (p. 11).

Malling et al.'s (2009) analysis of healthcare programmes in Denmark found no statistical difference between the experiment and control groups in a year after a leadership intervention, measured by comparison to a baseline figure, which they speculate may have been caused by a lack of organisational support and culture. McGurk's (2010) study of UK public sector social services middle managers asserts that there was little evidence that the leader development programme had any substantial or long-term impact on the organisation, which the author suggests might be due the programme's lack of anchorage in a specific business context. As mentioned earlier, Abrell et al. (2011) found that the German leadership development programme they analysed led to increases in supervisees' perceptions of transformational leadership six, nine, and twelve months later, but *not* at three months later. They contend that the reason for this is that a certain amount of time may be needed for complex leadership skills to become evident. One critic referred to this as an extraordinary example of post-hoc rationalisation. Ely et al. (2010) say that the effects of leadership interventions like coaching take years to ascertain. Pless, Maak, and Stahl's (2011) study of executives found no evidence of skill development and changes in behaviour in 40 per cent of the cases and Kwamie, Djik, and Agyepong's (2014) analysis of district healthcare managers in Ghana concedes that the leadership development programmes did not appear to contribute to the main goal of enhanced systems thinking. They further explain that despite an initial short-term achievement of clinical outcomes by healthcare teams, the results failed to maintain over the medium and long-term. Participants attribute this to a rigid organisational hierarchy, resulting in a post-programme return to its original equilibrium. Therefore, while many studies describe positive capability development and improved performance outcomes, there are several that indicate that programmes or interventions were ineffective. The discussions in the latter group are very helpful to direct future research and practice. This generalisation also further echoes the need for more research to be done, since many of the studies that included positively rated programmes do not address key variables or provide a deep or complete enough analysis, while others admit that their programmes were unsuccessful.

One final point is that the preparation of this section of examples and explanations of why leadership development programmes failed to meet expectations is what gave rise to the desire to pursue this topic further. The SEA methodology allows for this through its multi-level iterative analytical process, which manifested itself in this instance as leading to the second conclusion explored concerning the conditions that enable the successful transfer of leadership learning following programmes.

5.17 Relationships Among Variables – HEE Only

It has been mentioned previously that none of the extant systematic literature reviews analysed the relationships among the variables pertinent to leadership development beyond just in a descriptive analytical way. It is acknowledged that the bivariate linear regression analyses undertaken for HEE are based on a limited sample size, however, there is support defending the use of small samples (Fiedler & Kareev, 2010; J. P. Stevens, 2012) mentioned earlier and it was decided that testing this approach was necessary to enhance the review's transparency and credibility. As stated earlier, it was anticipated that the margins would be higher and that non-significant results would present; however, this analysis served the purposes of experimenting with this approach and attempting to add credibility to the investigation.

As already described, the explanatory variables (x axis) were: the MERSQI groupings, programme length (≤1 week; 1 month – 10 months; 1 year; >1 year), the Kirkpatrick outcome levels (Level 1 – 3a; 3b; 4a; 4b), and key developmental activities (simulations, 360s, lectures, action learning, case study analysis, and coaching). The dependent variables (y axis) were: the data type, methodology (case study), data collection methods (questionnaires and interviews), sample size, physician-only or interdisciplinary, selection criteria (applied and selected, nominated, and volunteered), faculty (internal or both internal and external), location (in-house or external), programme length, developmental activities, and the Kirkpatrick outcome levels.

The highlights of the results for each pairing are presented below in Table 5.24, Table 5.25, and Table 5.26, while the full set of results is included in the appendix on page 381. The PRISMA guidelines recommend that all analyses conducted be reported, not just those that are statistically significant, to avoid selective outcome reporting bias (Liberati et al., 2009). The tables below list the p value, R-Squared value, and whether the correlation is statistically significant at p = .05 and at p = .01 for each of the pairings.

5.18 Results of the Statistical Analysis – HEE Only

Despite the small sample and expectations of non-significant results, there were a number of surprises that will be highlighted below.

In terms of the **MERSQI groupings** (x axis), there was a *negative* correlation found to be statistically significant between the MERSQI groupings and qualitative only data collection (p = 0.0049, R-Sq = 0.2962), as well as a statistically significant correlation between the MERSQI groupings and collecting quantitative data only (p = 0.0020, R-Sq = 0.3463). These are partly attributable to the MERSQI instrument ascribing a higher score to objective data, but, in a separate category, higher outcome scores were not dependent on quantitative data. Not one of the good or moderate studies collected qualitative data only, whereas 7/11 anecdotal studies did. Conversely, both good evidence studies collected quantitative data only; whereas only one anecdotal study did. This confirms observations made earlier in the design section of the findings.

There was a *negative* correlation found to be statistically significant between the MERSQI groupings and Kirkpatrick outcomes 1 – 3a only (p = 0.0144, R-Sq = 0.2333). Neither the good nor the moderate evidence studies restricted themselves to these; however, 8/11 anecdotal studies did. This is partly attributable to the MERSQI outcomes rating preference for higher Kirkpatrick levels, along with the objective data requirement of 3b outcomes.

There was a statistically significant correlation found between the MERSQI groupings and Kirkpatrick Level 3b outcomes (p = 0.0022, R-Sq = 0.3404). This is to be expected given that it was a criterion for the good and moderate evidence studies categories, so naturally all of the studies in these categories included a Level 3b measurement, whereas only 1/11 of the anecdotal studies did.

In terms of the **programme length**:

A *negative* correlation between the programme length C grouping (1 year) and physician-only samples was found to be statistically significant (p = 0.0115, R-Sq = 0.2471), since of the six studies that were a year long, only one was physician-only (17%), compared to 14/19 studies (74%) of other lengths that were physician-only. Thus, the majority of year-long programmes in HEE featured interdisciplinary samples.

The correlation between the programme length C grouping (1 year) and Kirkpatrick Level 4a outcomes was found to be statistically significant (p = 0.0004, R-Sq = 0.4298), since of the six studies that were a year long, four reported organisational outcomes. Conversely, of the five studies that featured Level 4a outcomes, four of the them were a year in duration. This should be considered alongside the correlation between Level 4a outcomes and action learning projects to be described shortly.

The correlation between the programme length D grouping (>1 year) and interventions featuring internal faculty was found to be statistically significant (p = 0.0217, R-Sq = 0.2087). Of the four studies that were longer than a year, three featured internal faculty. One possible explanation for this is that internal faculty tend to cost less than external. Another possibility is that given that three of four longer programmes featured residents samples, it is perhaps easier to identify internal senior leaders for them compared to an intervention for executive-level leaders.

In terms of the **Kirkpatrick levels**:

There was a statistically significant correlation found between Kirkpatrick Level 1 – 3a outcomes only and qualitative data only (p = 0.0068, R-Sq = 0.2778), but not quantitative data only or both. Although it is possible that objective outcomes can be used in reference to an increase of knowledge and skills (Level 2b), the latter tend to generally be self-reports. Of the seven studies that were restricted to Kirkpatrick Levels 1 – 3a, five collected only qualitative data. The tabulations of outcomes for the different MERSQI groupings are listed above.

There was a statistically significant correlation found between Kirkpatrick Level 1 – 3a outcomes only and sample size (p = 0.0407, R-Sq = 0.1932). Of the seven studies that were restricted to Kirkpatrick Levels 1 – 3a, six were anecdotal and the mean sample size is 26.3, compared to the overall mean of 72.5.

There was a *negative* correlation found between Kirkpatrick Level 3b outcomes and qualitative data only that was statistically significant (p = 0.0129, R-Sq = 0.2402). Of the 15 studies that included a Level 3b outcome, only one used qualitative data only (7%). While it is possible to identify Level 3b outcomes through qualitative reports from other raters, for the most part, Level 3b outcomes are verified using quantitative data.

There was a correlation found between Kirkpatrick Level 3b outcomes and both qualitative and quantitative data was statistically significant (p = 0.0428, R-Sq = 0.1667). Of the 15 studies that included a Level 3b outcome, eight collected both types of data (53%). Of the 11 studies that featured mixed methods, eight (73%) also included Level 3b outcomes.

There was a *negative* statistically significant correlation found between Kirkpatrick Level 4a outcomes and physician-only samples (p = 0.0011, R-Sq = 0.375). Of the five studies that included a Level 4a outcome, not one of them was physician-only. This does not suggest that physician-only programmes cannot lead to Level 4a outcomes, but adds reinforcement to the effectiveness of interdisciplinary programmes.

There was a statistically significant correlation found between Kirkpatrick Level 4a outcomes and conducting interviews (p = 0.0359, R-Sq = 0.1776). Of the five studies that included a Level 4a outcome, three included interviews. As stated earlier, not one of these was a good or moderate evidence study.

There was a statistically significant correlation found between Kirkpatrick Level 4a outcomes and sample size (p = 0.0115, R-Sq = 0.2789). The overall mean sample size was 72.5; whereas the five studies that reported a Level 4a outcome featured samples of 15, 70, 125, 200, and 210 for a mean of 124. As mentioned previously, only two of five studies included Level 4a outcome metrics versus reported outcomes and with a larger sample size, there is a greater likelihood of a wider range of reported outcomes.

There was a statistically significant correlation found between Kirkpatrick Level 4a outcomes and action learning projects (p = 0.0085, R-Sq = 0.2647). Level 4a and 4b outcomes tend to come consequently with action learning projects when they involve initiatives in a clinical setting. In fact, all but one study that reported Level 4a outcomes (n = 5) featured action learning. The explanation for the exception is that in the Cherry et al. (2010) study the projects had not yet been implemented.

There was a statistically significant correlation found between Kirkpatrick Level 4a outcomes and coaching (p = 0.0359, R-Sq = 0.1776). Of the five studies that listed a Level 4a outcome, three featured coaching and only three of the six studies that offered coaching did not report an organisational benefit. This is partly attributable to the fact that coaching often reinforces action learning projects on development programmes and action learning projects are correlated with Level 4a outcomes, as was just explained.

There was a statistically significant correlation found between Kirkpatrick Level 4a outcomes and Kirkpatrick Level 4b outcomes (p = 0.0359, R-Sq = 0.1776). As mentioned previously, there were five studies that included a Level 4a outcome and six that reported a Level 4b outcome and three studies included both. Of the latter group, all involved action learning.

There was a statistically significant correlation found between Kirkpatrick Level 4b outcomes and combining qualitative and quantitative data collection (p = 0.0491, R-Sq = 0.158). Of the six studies that reported a Level 4b outcome, five included mixed methods.

There was a statistically significant correlation found between Kirkpatrick Level 4b outcomes and lectures (p = 0.0378, R-Sq = 0.1744). Of the six studies that reported a Level 4b outcome, four included lectures.

There was a statistically significant correlation found between Kirkpatrick Level 4b outcomes and action learning projects (p = 0.0010, R-Sq = 0.3824). Of the six studies that reported a Level 4b outcome, all but one integrated action learning projects into their intervention and, as mentioned previously, and only three of the eight studies that featured action learning did not reported 4b outcomes.

There was a statistically significant correlation found between Kirkpatrick Level 4b outcomes and coaching (p = 0.0035, R-Sq = 0.3152). Of the six studies that included a Level 4b outcome, five (83%) featured coaching and likewise, of the six studies that incorporated coaching, only one did not report a Level 4b benefit.

In terms of the **developmental activities**:

There was a statistically significant correlation found between simulations and internal faculty (p = 0.0206, R-Sq = 0.2119). Five of the nine studies (56%) that included simulations featured internal faculty. This is possibly because healthcare simulations often develop, among others, technical clinical skills led by experts, who are likely to be clinicians themselves.

There was a statistically significant *negative* correlation found between lectures and case study as a methodology (p = 0.0326, R-Sq = 0.1835). This is a slightly misleading finding, since although only two of the eight studies that included lectures used a case study methodology, four of them left it unclear whether they used a case study or action learning design, thus the number might be as high as six of eight. Likewise, of the 14 studies that used case study as their design, only two included lectures.

There was a statistically significant correlation found between lectures and in-house programmes (p = 0.0476, R-Sq = 0.1667). All eight studies that included lectures were in-house programmes, which is surprising, since many external programmes are run through business schools and tend to be lecture-centric (Frich et al., 2014; Rosenman et al., 2014).

There was a statistically significant correlation found between lectures and action learning projects (p = 0.0007, R-Sq = 0.3999). Six of the eight studies that included lectures also included action learning.

There was a statistically significant correlation found between lectures and case study analysis (p = 0.0068, R-Sq = 0.2778). Five of the eight studies that included lectures also included case study analysis used internal faculty.

There was a statistically significant correlation found between case study analysis and internal faculty (p = 0.0068, R-Sq = 0.2778). Four of the seven studies that included case study analysis featured internal faculty.

There was a statistically significant correlation found between action learning projects and coaching (p = 0.0378, R-Sq = 0.1744). Of the six studies that featured coaching, four also included action learning projects.

One comment on the previous relationships is necessary. Of the ten studies whose programmes involved lectures, action learning, or coaching, four were a year-long, two were longer than a year, two were 8 – 9 months, and none was shorter than a month. Thus, it is possible that these correlations are showing because longer programmes tend to offer more developmental activities, rather than that the activities have a reliant or beneficial relationship with each other.

Finally, there was a statistically significant correlation found between coaching and collecting both quantitative and qualitative data (p = 0.0491, R-Sq = 0.158). Of the six studies that included coaching, five collected both types of data.

In terms of a summary, the statistical analysis revealed that there were statistically significant correlations between high MERSQI groupings and not collecting qualitative data only, collecting quantitative data only, not restricting the outcome metrics to Kirkpatrick 1 – 3a levels, and including a Level 3b outcome. There was also a connection among anecdotal calibre studies, Kirkpatrick Levels 1 – 3a outcomes only, and small participant sample sizes.

Those programmes that reported organisational level outcomes (4a) tended to be longer programmes (a year or more), feature large, interdisciplinary samples, include action learning and coaching, and also report Level 4b outcomes.

The highlights of the statistically significant findings are presented below with the full version in the appendix. The version below is colour-coded: purple represents a statistically significant relationship at p = .05 and green is for p = .01 or lower.

Table 5.24

Bivariate Linear Regression Results Highlights (1/3)

		X Axis											
Y Axis	**Subgrouping**	**MERSQI Grouping** (Good, moderate, limited, anecdotal)		**Programme Length C: 1 year (Y/N)**		**Kirkpatrick Levels: 1 - 3a only (Y/N)**		**Kirkpatrick 3b Behaviour (objective) (Y/N)**		**Kirkpatrick 4a** Organisational (Y/N)		**Kirkpatrick 4b** Benefit to Patients (Y/N)	
		P-value	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**
Data type	Qualitative only (Y/N)	**Qualitative only (Y/N)***		Qualitative only (Y/N)		**Qualitative only (Y/N) ****		**Qualitative only (Y/N)**		Qualitative only (Y/N)		Qualitative only (Y/N)	
		0.0049	**0.2962**	0.2978	0.04704	**0.0068**	**0.278**	**0.0129**	**0.2404**	0.6836	0.007355	0.3767	0.03414
	Quantitative only (Y/N)	**Quantitative only (Y/N) ****		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)	
		0.002	**0.3463**	0.8241	0.002195	0.2428	0.05883	0.5333	0.01711	0.2281	0.0627	0.1737	0.07897
	Both (Y/N)	Both (Y/N)		Both (Y/N)		Both (Y/N)		**Both (Y/N)**		Both (Y/N)		**Both (Y/N)**	
		0.8605	0.00137	0.4307	0.02724	0.1292	0.09724	**0.0428**	**0.1669**	0.5674	0.01444	**0.0491**	**0.16**
MD's only (Y/N)		MD's only (Y/N)		**MD's only (Y/N)**		MD's only (Y/N)		MD's only (Y/N)		**MD's only (Y/N)**		MD's only (Y/N)	
		0.6243	0.0106	**0.0115**	**0.2473**	0.4878	0.02118	>0.9999	1.37E-35	**0.0011**	**0.377**	0.5851	0.01318
Methodology	Case study (Y/N)	Case study (Y/N)		Case study (Y/N)		Case study (Y/N)		Case study (Y/N)		Case study (Y/N)		Case study (Y/N)	
		0.1734	0.07903	0.1323	0.09577	0.9457	0.00023	0.7547	0.004331	0.2443	0.05846	0.2156	0.06587
Methods	**Interviews** (Y/N)	Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)		**Interviews (Y/N)**		Interviews (Y/N)	
		0.5485	0.01587	0.3196	0.04308	0.7512	0.00448	0.5851	0.01318	**0.0359**	**0.1778**	0.5587	0.0151
Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		**Sample size (# of participants) ***		Sample size (# of participants)		**Sample size (# of participants)**		Sample size (# of participants)	
		0.7286	0.006153	0.5587	0.0151	**0.0407**	**0.1934**	0.8519	0.001787	**0.0115**	**0.2791**	0.8041	0.00315
Faculty	Internal (Y/N)	Internal (Y/N)		Internal (Y/N)		Internal (Y/N)		Internal (Y/N)		Internal (Y/N)		Internal (Y/N)	
		0.7087	0.006184	0.0859	0.123	0.3224	0.0426	0.8630	0.001325	0.1292	0.09724	0.4988	0.02014
Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)	
		0.1990	0.07069	0.1828	0.07583	0.0547	0.1514	0.4878	0.02118	0.5242	0.01788	0.4988	0.02014

*** Statistically significant below the .05 level**

**** Statistically significant below the .01 level**

Table 5.25

Bivariate Linear Regression Results Highlights (2/3)

		X Axis							
Y Axis	**Subgrouping**	**Simulations** (Y/N)		**Lectures** (Y/N)		**Case Study Analysis** (Y/N)		**Coaching** (Y/N)	
		P-value	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**
Data type	Qualitative only (Y/N)	Qualitative only (Y/N)		Qualitative only (Y/N)		Qualitative only (Y/N)		Qualitative only (Y/N)	
		0.9190	0.00047	0.2010	0.07003	0.1004	0.115	0.3767	0.03414
	Quantitative only (Y/N)	Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)	
		0.4259	0.0278	0.0932	0.1178	0.1292	0.09724	0.1737	0.07897
	Both (Y/N)	Both (Y/N)		Both (Y/N)		Both (Y/N)		**Both (Y/N)**	
		0.8012	0.002814	0.8963	0.00077	0.7605	0.00414	**0.0491**	**0.16**
MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)	
		0.1881	0.07409	0.8681	0.001227	0.863	0.001325	0.5851	0.01318
Methodology	Case study (Y/N)	Case study (Y/N)		**Case study (Y/N)**		Case study (Y/N)		Case study (Y/N)	
		0.4039	0.0305	**0.0326**	**0.1837**	0.9457	0.0004	0.7470	0.00463
Methods	**Interviews** (Y/N)	Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)	
		0.2562	0.05568	0.9601	0.00013	0.2681	0.05305	0.3196	0.04308
Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)	
		0.4306	0.03137	0.4962	0.02348	0.1676	0.09301	0.2963	0.05441
Faculty	Internal (Y/N)	**Internal (Y/N) ***		Internal (Y/N)		**Internal (Y/N)**		Internal (Y/N)	
		0.0206	**0.2121**	0.1004	0.115	**0.0447**	**0.164**	0.4988	0.02014
Location (In-house v external)		Location (In-house v external)		**Location (In-house v external)**		Location (In-house v external)		Location (In-house v external)	
		0.1719	0.07961	**0.0329**	**0.185**	0.0547	0.1514	0.7512	0.004458

*** Statistically significant below the .05 level**

Table 5.26

Bivariate Linear Regression Results Highlights (3/3)

		X Axis									
Y Axis	**Subgrouping**	**MERSQI Grouping** (Good, moderate, limited, anecdotal)		**Kirkpatrick 4a** Organisational (Y/N)		**Kirkpatrick 4b** Benefit to Patients (Y/N)		**Lectures** (Y/N)		**Action Learning** (Y/N)	
		P-value	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**
Programme length	1 year (Y/N)	Length: 1 year (Y/N)		**Length: 1 year (Y/N)**		Length: 1 year (Y/N)		Length: 1 year (Y/N)		Length: 1 year (Y/N)	
		0.8960	0.0008	**0.0004**	**0.43**	0.5587	0.0151	0.9393	0.00026	0.2978	0.04704
Developmental activities	**Lectures** (Y/N)	Lectures (Y/N)		Lectures (Y/N)		Lectures (Y/N)		N/A	N/A	N/A	N/A
		0.1852	0.07504	0.6836	0.007355	0.0378	0.1746				
	Action learning (Y/N)	Action learning (Y/N)		**Action learning (Y/N)**		**Action learning (Y/N)**		**Action learning (Y/N)**		N/A	N/A
		0.6559	0.008782	**0.0085**	**0.2649**	**0.0010**	**0.3826**	**0.0007**	**0.4001**		
	Case study analysis (Y/N)	Case study analysis (Y/N)		Case study analysis (Y/N)		Case study analysis (Y/N)		**Case study analysis (Y/N)**		Case study analysis (Y/N)	
		0.3324	0.04089	0.5242	0.01788	0.7512	0.004458	**0.0068**	**0.278**	0.4888	0.02109
	Coaching (Y/N)	Coaching (Y/N)		**Coaching (Y/N)**		**Coaching (Y/N)**		Coaching (Y/N)		**Coaching (Y/N)**	
		0.5485	0.01587	**0.0359**	**0.1778**	**0.0035**	**0.3154**	0.9393	0.00026	**0.0378**	**0.1746**
Kirkpatrick levels	**Only 1 - 3a** (Y/N)	**Kirkpatrick only 1 - 3a (Y/N) ***		N/A	N/A						
		0.0144	**0.2333**								
	3b (Y/N)	**Kirkpatrick 3b (Y/N) ****		Kirkpatrick 3b (Y/N)							
		0.0022	**0.3404**	>0.9999	0.1.156e-035						

*** Statistically significant below the .05 level**

**** Statistically significant below the .01 level**

Above are tables listing the colour-coded highlights of the statistically significant correlations between variables.

A summary of the statistically significant correlated findings of the statistical analysis is depicted in Table 5.28 and Table 5.30 below, which also depict the analysis of whether statistically significant correlations were reciprocal.

When a statistically significant correlation was found, the occurrences of y in x were tabulated and the inverse (the number of occurrences of x (the dependent variable) in y (the explanatory variable)) was tested to examine whether it was a reciprocal correlation. For example, a statistically significant correlation was found between Kirkpatrick Level 4b outcomes and action learning as a developmental activity. Thus, participants who claimed to have impacted patient outcomes (Level 4b) tended to have done so by implementing action learning projects. Notably, of all the HEE programmes that reported a 4b outcome (n = 6), the explanatory variable, five featured action learning projects, the dependent variable. The inverse was also examined: of all programmes that included action learning projects (n = 8), five reported Level 4b outcomes. When the inverse occurrence percentage was 80 per cent or higher or lower than 20 per cent for negative correlations, it was noted below. When appropriate, observations and speculations regarding these findings are made.

Statistically significant correlations are coloured green, negative statistically significant correlations are coloured red, inverse relationship occurrences that were unremarkable because they were less than 80 per cent for correlations and higher than 20 per cent for negative correlations are coloured yellow, and inverse relationship occurrences that were noteworthy because they were higher than 80 per cent for correlations or lower than 20 per cent for negative correlations are coloured orange. Blue and purple are only used for the reader's convenience, but have no deeper significance. Table 5.27 below outlines the colour coding system.

Table 5.27

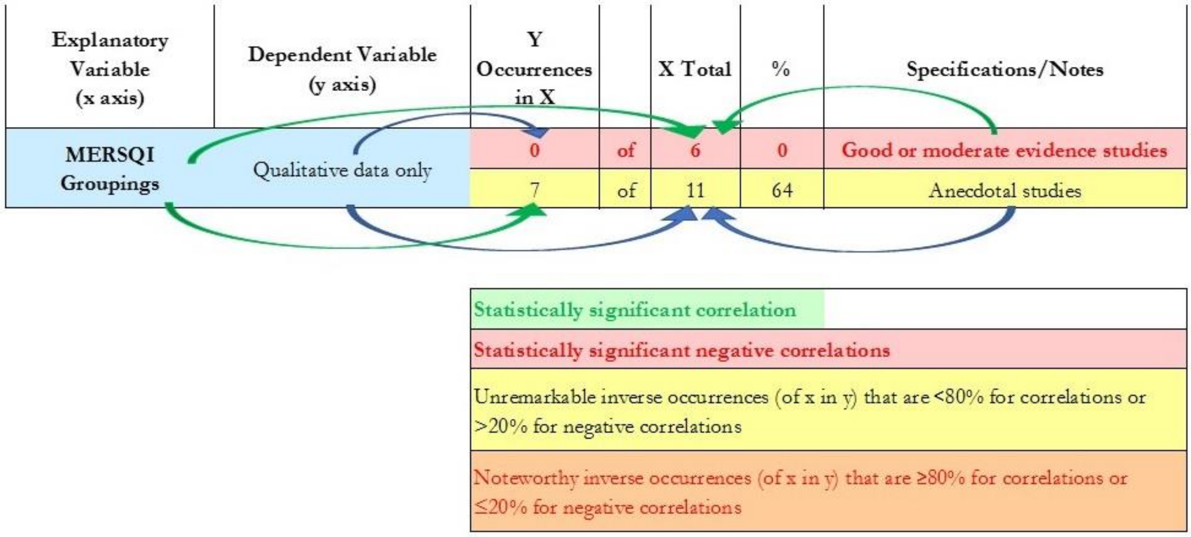
Colour Coding for the Linear Regression Results

Statistically significant correlations	Green
Statistically significant negative correlations	Red
Unremarkable inverse occurrences (of x in y) that are <80% for correlations or >20% for negative correlations	Yellow
Noteworthy inverse occurrences (of x in y) that are ≥80% for correlations or ≤20% for negative correlations	Orange
Colour divides for the reader's convenience	Purple
Colour divides for the reader's convenience	Blue

The table above outlines the colour-coding system used for the summary of the statistical analysis below. Before moving on to the summary, an explanation is provided in Table 5.28 below for the reader’s convenience.

Table 5.28

Explanation of the Summary of Statistically Significant Correlations Table



Explanatory Variable (x axis)	Dependent Variable (y axis)	Y Occurrences in X		X Total	%	Specifications/Notes
MERSQI Groupings	Qualitative data only	0	of	6	0	Good or moderate evidence studies
		7	of	11	64	Anecdotal studies

Statistically significant correlation
Statistically significant negative correlations
Unremarkable inverse occurrences (of x in y) that are <80% for correlations or >20% for negative correlations
Noteworthy inverse occurrences (of x in y) that are ≥80% for correlations or ≤20% for negative correlations

The table above depicts one example of a statistically significant correlation found in the bivariate linear regression analysis, along with an explanation below of what the various components mean. The information above indicates:

- The colour red in the first row indicates that there was a *negative* statistically significant correlation between top two MERSQI groupings (good and moderate evidence studies) and qualitative only data collection
- Following that row along, there were six good and moderate calibre studies (x axis) and, of those, not one featured only qualitative data (y axis)
- The colour yellow in the second row indicates that the analysis of whether the inverse correlation between the two variables was also true was found to be unremarkable. This means that of the 11 anecdotal studies (the inverse of the good and moderate studies), seven (64%) collected only qualitative data

Table 5.29

Summary of Statistically Significant Correlations (1/2)

Explanatory Variable (x axis)	Dependent Variable (y axis)	Y Occurrences in X		X Total	%	Specifications/Notes
MERSQI Groupings	Qualitative data only	0	of	6	0	Good or moderate evidence studies
		7	of	11	64	Anecdotal studies
	Quantitative data only	2	of	2	100	Good evidence studies
		1	of	11	9	Anecdotal studies
	Kirkpatrick Levels 1 - 3a only	0	of	6	0	Good or moderate evidence studies
		6	of	11	55	Anecdotal studies
	Kirkpatrick Level 3b	6	of	6	100	Good or moderate evidence studies
		3	of	11	27	Anecdotal studies
Programme Length C: 1 year	Physician-only samples	1	of	6	17	
		5	of	10	50	
	Kirkpatrick Level 4a	4	of	6	67	
		4	of	5	80	
Programme Length D: >1 year	Internal faculty	3	of	4	75	
		3	of	7	43	
Kirkpatrick Levels 1 - 3a only	Qualitative data only	3	of	7	43	
		3	of	8	38	
	Sample size	4	of	7	57	4/7 did not include sample information and the other sizes were 16, 21, and 53
		-	-	-	-	
Kirkpatrick 3b Behaviour (objective)	Qualitative data only	1	of	15	7	
		1	of	8	13	
	Both qualitative and quantitative data	8	of	15	53	
		8	of	11	73	

Statistically significant correlation

Statistically significant negative correlations

Unremarkable inverse occurrences (of x in y) that are <80% for correlations or >20% for negative correlations

Noteworthy inverse occurrences (of x in y) that are ≥80% for correlations or ≤20% for negative correlations

Table 5.30

Summary of Statistically Significant Correlations (2/2)

Explanatory Variable (x axis)	Dependent Variable (y axis)	Y Occurrences in X		X Total	%	Specifications/Notes
Kirkpatrick 4a Organisational	Physician-only programmes	0	of	5	(0%)	
		0	of	15	(0%)	
	Interviews	3	of	5	(60%)	
		3	of	5	(60%)	
	Sample size	3	of	5	(60%)	3/5 programmes featured samples larger than the review mean of 72.5
		3	of	9	(33%)	
	Programme Length C: 1 year	4	of	5	(80%)	
		4	of	4	(100%)	
	Action learning (developmental activity)	4	of	5	(80%)	
		4	of	8	(50%)	
	Coaching	3	of	5	(60%)	
		4	of	6	(67%)	
	Kirkpatrick Level 4b	3	of	5	(60%)	
		3	of	6	(50%)	
Kirkpatrick 4b Benefit to Patients	Both qualitative and quantitative data	5	of	6	(83%)	
		5	of	11	(45%)	
	Lectures	4	of	6	(67%)	
		4	of	8	(50%)	
	Action learning (developmental activity)	5	of	6	(83%)	
		5	of	8	(63%)	
	Coaching	5	of	6	(83%)	
		5	of	6	(83%)	
Simulations	Internal faculty	5	of	9	(56%)	
		5	of	7	(71%)	
Lectures	Case study methodology	2	of	8	(25%)	
		2	of	14	(14%)	
	In-house programmes	8	of	8	(100%)	
		8	of	18	(44%)	
	Action learning (developmental activity)	6	of	8	(75%)	
		6	of	8	(75%)	
	Case study analysis	5	of	8	(63%)	
		5	of	7	(71%)	
Case Study Analysis	Internal faculty	4	of	7	(57%)	
		4	of	7	(57%)	

🟩 Statistically significant correlation

🟥 Statistically significant negative correlations

🟨 Unremarkable inverse occurrences (of x in y) that are <80% for correlations or >20% for negative correlations

🟧 Noteworthy inverse occurrences (of x in y) that are ≥80% for correlations or ≤20% for negative correlations

The tables above provide a summary of the colour-coded statistically significant correlations among variables, as well as the occurrences of each of the variables in each pairing and the inverse.

5.19 Not Correlated with Credible MERSQI Ratings

There was an unexpectedly high number of variables that were not correlated with high (or low) MERSQI ratings in terms of bivariate linear regression analyses. This is not, however, to suggest that these factors do not matter in leadership development – many likely do – but merely that in this study they did not influence the MERSQI scores notably. They are:

- Methodology, except for experiments, though there was only one
- Methods
- Sample size
- Gender of sample participants
- Level of seniority of sample participants
- Doctors-only versus interdisciplinary samples
- Faculty, internal versus external
- Location, in-house versus external
- Length of the programme. There was also no apparent correlation between length and level of seniority, except for the fact that three programmes for residents were a year or longer, which could be attributable to the fact that residency programmes, like MBA programmes, make it more feasible to establish longer interventions to which participants can commit than attempting something similar with fully qualified professionals who are participating independently
- Level 4a and 4b outcomes. As has been mentioned previously, MERSQI ascribes higher ratings for these outcomes; however, despite that fact, few of the most credible studies attempted to measure at these important outcome levels

Of note in the above, it has been suggested elsewhere that there are different leadership requirements at different levels of the organisation (Mumford et al., 2007; Van Aerde, 2013), which raises questions about how and to what extent the content, programme structure and components, and developmental activities should change according to these various levels.

Finally, the descriptive analysis in MULTI revealed no apparent consistencies among the variables of samples, programme goals, topics, or formats, which means that in as much as there was a range of each of those features overall, there were also no patterns among them, such as most programmes for physicians targeting the same goals or featuring largely similar length of interventions. This feature matches the heterogeneity of these variables cited in Frich et al. (2014). This echoes the need for further research into optimal combinations of programme content and structure (length, location, developmental activities) for various desired learning outcomes and participants.

6 Chapter Six: Conclusions from the Best Available Evidence and Conclusions Explored

This chapter examines the heart of the systematic evidence analysis (SEA) methodology, which is the conclusions of the best available evidence and the two conclusions explored. As has been mentioned throughout, the conclusions of the best available evidence are based primarily on the findings of the HEE review and are tiered in favour of the most credible of the included studies. Thus, the overall thesis conclusions outlined in the following chapter have at their core the findings from the good and moderate evidence HEE studies, in line with the priority of isolating the best available evidence. The second section of this chapter discusses the further, in-depth investigation afforded by the SEA, which involved the two conclusions from the best available studies that emerged as being worthy of further exploration, as mentioned in the earlier chapters. A secondary investigation involved analysing the included studies in detail with these two topics in mind. The first conclusion explored is how Knowles's (1984) principles of adult learning, plus two principles that surfaced as key additions during the analysis, apply to leadership development. Further research unveiled the relevance of Dale's (1969) educational theory, the Cone of Experience, to the topic of the chief function of each developmental activity and how these functions relate to one another. The second conclusion explored is a set of characteristics of the design of leadership development programmes and organisational culture that are thought to significantly influence the success of applying leadership to the workplace following interventions. While this topic has been addressed by other authors, the set that follows is more extensive and provides illustrative examples from studies, particularly from those interventions that reportedly failed. A final stage of the SEA investigation involved connecting related points from the previous sections of this chapter into a coherent sequence of factors that appear to contribute to optimising the effects of leadership development. The chapter concludes with this sequence described in the form of a prototype model of leadership development design and implementation. With this, the attention turns to the conclusions.

6.1 Findings: Conclusions from the Best Available Evidence – Summary

The findings that follow are drawn from the HEE studies that qualify as good or moderate evidence according to their MERSQI grouping, answering the questions "what we know" regarding optimal leadership development and measurement principles and "with what evidence." In a field marred by unclear, equivocal, and at times conflicting information, it is essential to clarify the evidence base for the sake of practice and future research. This foundation will allow practitioners to design programmes with confidence and experiment applying these principles and trialling novel approaches in their own context. This clearly defined evidence base is particularly important given the enormous investment in leadership development and the number of studies identified in this review that purportedly failed to meet expectations. Even though some of the points below may not seem especially innovative, this review has demonstrated that many are not being implemented or measured consistently. One possible reason for the lack of consistency may be the lack of clarity regarding the evidence base. In addition to explicating what is known in the field, it is important to identify both what *is not* clear and those aspects of leadership development for which there is not (yet?) good evidence. For example, there is a common conception that interdisciplinary healthcare leadership development programmes are favourable to physician-only interventions; however, that conclusion has no support in the literature within the parametres of this study. The points below also include explanations of the ways in which certain principles facilitate the application of learning or nuances to this effect, which can also inform design choices. With the information described below, researchers can proceed with better-informed questions and awareness of knowledge gaps and areas of research that have not yet been adequately addressed in order to undertake the kind of work that needs to be done to advance the field.

Below is a summary of the major findings according to the calibre of evidence.

Good evidence

In terms of **outcomes,** there is good evidence that a variety of factors can be improved through leadership development, including at the individual level:

- Increased self-ratings of competence, self-efficacy, and confidence in leadership knowledge and skills
- Increased leadership knowledge
- Increased frequency of observable leadership skills and behaviours
- Positive impact on career progressions, including motivation to pursue leadership roles

Leadership development can also lead to achieved benefits to patients outcomes through improving leadership behaviours and action learning projects, though there was little mention in the five most credible HEE studies of benefits to patients.

In terms of programme **design**, **outcomes-based** interventions that link the goals, content, delivery, and evaluation appear to be optimal.

Knowles's (1984) principles of adult learning can be applied successfully to leadership development interventions, as can capability frameworks.

Certain **factors in the programme design** and the **organisational culture** can significantly affect the effectiveness of the programme and most importantly, the application of leadership development afterward.

Finally, interdisciplinary programmes and physician-only programmes can be effective in enabling leadership outcomes. No study was identified that provided evidence that one is superior to the other.

In terms of **developmental activities**, workshops, followed by videotaped simulations with expert feedback, reinforced by coaching and 360s, can be effective in increasing leadership behaviours.

In terms of **evaluation**, the following components can enhance the credibility of results:

- Collecting objective data through either external raters, statistics, or verifiable outcomes, such as leadership role attainment
- Collecting quantitative data
- Comparing baseline measurements to post and post-post measurements to track relative progress
- Comparing individual performance to those in a balanced control group or a non-intervention population
- Comparing self-ratings to external ratings, including those of colleagues and experts
- Targeting Level 3b (objective behaviour change) outcomes
- Targeting Level 4b (benefit to patients) outcomes
- Evaluating the programme as well as the participants' development
- Including several iterations of a programme to broaden the scope of the results

A full description of each of the best six studies' evaluations is included in the appendix on page 387.

Moderate evidence

In terms of **outcomes**, the following self-reports of individual outcomes were also included:

- Increased clinical skills (Level 3a)
- Increased motivation (Level 2a)

Level 4a (organisational benefits) can also be achieved, particularly through action learning projects.

In terms of **programme design**, there is moderate evidence that the following components can be effective in enabling leadership outcomes:

- Selecting developmental activities and topics that are intentionally directed to workplace needs and clinical experiences to maximise their relevance
- Conducting a needs assessment prior to the intervention
- Embedding leadership interventions in medical residency programmes

In terms of **developmental activities, action learning projects** can be effective especially in terms of enabling outcome attainment at the Kirkpatrick 3b, 4a, and 4b levels.

Mentoring can increase self-ratings of leadership competencies and correlate with career advancement.

Finally, there is **no credible evidence** to inform the selection of other programme components, such as size, length, structure, or location of programmes, faculty characteristics, or optimal sample makeup, such as level of seniority or profession-specific versus interdisciplinary.

When points were raised in the combined included studies concerning these variables, they were added as nuances to the findings, conclusions from the best available evidence, or conclusions explored sections.

As will be explained in the implications for research, given the paucity of good evidence in the field, much of the reported wisdom regarding optimal leadership development and measurement needs to be subject to more credible testing to expand the knowledge base with better evidence.

6.2 Findings: Conclusions from the Best Available Evidence – Full Version

The conclusions in bold below derive from the most credible studies, whereas the points that follow each conclusion are related elaborations or nuances from these and other included HEE studies, MULTI, or the EMD reviews. This follows Cook and West's (2012) recommendation that conclusions relating to similar points should be grouped together. One caveat is that what qualifies as "good evidence" below results from at least one good calibre study, but has not necessarily been corroborated by testing in other contexts. The conclusions are organised below by the MERSQI score groupings of good, moderate, and limited evidence.

6.2.1 Good Evidence (two studies)

(Dannels et al., 2008; Ten Have et al., 2013)

Outcomes

- **Participants have reported increased leadership knowledge, skills, competence, self-efficacy, and confidence** (Bergman, Fransson-Sellgren, Wahlstrom, & Sandahl, 2009; Cherry, Davis, & Thorndyke, 2010; Dannels et al., 2008; Day et al., 2010; Edmonstone, 2009; Fernandez et al., 2016; MacPhail, Young, & Ibrahim, 2015; Miller et al., 2007; Patel et al., 2015; Satiani, Sena, Ruberg, & Ellison, 2014). This echoes previous research suggesting that leadership self-efficacy is said to predict leadership behaviour and effectiveness (Seibert, Sargent, Kraimer, & Kiazad, 2017), as well as to distinguish leaders from non-leaders (McCormick, 2001; McCormick, Tanguma, & López-Forment, 2002).
- **Increased frequency of observable leadership skills and behaviours** (Ten Have et al., 2013). These skills include those related to technical performance, decision-making, communication, teamwork (Ten Have et al., 2013), as well as interpersonal skills (Edmonstone, 2011; Miller et al., 2007; Satiani et al., 2014), communication skills (Fernandez et al., 2016), and networking skills (Edmonstone, 2011; H. Korschun et al., 2010).
- **Leadership development is correlated to a variety of career outcomes** for participants, including promotions (Dannels et al., 2008; Day et al., 2010; Fernandez, Noble, Jensen, & Chapin, 2016; Korschun, Redding, Teal, & Johns, 2010; Kuo, Thyne, Chen, West, & Kamei, 2010; Sanfey, Harris, Pollart, & Schwartz, 2011), increased academic rank and hospital administrative rank (chair or chief) (Day et al., 2010), chairing committees (Day et al., 2010; Korschun et al., 2010), grants secured and publications produced (C. S. Day, Tabrizi, Kramer, Yule, & Ahn, 2010; Kuo et al., 2010), and increased participation in further leadership development programmes (Dannels et al., 2008). It is possible that a latent contributing factor is that some organisations are more likely to promote employees in whom they have invested by way of leadership development; however, the two studies in this review (Dannels et al., 2008; Day et al., 2010) that compared successful applicants to those rejected from leadership programmes showed that motivation alone without leadership development is not enough to progress equally.
- **Participants have reported increased aspirations to lead** following a leadership programme (Dannels et al., 2008; H. Korschun et al., 2010; MacPhail et al., 2015; Patel

et al., 2015). Inversely, aspirations to lead among the subjects in the control group in the Dannels et al. (2008) study who applied and were not accepted actually *decreased* following their rejection.

For both of the previous points, it would be interesting to know if these perceptions were sustained in the medium and long-term and the extent to which this is correlated with organisational culture.

- **Benefit to patients outcomes** (Level 4b) **and quality improvement** can be achieved through increased leadership skills and behaviours (Ten Have et al., 2013). This is also possible through action learning projects. Other examples of successfully improved Level 4b outcomes are length of stay in hospital, clinical error rates, mortality/patient survival rates (Husebø & Akerjordet, 2016; Rosenman et al., 2014), and used innovative approaches to improve healthcare delivery (Nakanjako et al., 2015).

Programme design

- **Outcomes-based** programmes that link the goals, content, delivery, and evaluation seem to be optimal in terms as a design approach (Dale, 1969; Fayolle & Gailly, 2008; Kuo et al., 2010; MacPhail et al., 2015; Nabi, Liñán, Fayolle, Krueger, & Walmsley, 2017). These are maximised when combined with organisational support and follow-up.
- **Capability frameworks** have also been included in the design of programmes (Kuo et al., 2010; Ten Have et al., 2013).
- **Knowles's** (1984) **principles of adult learning** have been incorporated successfully into the design of leadership interventions and are said to enhance their effectiveness (Bearman et al., 2012; MacPhail et al., 2015; Ten Have et al., 2013). These principles appeared to be the most common theoretical support for leadership development identified in the included studies and were consequently featured as the first conclusion explored. The results of this in-depth analysis and examples of how they can be applied to leadership development are explained later in this chapter.
- **Certain factors in the programme design and the organisational culture can significantly influence the effectiveness of the programme** and most importantly, the application of leadership development afterward (Kuo et al., 2010; Kwamie et al., 2014; Malling et al., 2009; K. E. Watkins et al., 2011). As mentioned previously, successful transfer of leadership to the workplace is the chief concern in leadership development (Edmonstone, 2009; Raelin, 2011; K. E. Watkins et al., 2011). This was made most evident by studies of programmes that reportedly failed, as are described below. These

influencing factors were considered so important that they were investigated as the second conclusion explored, which is described in detail below. Other factors, such as drawing clear and explicit connections between curricular goals, relevant programme activities, outcomes, and post-programme measurements, form the basis of the prototype theoretical model also outlined later in this chapter.

- **Interdisciplinary leadership development** (eg physicians, nurses, pharmacists, social workers) **can be effective** (Bergman et al., 2009; Edmonstone, 2011; Korschun et al., 2010; MacPhail et al., 2015; Nakanjako et al., 2015; Rose, 2015), **as can physician-only** programmes (C. S. Day, Tabrizi, Kramer, Yule, Ahn, et al., 2010; Kuo et al., 2010; Patel et al., 2015; Ten Have et al., 2013; Vimr & Dickens, 2013). There was no study however that compared the two. Interdisciplinary programmes are said to be beneficial for generating collaboration and breaking down silos. Korschun et al. (2010) add that participants in their study reported planning to do more interdisciplinary work following the programme and had developed a network to make doing so successful. Only one HEE study, Vimr and Dickens (2013), argued for physician-only programmes. Five of the six good and moderate evidence HEE studies were physician-only.

Developmental activities

- **Workshops followed by videotaped simulations with expert feedback**, reinforced by coaching and 360s, can be effective in increasing observable leadership behaviours, including decision-making and communication skills (Harden, Grant, Buckley, & Hart, 1999; Hunziker et al., 2010; Ten Have et al., 2013; Weaver et al., 2014). Simulations were also utilised to facilitate clinical quality improvement (Patel et al., 2015), improve patient outcomes (Weaver et al., 2014), improve interpersonal skills, and enhance self-awareness (Bearman et al., 2012; Miller et al., 2007), and community leadership (Getha-Taylor & Morse, 2013). Rosenman et al. (2014) stated that simulations were particularly effective for *task-centric* and *directive* objectives and that when teams take part in simulations together, their technical and teamwork skills are enhanced, as is their team effectiveness. Simulations ideally offer repetition, structured reflection, and mastery learning, which they describe as essential elements of effective training (Husebø & Akerjordet, 2016; Rosenman et al., 2014; Ten Have et al., 2013; Weaver et al., 2014). Bearman (2012) suggests that learning occurs best when learners can actively engage in concrete, realistic experiences in authentic contexts, followed by reflective observation on their own and others' experiences. Simulations are used routinely in other domains, such as team sports and the military. The Royal Marines, for example, use combat simulations extensively to train officers in decision-making, problem-solving, leadership, and adaptability under pressure, since the nature of combat requires a constant ability to react to hostile and changing conditions (Woodward, 2016). Similar skills are essential for leaders at all levels (Heifetz, 1994).

It is surprisingly that of the 17 unique studies in this review that included simulations other than Ten Have et al.'s (good evidence), only one included video taping and only 4/16 reported facilitator feedback as a measurement. Therefore, there is good evidence that simulations can improve observable technical, teamwork, decision-making, and leadership skills, and can facilitate self-awareness and adaptability.

Evaluation

The following components can enhance the credibility of results:

- Collecting objective data through either external raters, statistics, or verifiable outcomes, such as leadership role attainment (Dannels et al., 2008; Malling et al., 2009; Ten Have et al., 2013). The limitations of self-ratings alone are best demonstrated by the Malling et al. (2009) study in which although participants' self-ratings increased following the intervention, those of external raters did not. More information regarding self-reports is included in the previous chapter.
- Collecting quantitative data (Dannels et al., 2008; Ten Have et al., 2013)
- Comparing baseline measurements to post and post-post measurements to track relative progress (Dannels et al., 2008; Ten Have et al., 2013). Omitting a post-post measure misses the true test of actual application to the workplace (Edmonstone, 2013), which takes time (Abrell et al., 2011; Dannels et al., 2008; Hirst et al., 2004). This application can also challenging and self-reports of leadership capabilities, confidence, and self-efficacy can decrease from the post to post-post rating (Fernandez et al., 2016; Sanfey et al., 2011), as can clinical outcomes and team performance (Kwamie et al., 2014). Finally, post-post measures can prompt participants to reflect on how they have developed and applied their learning, which is reported as itself serving as an additional development tool (K. E. Watkins et al., 2011).
- Comparing individual performance to those in a balanced control group or a non-intervention population (Dannels et al., 2008; C. S. Day, Tabrizi, Kramer, Yule, & Ahn, 2010)
- Comparing self-ratings to external ratings, including those of colleagues and experts

- Targeting Level 3b (objective behaviour change) outcomes (Dannels et al., 2008; Ten Have et al., 2013)
- Targeting Level 4b (benefit to patients) outcomes (Ten Have et al., 2013)
- Evaluating the programme as well as the participants' development
- Including several iterations of a programme to broaden the scope of the results (Dannels et al., 2008)

6.2.2 Moderate Evidence (four studies)

(Day, Tabrizi, Kramer, Yule, & Ahn, 2010; Kuo et al., 2010; Malling, Mortensen, Bonderup, Scherpbier, & Ringstead, 2009; Patel et al., 2015)

Outcomes

- **Leadership development can lead to increased leadership knowledge** (Level 2b), measured by a pre and posttest (Patel et al., 2015).
- **Organisational** (Level 4a) outcomes can be achieved, particularly through **action learning** projects (Edmonstone, 2011; Husebø & Akerjordet, 2016; Kunzle et al., 2010; Nakanjako et al., 2015; Patel et al., 2015; Rosenman et al., 2014; Strasser et al., 2008; Vimr & Dickens, 2013; Weaver et al., 2014). Examples of these outcomes that authors claim can be improved by leadership development are: developed organisational capacity and policy change (Edmonstone, 2011), workplace satisfaction, retention, lack of absenteeism (Doran et al., 2004; Gagnon et al., 2006; Hayes, 2007; Jeon et al., 2013; Artz, Goodall, & Oswald, 2016)

Programme design

- Programmes in which **developmental activities** and topics are **intentionally selected and directed** to workplace needs and clinical experiences to maximise their relevance (Kuo et al., 2010). As will be explained in the conclusions explored, building on Dale's (1969) model the Cone of Experience, different developmental activities have different chief functions and relations to each other and are best directed toward specific goals. As mentioned with simulations, the more realistic and relevant the components are, the better, in line with the principles of adult learning.
- **Conducting a needs assessment** prior to an intervention has also been introduced to inform the programme design and content, as well as encourage buy-in from key stakeholders whose input is solicited (Kuo et al., 2010; Malling, Mortensen, Bonderup, Scherpbier, & Ringstead, 2009).
- **Embedding a leadership programme in a medical residency programme can work well** without detracting from the participants' clinical education and skill development

(Kuo et al., 2010; Patel et al., 2015). In fact, the participants in the Kuo et al. (2010) study reported an *increase* in their clinical skills as a result of the programme. Blumenthal et al. (2014) assert that designing a tailored programme especially for residents and using internal faculty is optimal. This assertion supports claims by Rose (2015) and others that leadership development is needed at early as well as later stages of physicians' careers (Van Aerde, 2013)

A final point is raised by Kuo et al. (2010) who endorse the assertion in the bullet point below. This is not a finding based on empirical data; however, merely reflections of the authors on why they believe their programme was successful. It is included here because it is reminiscent of similar points raised in other included articles.

- **Exposing participants to as many senior leaders as possible, particularly in different paths to and roles of physician leadership,** through job shadowing, guest lectures by in-house faculty, mentoring, and networking can be beneficial from a role modelling and networking perspective (Hernez-Broome & Hughes, 2004; Leskiw & Singh, 2007; Zenger & Folkman, 2003).

Developmental activities

- **Action learning projects are an effective leadership developmental activity** (Cherry et al., 2010; Day, 2000; Dickey et al., 2014; Edmonstone, 2011; H. Korschun et al., 2010; Kuo et al., 2010; MacPhail et al., 2015; Miller et al., 2007; Mountford & Webb, 2009; Nakanjako et al., 2015; Patel et al., 2015; Suutari & Viitala, 2008; Vimr & Dickens, 2013; Watkins et al., 2011), **especially in terms of enabling outcome attainment at the Kirkpatrick 3b, 4a, and 4b levels**. One of its greatest strengths is that action learning requires leadership development to move beyond the individual-level to team and organisational outcomes. Action learning is also the best approach to toppling the "knowledge transfer problem," a phrase coined by Watkins, Lysø, and deMarrais (2011) to denote the difficulty of applying one's learning to the organisational context. Action learning by definition centres on engineering this transfer, which is a hybrid solution to the on-the-job learning versus formal intervention debate. Some, such as Daimler (2016), favour the former and others, such as Allio (2005), contend that although leadership can be learned and developed experientially, it cannot be taught or acquired cognitively. Action learning essentially encompasses the benefits of both approaches: the real, direct application of leadership and the experience gained by doing so with all the resources offered through formal programmes (Jesuthasan & Holmstrom, 2016). The natural connection between action learning and outcomes likely explains why there was a statistically significant correlation in HEE between action learning and Level 4a and 4b outcomes. One such example is Suutari and Viitala (2008) who report that managers in their study who participated in developmental projects achieved increased levels of leadership competencies (Level 3b). Likewise, Frich et al. (2014) conclude that programmes with action learning at their core, supported by other activities, are likely to yield the largest impact in leadership development for physicians.

Action learning can also be enhanced when offered in conjunction with support interventions and is said to benefit participants in many ways. For example, Watkins, Lyso, and deMarrais (2011) suggest that action learning, along with 360-degree feedback, is highly effective in terms of the application of leadership skills and long-term impact. Similarly, McCauley (2008) asserts in her review that job experiences, which are related to action learning, supported by coaching, are frequently deemed to be the most effective forms of leadership development. Support mechanisms such as coaching and mentoring, which participants can receive while pursuing their action learning projects can enable participants to develop and apply skills and address weaknesses or implementation challenges whilst interacting with team members to successfully implement their projects (Kuo et al., 2010; Miller et al., 2007). There are also networking benefits that go along with action learning as participants collaborate with colleagues and become more familiar with aspects of their work environment, such as protocols and available resources. An example is Cherry et al. (2010), who suggest that mentors were instrumental in enabling participants in their study to complete their projects. In addition to developing knowledge and skills of both leader and leadership development, which Day (2000) outlined in his seminal work, action learning is also said to build human and social capital. Leading such initiatives is reported to heighten participants' personal accountability and visibility within the organisation (Morahan et al., 2010).

These examples make it unsurprising that in HEE, the correlation between action learning and coaching was also found to be statistically significant, as it was between coaching and organisational (Level 4a) and clinical (Level 4b) outcomes. The utility of action learning projects was not found to be dependent on physician-only samples, as half the HEE programmes (n = 4) that included an action learning project featured interdisciplinary samples. Likewise, Frich et al. (2014) state that the small number of studies in their review that reported organisational outcomes featured action learning projects with interdisciplinary teams. In the HEE included studies, action learning was also not restricted to a certain level of seniority. A further observation is that 7/8 HEE studies that included action learning were in-house programmes. Finally, it has been suggested that enabling participants to select their own action learning projects yields positive outcomes. For example, Samani and Thomas (2016), who study pioneering, innovative companies like Disney and Unilever, suggest that when developing leaders select their own action learning projects they create more value for the organisation and are less likely to quit. Therefore, there is moderate evidence that action learning projects, particularly when supported by 360s, coaching, and mentoring, can lead effectively to a variety of outcomes at the individual, team, organisational, and benefit to patients/clients levels.

- **Mentoring for surgeons can increase self-ratings of leadership competencies and correlate with career advancement** (Day et al., 2010). Formal mentoring, involving a long-term relationship with a veteran, is said to be a useful facilitator of leadership development application (Zenger & Folkman, 2003; Hernez-Broom & Hughes, 2004; Leskiw & Singh, 2007) and very effective in contributing to ongoing development, especially when combined with 360-degree feedback (Zenger & Folkman, 2003). McCauley (2008) states in her review that there is evidence that mentees gain organisational and technical knowledge, develop technical, interpersonal, time management, and self-management skills, and report increased self-confidence. What her "evidence" is and its credibility is unclear, though she references a review of mentoring best practice literature by Finkelstein and Poteet (2007). Mentoring was also mentioned in several included studies as supporting other interventions, such as 360s and action learning, and can help facilitate project completion (Nakanjako et al., 2015; Steinert et al., 2012). One caveat is that Korschun et al. (2010) report that the success of mentoring can depend significantly on the quality of the mentors. There was no mention in any of the studies of training for mentors, which is worthy of further investigation.

6.2.3 Limited Evidence

Although the points that follow are not an exhaustive list of all points listed in the conclusions of the limited evidence studies, two key items emerged consistently:

- **Coaching, 360-degree feedback tools, and assessment tests were highly rated** (Drew, 2009; Edmonstone, 2011; Miller et al., 2007; Pradarelli et al., 2016; Vimr & Dickens, 2013) and every HEE study that included coaching (n = 6) also included 360s. This was also true of two-thirds of the MULTI studies (n = 15), though from the more recent studies (2010 – 2015), only two MULTI studies that included coaching (n = 13) did not include 360s. Coaching, in terms of goal-focused counselling and support, has proven to be an excellent facilitator of leadership development application (Zenger & Folkman, 2003; Hernez-Broom & Hughes, 2004; Leskiw & Singh, 2007). For example, participants who received coaching in the Bowles et al. (2007) study demonstrated significantly higher levels of quota achievement than their non-coached counterparts who had simply learned from job experience over time. McCauley (2008) cites several examples of credible studies reporting coaching interventions led to improved performance compared to control groups. As with mentoring, the effectiveness of coaching is allegedly dependent on the quality of the coaches and facilitation (Drew, 2009; McCauley, 2008; Pradarelli et al., 2016). Further, coaching, when coupled with formal feedback, in particular through 360-degree performance instruments, is also said to be very effective in contributing to ongoing development (Zenger & Folkman, 2003). Straus et al. (2013) also supported multi-source feedback (MSF), of which 360s are an example, as a useful developmental tool. This builds on other research correlating 360s with improved performance appraisal ratings and objective performance measures, though the magnitude is said to be small (McCauley, 2008). Despite many claims that coaching, 360s, and assessment tests facilitate effective leadership development, 360s and assessment tests were only included in one moderate and no good evidence HEE studies and coaching was not mentioned in any of these groups. This absence of empirical data suggests the need for further testing of the utility and optimal use of these developmental activities.

- **Leadership development increases participants' self-awareness** (Bergman et al., 2009; Blumenthal et al., 2014; Miller et al., 2007; Pradarelli et al., 2016; Sanfey et al., 2011; Satiani et al., 2014; Steinert et al., 2012). Even though increased self-awareness has been linked to better leadership performance (Steinert et al., 2012), only two studies in the HEE review used this as a programme-wide outcome metric, though it was reported as a benefit in 19 unique studies in the combined findings. Frich et al. (2014) also noted a lack of self-awareness outcome metrics. Self-awareness was considered the dominant theme of the benefits of feedback collected in the Bergman et al. (2009) study and Rowland (2016) argues that being aware of and regulating one's emotion and mental states is an essential leadership skill. This assertion echoes the work of educational theorist Parker Palmer (1998), who writes, "The quest for leadership is first an inner quest to discover who you are" (p. 160). As mentioned earlier, Miller et al. (2007) suggest that activities that increase self-knowledge, such as assessment tools and coaching, might be most effective when combined with an action learning project so that participants can address weaknesses during interactions with other learning project team members.

6.2.4 No Evidence: What Was Not Highlighted, Innovation

There are many key questions related to leadership development and measurement for which no evidence emerged in this study. These include those related to the sample: size, gender, level of seniority, and domain- or profession-specific versus interdisciplinary makeup of sample participants, faculty characteristics, including internal versus external, location (internal, external, hybrid), length of the programme, and optimal combinations of developmental activities. These have been listed as opportunities for further research in section 7.3; however, one is mentioned here as a further point of interest. Innovative approaches to leadership development did not appear in the findings of the good and moderate evidence studies. This is not to imply that innovation cannot add value to the design and delivery of programmes, but rather that novel approaches have not yet been documented in the academic literature. This omission could also suggest that the core principles of leadership development are to some extent universal and on which innovative and existing approaches could draw. More research needs to be done on this topic before one can say with confidence to what extent it is the one or the other.

The following chapter includes further discussion of measurement, followed by a treatment of the implications for research.

6.3 Conclusions Explored

Two of the conclusions from the best available evidence emerged as demanding further investigation for reasons explained earlier.

The first conclusion explored examines deductively how the theoretical model that was mentioned most commonly in the included studies, Knowles's (1984) principles of adult learning, applies to leadership development. This theory was also addressed in the Steinert et al. (2012) review. In-depth analysis of the included studies from the perspective of this theory evoked the question of which developmental activities best meet the needs of adult learners. To most effectively answer this question, a second theoretical model, Dale's (1969) Cone of Experience, was also interwoven into the set of principles. The second conclusion explored is inductive and outlines factors in programme design or workplace culture that either enable or inhibit the application of leadership development following interventions. This is a key set of features that McCauley (2008) identifies as being under-researched and was selected for this thesis mainly because of the number of studies that claimed that they failed to meet expectations, often for similar reasons. Even if a provider were to follow an empirically-based set of principles of optimal leadership development, it is possible that participants' application of leadership to the workplace may yet be stifled if the factors from this set, including those relating to organisational culture, are not addressed.

It should be noted that although the points that follow are clearly referenced, they are not tiered according to the credibility of the studies as the conclusions above were. In cases of studies that reportedly failed, the explanation for the failures may not be exactly as the authors propose; however, it is unlikely that studies would fabricate poor outcomes. There is a theological scholastic term, the "criterion of embarrassment", which postulates that aspects of early Christian writings are more likely be authentic if they run contrary to the Christian tendency to glorify Jesus (Hägerland, 2015). The same thinking can be extended to the research value of leadership studies reporting negative results, particularly given concerns over perceived publication bias in favour of positive outcomes. While the points in the conclusions explored have been compiled through a less scientific process than those in conclusions from the best available evidence section, they represent strategies for optimising leadership development based on commonly held facilitators of and barriers to successful implementation. Programme providers who follow the scientific findings described in the conclusions and ignore the points in this section expose their interventions to a real risk of sub-optimisation or failure.

Thus, while further empirical testing would add value, many of the following points are already said to be as important as the interventions themselves in influencing post-programme impact.

6.4 Conclusion Explored One: Principles of Adult Learning

Knowles's (1984) widely-used set of principles is based on his belief that until the second half of the twentieth century, there was only one dominant model of assumptions about learners and the process of learning in the West. He believed that this model centred on the practice of educating children didactically (Knowles, 1981). Knowles perceived that over the course of the last century, this prescriptive, content-based approach did not seem to work well with adults. In the late 1960's, he explains, European educators coined the phrase "andragogy," meaning "man not boy" or "adult," in contrast to pedagogy, the art and science of teaching children (Knowles, 1981). Knowles then undertook to define a set of principles describing the climate in which adults learn best and the features of effective practice. The result of his work is four principles, along with a precursor of effective adult learning. The following section describes how each of Knowles's principles, plus two additional principles that were added based on the analysis of the combined included studies, apply to adult learning in the context of leadership development. These descriptions are accompanied by examples of HEE and MULTI studies that incorporated those principles or whose underperformance was attributed to allegedly not doing so. The full set of principles described has the potential to be relevant for leadership programme designers, facilitators, stakeholders, and participants alike and appear to be equally applicable to individual interventions as to long-term leadership development over the course of a career.

Pre-Condition: Motivation to Learn

As a pre-condition for effective adult learning, Knowles (1984) believed that adults have to be convinced of the need to learn and develop and understand the relevance of the intervention to their current work or responsibilities. Berman et al. (2010) add that the best results ensue when participants are galvanised and *interested* in self-improvement and improved decision making, the latter of which is integral to leadership. The research supports this assertion as well, since motivation to learn has also been correlated with improved outcomes (DeRue, Nahrgang, Hollenbeck, & Workman, 2012; Hassan, Fuwad, & Rauf, 2010; Suutari & Viitala, 2008).

Of the HEE included studies, 20 of the 25 featured participants who either volunteered, were nominated, or who applied and were selected; whereas only two studies involved participants who were required to attend. (Selection criteria was not recorded for MULTI). In the programme MacPhail et al. (2015) studied, attendance and programme completion rates were significantly higher the year after providers transitioned from required attendance to self-nominations supported by a line manager. Likewise, Kwamie et al. (2014) contend that participants may experience an increased sense of gratitude and obligation when their supervisor personally recommends them and may be lessened if participation is imposed by the organisation as a mandatory “extra”. Furthermore, managers in the Gilpin-Jackson and Bushe (2006) study reported feeling motivated and even obliged to apply what they had learned since the organisation had shown a commitment by investing in them, especially at a time of strained budgetary resources. Blumenthal et al. (2014) and Vimr and Dickens (2013) echo the importance of making programmes voluntary and conversely, Edmonstone (2009) claims that the effectiveness of the programme he studied was eroded because participants lacked an understanding of the overall purpose, detracting from their motivation. He also cites an inadequate selection process as contributing to the under-performance. In another study, DeRue and Wellman (2009) assert that individuals with a higher learning orientation – those inclined to want to learn and develop – and greater access to feedback experienced linear development; whereas, those with a lower learning orientation experienced diminishing returns in developmental challenge experiences. Hassan, Fuwad, and Rauf (2010) found similar correlations between stronger training motivation and better outcomes, as did Suutari and Viitala (2008) with greater learning goal orientations. Rose (2015) makes the recommendation that individuals should be encouraged to increase their personal accountability for their own developmental needs. Thus, it seems that adult learning is affected by one's initial motivation to learn, one's understanding of, and agreement with, the purpose of the intervention, and one's general openness to learning and developing.

This motivation to learn can be prompted by a variety of pre-programme factors. These factors include recognising the need to develop based on a lower than expected 360-degree feedback report, a performance review, or a desire to extend one's skills further following positive assessments. Alternatively, having an action learning project idea in mind may motivate employees to attend a programme with that focus. This motivation can also be triggered by a potential promotion or position that participants have identified themselves or by the organisation having earmarked for them based on an organisational needs assessment or as part of their leadership pipeline. McGurk (2010) argues that another way to increase motivation is by enabling participants to take ownership of their own development by allowing them to be involved in generating or selecting their own goals and outcome metrics. The advantages of this approach have been mentioned previously. Another incentive was demonstrated by trainees in the Gilpin-Jackson and Bushe (2007) study who asserted that they were even more inspired when their supervisor and peers had also attended the programme and therefore knew the language and modelled the behaviours. The authors say that this situation resulted in a high correlation to leadership learning transfer. This result might also suggest a benefit of having multiple people from an organisation undertaking leadership development at the same time, in line with Day's (2000) distinction between individual leader and broader leadership training. Gilpin-Jackson and Bushe (2007) propose that perhaps the most effective progression is for senior leaders to take the programme first so they can model the leadership norms and begin to spread organisational support of the enterprise, demonstrating that such training is valued before it is introduced sequentially through lower levels of management. It is clear that there is a multitude of ways that participants' motivation can be developed both through their own individual ambitions or by virtue of positive organisational support.

With the motivation to learn in place, Knowles identified four principles of effective adult education, which are as follows:

1. Self-directed

Knowles suggests that adult learning is augmented when participants are aware of their own learning needs and take responsibility for their own development by being involved in the planning and evaluation of the learning process, a point that is echoed by Rowland (2016) and Berman et al. (2010). As suggested earlier, awareness of one's individual developmental needs can arise from 360-degree reports, which were included in 28 of the combined included studies, performance reviews, or as one prepares to take on a new position or responsibility. Adult learners are often aware of their preferred style of learning and it is to be expected that people will engage and retain more when learning by the methods that best suit them. McCauley (2008) stresses in her review that best-practice organisations tend to offer development that is highly customised according to employees' individual strengths, developmental needs, and career potential. A high level of personalised programme content not only accommodates various learning preferences, but also helps participants see the relevance to their own situation, increasing the likelihood that they will apply their learning (Blumenthal et al., 2014; Edmonstone, 2009; Leskiw & Singh, 2007; Van Aerde, 2013). The Blumenthal et al. (2014) and Dickey et al. (2014) interventions with medical residents demonstrate that participants can be involved in selecting the programme content, structure, and goals, as well as their own action learning projects (MacPhail et al., 2015; Nakanjako et al., 2015). Similarly, participants in the Rowland (2016) study organised site visits outside their own organisations to key stakeholder groups and Dickey et al. (2014) stated that involving medical residents in the design of the intervention enhanced their ownership and feeling of responsibility for the quality of it. Personalising aspects of interventions can therefore enhance participants' learning, the perceived relevance to their organisational context, and their sense of ownership of the programme. This principle can be extended to personalising development for teams, as well as to through-career development in addition to one-time, individual programmes.

In addition to personalising learning approaches and content, allowing participants or teams to devise their own individual goals at all three levels can be advantageous. This measure offers another opportunity for them to match their own developmental needs with those of the organisation and to build on feedback from 360s or performance reviews, to take on an interesting project, or to work towards a desired role. Samani and Thomas (2016) assert that when participants choose their own action learning projects, they create more value for the organisation and are less likely to quit. Similarly, Ardts, van der Velde, and Maurer (2010) report that participants' perceived control was the most influential factor in determining positive organisational outcomes. Whether or not participants select the content, developmental activities, and outcome metrics themselves, outcomes are improved when those elements are tailored specifically to their role and level of experience (Blumenthal et al., 2014). This an oversight that Pradarelli et al. (2016) and Edmonstone (2009) claim led to participant dissatisfaction in the programmes they studied. Another reason for involving participants in selecting goals, outcomes, and measurements is the growing number of claims that individuals' development is significantly affected by their learning orientation and capacity to handle challenges and develop (DeRue et al., 2012; Hassan, Fuwad, & Rauf, 2010; Suutari & Viitala, 2008). It is therefore likely that adult learners will select realistic, achievable goals and measurements based on their preferences and sense of their own abilities. Ardts, van der Velde, and Maurer (2010) also determined that participants' perceived control emerged as the most influential factor in improved organisational outcomes, which is intimately linked to self-direction. Sixteen of the combined included studies formalised self-direction through a Personalised Development Plan (PDP), which can be completed before or during the intervention and reviewed or updated at the end, with the intention of it extending after the programme finishes.

Self-direction can also come into play while the programme progresses. The second conclusion explored in this thesis describes the process of maximising the effect of experiences by: setting goals, experiencing an activity, having a discussion/receiving feedback, reflection, re-evaluating goals, receiving support, and repeating the process. This process encourages participants to consider how their learning applies to their role and specific work context (Blumenthal et al., 2014; MacPhail et al., 2015; Patel et al., 2015), contemplate their actual versus intended progress, and to experience increased self-awareness. Participants can be given the opportunity to adapt their personal goals based on their involvement in simulations, 360s, peer and facilitator feedback, action learning, and structured reflection time. Rowland (2016) suggests that it is important to include "stillness and space for intentional, nonobstructed contemplation" (p. 3). This endeavour can also contribute to self-awareness, which is a key component of leadership development (Miller et al., 2007; Rowland, 2016). Of the nineteen unique studies that included self-awareness as a reported benefit, all included at least one of the aforementioned activities. It is apparent that self-direction coupled with developmental experiences, feedback, and reflection is key to maximising the effects of activities and interventions.

The more that participants can take personal ownership of various aspects of their own development, especially in an iterative way, the more beneficial interventions are likely to be.

2. Participants’ experience as the basis

Valuing adult learners’ experience is an important undergirding adult learning principle, as is treating their experience as an educational resource (Berman et al., 2010).

There are several ways that this principle can apply to leadership development. The first is ensuring that programmes are appropriate for the participants’ role and level of experience (Edmonstone, 2009; Pradarelli et al., 2016) and having expert facilitators and speakers who are considered credible by the participants (Murdock & Brammer, 2011). The approach is enhanced if the facilitators exhibit a reciprocal respect for the experience of the participants. In this approach to adult education, participants can join new learning with what they already know, consider how it relates to their current work situation, and apprehend how it can be a tool to enable them to achieve their programme outcomes. This principle suggests that professionals in leadership development programmes should not be treated as “blank slates”, nor offered something as basic as “Leadership 101”. Educational theorist Paulo Freire (2007) coined the term the “banking approach” to describe the idea that the teacher is the sole possessor of knowledge and students are passive receivers, an approach he considered suppressive. An alternative to this approach is to offer adult learners the opportunity to engage with appropriate theoretical models or case studies, which they can discuss and test against prior knowledge (Vimr & Dickens, 2013). Finally, expert facilitators are best suited to enable participants’ development beyond their current capacity, in line with educational theorist Vygotsky’s (1978) theory of the Zone of Proximal Development. This term refers to the area between a person’s current developmental stage and the growth she/he can potentially achieve, which Vygotsky believed could be extended through the guidance of an expert. When considering adult leadership development, this theory implies that faculty should see their role as to facilitate rather than prescribe learning, recognising that there will be differences in participants’ potential results. Therefore, programmes and faculty who value the experience of adult learners and offer theory and information with which learners can engage as co-investigators are said to be more effective than traditional, paternalistic approaches.

In terms of valuing participants’ experience, this can beg the question of whether it is optimal, as some suggest, to restrict programmes to participants of equal seniority due to their similar career experience and role-specific challenges. Unfortunately, no studies were identified that compared mixed-level to level-specific programmes. In the combined sample, many were unspecified, only one study identified a mixed sample, and one combined middle and senior leaders, a sample so small that it precludes comparison to level-specific programmes. Chochard and Davoine (2011) report that in their multi-programme study, the

ROI for entry-level managers was significantly higher than that for training middle managers. The question of equal versus mixed level seniority is worthy of further investigation, since the latter can present challenges. For example, in the Pradarelli et al. (2016) study, many participants commented that mid-level participants got a lot out of it, while for senior faculty, the intervention was perceived to have no impact. Some senior level participants in the same study felt that including residents would dilute the programme because of their early career stage. Two strategies for accommodating groups of mixed experience are 1) including the various measures of personalisation described in the previous section; and 2) a hybrid option of combined mixed-level sessions followed by role-specific syndicate breakout groups.

Action learning projects, simulations with peer feedback and coaching, and case study analysis are development activities that can build effectively on participants' experience (Getha-Taylor & Morse, 2013; Patel et al., 2015; Steinert et al., 2012; Ten Have et al., 2013). These measures are particularly useful when followed by support tools such as coaching and mentoring (Edmonstone, 2011; Hernez-Broome & Hughes, 2004; Kuo et al., 2010; Leskiw & Singh, 2007; Zenger & Folkman, 2003). Another important aspect of this principle is ensuring that lectures or workshops are followed by some form of small group discussion where participants can share opinions, discuss, and benefit from each other's perspectives and experience (Blumenthal et al., 2014). Vimr and Dickens (2013), for example, used cohort sessions and others, such as the Joint Services Command and Staff College (JSCSC) for military leadership and The Staff College for medical leadership in the UK, who have employed syndicate breakout groups. In these follow-up sessions, reminiscent of Freire's (2007) pedagogy, participants become valued active learners and co-investigators and contributors, rather than just passive listeners in a lecture hall. While the teachers and learners are not necessarily on the same professional footing, this approach allows for reciprocal respect and enables both groups to be active participants in the discovery of new knowledge (Freire, 1992).

Deeper-level development and self-awareness occurs when, in a confidential and supportive environment, the biases underpinning participants' experience can be identified and challenged (Getha-Taylor & Morse, 2013). This reflection can be facilitated through small group discussions, as well as through psychometric tests, 360s, coaching, peer and facilitator feedback, reviewing videotaped simulations, and mentoring. It is important that some of these tools are in place following activities and programmes to provide ongoing support for participants' continued development.

Self-direction, combined with participants' involvement in choosing their own goals and outcome metrics, can also enable participants to understand how individual interventions relate to their career projection or PDP. This can help participants recognise areas to develop or of weakness and pinpoint the knowledge, skills, or experience needed for the next step in their progression.

It is apparent therefore that incorporating participants' experience into seniority-appropriate developmental activities enables them to build on their existing knowledge, challenge their underlying assumptions, and consider how the intervention learning can apply to their current responsibilities and outcome metrics.

3. Content that is practical and relevant to participants' careers

Echoing earlier points, adult learners need to feel that developmental interventions are concrete and relevant to their careers and leadership context, current or anticipated. Relevance can be addressed in programme design by involving stakeholders in the design and personalising the content, goals, components, and outcome metrics, as mentioned above. For example, Blumenthal et al. (2014) suggest that one of the key strengths of their intervention for residents was tailoring the content to the participants' context and stage of career. Programme content also needs to be applicable to the participants' organisational context (Van Aerde, 2013), including their current business situation. McGurk (2010) suggests that the lack of anchorage in the organisations' specific business contexts might be why there was little evidence that their leadership programme for middle managers in UK public social services had any substantial or long term organisational impact. Selecting the right faculty can also augment the perceived relevance of the content. Blumenthal et al. (2014), for example, assert that using internal faculty enhanced the perceived relevance of the programme, in part because they are directly familiar with how the organisation works, including its structures, processes, available resources, and challenges. Using internal faculty also offer networking benefits that can carry forward after the programme.

In addition to relevant programme content, developmental activities are also maximised when they are practical and relevant. There is good evidence that videotaped simulations with peer and expert feedback can be effective to train practical skills, as in the Ten Have et al. (2013) study, or develop broader skills, such as communication, This type of practical exercise is effective for both individuals and teams training together. There are also ways of enhancing the relevance of interventions. For example, Bearman et al. (2012) assert that the realistic nature of their simulations that involved simulated patients and accurately reproduced clinical environments contributed significantly to participants suspending disbelief and finding it an authentic experience. Similarly, Rowland's (2016) best practice example is of using simulations to replicate the precise contexts in which participants lead and Yeo (2007) claims that defining a realistic situation is a key feature of problem-based learning. Other effective developmental activities are case study analysis (Blumenthal et al., 2014; Getha-Taylor & Morse, 2013; Steinert et al., 2012), guest speakers (MacPhail et al., 2015), and site visits, either to locations in the same field or those in different fields with centres facing similar challenges (Edmonstone, 2011; MacPhail et al., 2015; Rowland, 2016). These experiences are solidified when participants are prompted to reflect on how the learning and ensuing discussions relate to their current role (Vimr & Dickens, 2013). In terms of individual development, psychometric tests, 360s, coaching, mentoring, structured reflection, and journals help participants to understand how the programme content and activities apply to their individual goals and professional context (Blumenthal et al., 2014; MacPhail et al., 2015; Patel et al., 2015). Several studies included Personal Development Plans (PDPs) as a formal culmination of these developmental activities, which can consolidate learning and goals to map out a strategy of continued development and application following individual interventions.

Using "wicked problems" for discussions or action learning projects can also be an effective leadership tool. The term wicked problems refers to complex problems involving a good deal of uncertainty, imperfect knowledge, and no clear solution (Rittel & Webber, 1973). Examining wicked problems facing an organisation can enable participants to channel the learning from the development programme's various intervention activities and personnel in order to engineer a strategy for tackling them. This process lends practical direction and relevance to the intervention. Surprisingly, only one of the combined included studies featured this tool. In the Yeo (2007) study, the CEO selected the wicked problem and all participants reported finding it meaningful and relevant. One final method to guarantee that leadership development is practical and relevant is to have participants choose their own action learning projects and implement them at their current workplace (Miller et al., 2007; Patel et al., 2015). As has been explained earlier, there is moderate evidence that the action learning projects can contribute meaningfully to organisational and clinical outcomes (Edmonstone, 2011; Patel et al., 2015). This tool can be further optimised by having participants select their own goals and desired outcomes as well.

Leadership interventions can therefore be enhanced when the content is made practical and relevant through the choice of topics, faculty, developmental activities, goals, and outcomes, as well as by building in discussion, feedback, and reflection sessions to connect the learning to the workplace context.

4. Outcomes-based learning

Knowles's fourth principle is "problem-based learning", which is a common approach used in Yeo (2007) and other studies. This principle resembles Freire's (2007) educational theory, whose premise involves presenting issues from the learners' environment in the form of problems and challenging them to analyse the issues critically as co-investigators. Freire believed that this process enabled learners to take ownership of and responsibility for that reality and for its improvement. Outcomes-based learning is similar to problem-based but specifically-directed rather than open-ended. It implies that there are expectations of tangible applications of learning from the beginning that form the focus of the intervention. Problem-based interventions can either have participants consider *hypothetical* problems and generate *theoretical* solutions, or can address actual problems in their organisations without necessarily implementing them. In both cases, there is potential for decision-making and teamwork skills development as well as self-awareness and networking benefits. Certain elements of outcomes-based learning would be missed however, including explicit evidence of the impact of programmes (by outcomes attained) and the learning associated with the process of actually attaining them, which could exceed that of simply devising a strategy that is not necessarily implemented.

An outcomes-based approach is more suitable for leadership development interventions than purely content-oriented programmes for the same reason: *application* is of the utmost importance (Edmonstone, 2009; Raelin, 2011). These programmes are not intended purely for personal development, since, as defined in chapter two, leadership is not a solitary enterprise. Leadership interventions are most effective when they prepare participants to achieve a variety of outcomes in their leadership context after and, in some cases, as the programme progresses. Adult learning in a leadership development capacity would therefore transfer the needs or priorities that the participants and their organisations have identified into programme objectives and outcome measures (Mountford & Webb, 2009). Further theoretical backing for this approach is found in Edgar Dale's (1969) educational theory, which will be discussed in more detail in the sixth part of this section. Dale advocates taking a systems approach when designing interventions and first determining the "desired terminal outcomes of instruction, the exit behaviour" (p. 7), which is supported by others (Fayolle & Gailly, 2008; Nabi et al., 2017). From there, he asserts, each developmental activity is seen as an interrelated part of an orchestrated learning programme directed toward these outcomes (Dale, 1969). As will be presented in detail in the prototype theoretical model at the end of this chapter, the design of an outcomes-based programme begins with identifying targets at the individual, team, organisational, and clinical/benefit to clients levels, from which point the goals of the programme, content, developmental activities, outcomes, and measurements can be devised so that all five link together symbiotically. MacPhail et al. (2015) provide a good visualisation of how the various components connect to one another (see Figure 6.3). This approach elucidates the relevance and application of the development intervention at the outset, which can be reinforced in each session and by way of support methods and reflection and gives the programme a clear direction and purpose. As mentioned previously, it is helpful if participants, with input from their supervisors and stakeholders, also contribute to identifying these outcomes.

Johnson et al.'s (2012) research regarding goal setting reinforces idea, suggesting that it increases motivation, energy, commitment, and persistence toward achieving goals. They also report that having established outcomes has been found to lead to increased transfer of learning, since striving to meet goals ignites self-regulatory behaviours such as self-monitoring and evaluation, reflective self-appraisal, and constructive reactions to performance standards. Likewise, Richman-Hirsh (2001) asserts that participants who set goals following interventions apply their learning more than those that do not. Furthermore, Latham and Locke (1983) provide strong evidence that participants with specific and challenging goals consistently outperform those who are given vague, less challenging goals. This motivation is enhanced when participants or teams are held accountable to their goals, an expectation which is most effective when it extends to the medium and long term, as described in the measurements section to follow. Yeo's (2007) practical steps of successful problem-based learning are similar to outcomes-based learning: a) securing an appropriate problem by defining a realistic situation, b) fostering team ownership through open communication, and c) utilising relevant resources and expertise. The latter point implies application, which an outcomes-based approach makes explicit. To concretise the link between the programme and outcomes, the leadership programme described in the Gilpin-Jackson and Bushe (2007) study devoted a session at the end to planning strategies of implementation of leadership skills at the workplace and related goal setting, which they claim has proven to lead to successful training transfer. This will be discussed further in the following section.

An outcomes-based approach to leadership development can also be connected to participants' career progression as an evolving process aligned with the long-term organisational strategy; this approach can help situate individual interventions in a larger context and trajectory, rather treating them as isolated occurrences.

Therefore, an outcomes-based approach provides purpose and direction to leadership development by explicitly outlining the focus that permeates all aspects of the programme. This approach also interweaves the previous three principles of adult learning together. Participants are held accountable to their outcomes, which involves self-direction, particularly if they are involved in selecting those outcomes. It also values their existing experience and challenges them to extend their capabilities further. Finally, the entire approach reinforces the practical and relevant nature of the intervention.

After the in-depth analysis of the HEE and MULTI included studies with the principles of adult learning in mind, and following discussion between this thesis's author and the HEE collaborating researcher, it became evident that two principles of adult learning should be added to this list: measurement and experiential and application-centred.

5. Additional principle: Measurement

Given the importance of application in leadership development, measuring outcomes is instrumental in maximising the effects of programmes and benefiting individuals, teams, and organisations concomitantly (Leskiw & Singh, 2007). McCauley's (2008) review lists measuring outcomes as a key best practice theme and recommendation. As mentioned previously, interventions with specific and challenging goals deliver superior results compared to those who do not (Latham & Locke, 1983). Understandably, stakeholders, including those funding, designing, and delivering programmes, as well as participants' supervisors, anticipate evidence of the impact of programmes (Beer et al., 2016). And yet, as mentioned in the introduction, it is surprising how few providers attempt to collect data on this at all, let alone do so comprehensively. Equally important, failing to evaluate interventions leaves them at risk of stagnation and falling short of their potential (Boaden, 2006).

The previous section stressed the importance of establishing clear outcome metrics at the individual, team, organisational, and clinical levels that are announced before the programme starts. Ideally, these metrics should address all four levels of the Kirkpatrick model and be reinforced with objective data, a combination which only two studies in in the HEE review and not one unique MULTI study demonstrated. Again, this type of measurement is most effective when it compares a baseline measure, measures during the intervention (depending on the length of the programme), ones after the programme, and post-post tests to track the long-term impact of learning application. Only six studies (8 per cent) in the combined HEE and MULTI study included this combination of measurements. Post-post measures should be collected a minimum of six to nine months following interventions in order to allow for the application of learning, which the Abrell et al. (2011) study demonstrated can take time, as well as to ensure that immediate post-programme improvements are sustained, the failure of which was seen in Kwamie et al. (2014). Although some less tangible cognitive, soft, and personal leadership skills may be challenging to quantify, an effort should be made to do so in order to maximise their development. To ensure a 100 per cent response rate, interventions that collect data during the programme should build in time for that function as part of their itinerary, not as an extra. Ongoing measurement also allows providers and facilitators to make adjustments to the programme as it progresses based on participant feedback and enables participants to reflect on and possibly modify their goals.

Having measurable goals benefits the participants themselves as well. Measuring is an integral part of goal setting and includes the advantages mentioned above in terms of focusing participants' learning and encouraging them to reflect on how programme content and components relate to their end goals. This has already been mentioned as serving as an additional development tool (K. E. Watkins et al., 2011). Having specific and challenging goals has also been found to improve their participants' performance (Latham & Locke, 1983). Furthermore, on an individual, team, and organisational level, measuring also provides clear evidence of the progress being or not being made during and following interventions. This evidence can be reassuring when progress is good and it can also indicate if further support is needed to attain the desired results. Mountford and Webb (2009) add that making performance data transparent can motivate clinicians to be involved in improvement efforts. The rationale is that publishing performance data, when done constructively, can serve as a social contract whereby people are incentivised to perform well and improve out of a sense of self-respect. This approach is more constructive for goals at the team, organisational, and benefits to clients levels than the individual, since publishing individual data such as 360 report results could be damaging to participants' self-confidence. Good leadership from participants' managers includes keeping apprised of their progress and offering support when necessary (McCauley, 2008). It also bears repeating that having participants, supported by their supervisors, select their own goals, is an underutilised form of measurement that relates to many of the principles of adult learning described above (McGurk, 2010).

In addition to maximising leadership development as an enterprise, measuring effects can provide the justification of the return on investment (ROI) that various stakeholders desire, particularly when economic benefits are included. The Jeon et al. (2013) study, which will be described in the following chapter, is a good example of this. Without discounting the value of intangible benefits such as increased self-efficacy in participants, significant monetary savings through lower turnover or absenteeism or achievement of objective Level 4a or 4b outcomes, such as a significant drop in preventable in-patient deaths, can seem immediately convincing. In addition to allaying concerns over ROI, such evidence can contribute to making the organisational culture more supportive of leadership development. The information can also be used to refine programmes and maximise their impact, particularly based on the combination of participants' feedback and the evolving needs and situation of the organisation.

Finally, measuring leadership programmes can also contribute to the overall organisational strategy. In addition to aligning the goals of leadership development with human resources and organisational strategy, action learning projects and the kinds of wicked problems discussed earlier can also contribute on a systemic level. For example, Rose (2015) suggests that when an organisation is without a systematic system of appraisal, developing key talent is almost impossible. Similarly, Steinert et al. (2012) assert that given the role of leadership in creating social change, assessment over time is critical. Finally, Mountford and Webb (2009) recommend that all healthcare organisations track measures of clinical leadership development and correlate them with their impact on quality and costs, which has been described as the classic tension in the medical domain.

Therefore, measuring the outcomes of leadership development can help maximise its effectiveness, enhance participants' experience, provide legitimacy through evidence of ROI, privilege information to refine programmes, and contribute to organisational strategy.

6. Additional principle: Experiential and application-centred

It has been argued that the traditional lecture-centric approach to leadership development seems to be giving way to more experiential forms where participants can apply their learning as part of the intervention, rather than afterwards. A lack of experiential and application focus of leadership development is analogous to attempting to teach people to fly a helicopter without ever using a helicopter. (The situation would be even more absurd without even using a helicopter simulator). Getha-Taylor (2013) suggests that traditional approaches to development are less effective given the needs of adult learners. This is not to imply that lectures and didactic workshops have no place in such programmes, but rather that their role should be carefully considered as opposed to being implemented as the default core of every programme. Rowland (2016) agrees, saying that if leadership development begins in the head, learning will stay in participants' heads, implying that it may not translate into action. Similarly, Dickey et al. (2014) contend that didactic components can never replace the actual experience of leading. Thus, application-oriented programming in the context of participants' actual work or leadership contexts is optimal (Gilpin-Jackson & Bushe, 2007; Van Aerde, 2013).

Making interventions experiential is the first factor that Rowland (2016) contends lies at the heart of effective, practical leadership development. It can be argued that *all* leadership development should be practical. This approach need not exclude the role of theory, but purely theoretical programmes should be recognised as having limited value given the advantages of outcomes-based development described in the previous section. Rowland (2016) argues that neuroscience demonstrates that people learn and change behaviour most when emotional circuits in the brain are activated, which happens most effectively through visceral, lived experiences. She notes that adults often learn better actively, rather than passively, and that novel experiential activities engage learners' intentional mind to make conscious decisions about their behaviour. The behaviour change that Rowland describes as a result of experiential activities can occur in leadership development through simulations or role plays, as well as through action learning projects, in which case personal development happens as progress towards outcomes is made. The effectiveness of experiential activities is reinforced by Berman et al. (2010) and ties in with long-standing wisdom from the field of education (Knowles, 1981). One example that was mentioned earlier is Dale's 1969 model, "the Cone of Experience", which has been adapted countless times. Although the original is admittedly not based on scientific evidence, and should be considered with that in mind, it depicts a useful classification system of different forms of pedagogy (and by extension, andragogy). Two such adaptations are included below in Figure 6.1 and Figure 6.2. The first is called The Learning Pyramid (NTL Institute, n.d.), which was developed by the National Training Laboratories. The second derives from a course at Queen's University in Kingston, Canada (Anderson, n.d.).

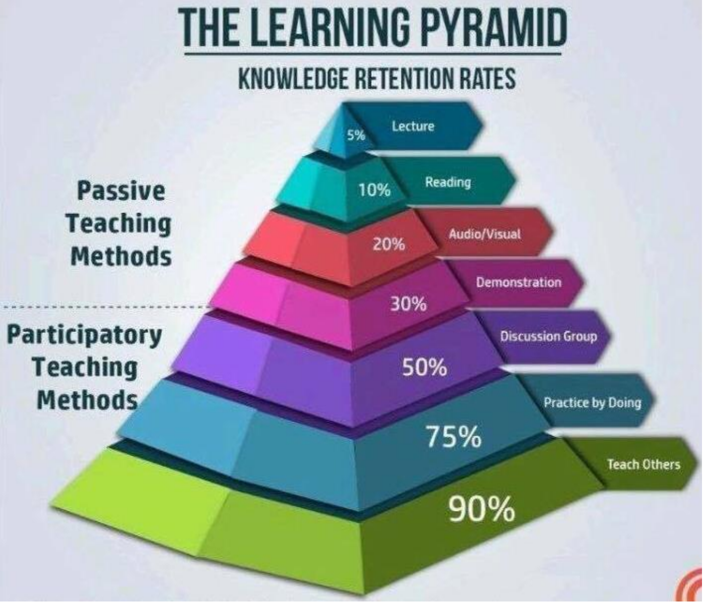


Figure 6.1 The Learning Pyramid by the National Training Laboratories.

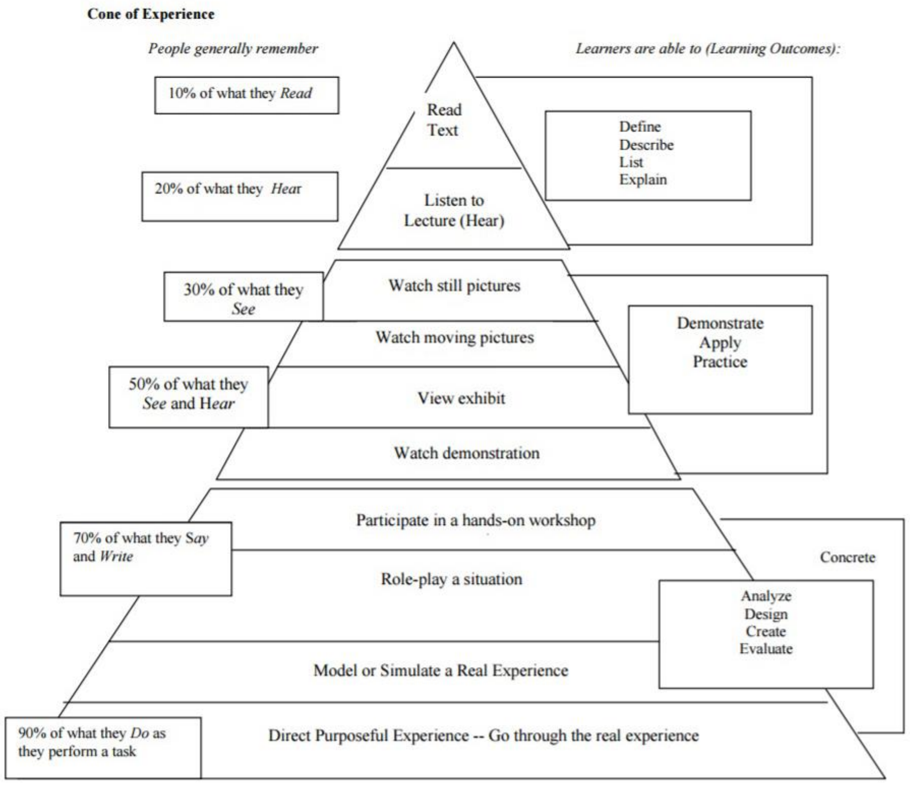

Cone of Experience
People generally remember
Learners are able to (Learning Outcomes):
10% of what they Read
Read Text
Define
Describe
List
Explain
20% of what they Hear
Listen to Lecture (Hear)
30% of what they See
Watch still pictures
Demonstrate
Apply
Practice
Watch moving pictures
50% of what they See and Hear
View exhibit
Watch demonstration
Participate in a hands-on workshop
70% of what they Say and Write
Concrete
Role-play a situation
Analyze
Design
Create
Evaluate
Model or Simulate a Real Experience
90% of what they Do as they perform a task
Direct Purposeful Experience -- Go through the real experience

Figure 6.2 Anderson's Adaptation of the Cone of Experience.

Above are two depictions of recent adaptations of Dale's Cone of Experience, which classifies different forms of developmental activities.

It is unclear what evidence reinforces the percentages in both figures; however, they represent the widespread notion that experiential activities often enable learners to consolidate and retain more information than do passive forms. The second model above provides interesting learning outcomes at all levels of the pyramid within the same overall structure. The notion of "practice by doing" in the first adaptation and "direct purposeful experience" in the second mirrors the finding from the conclusions from the best available evidence section regarding action learning. Similarly, the references to role plays and simulations in the second model matches the good evidence findings described earlier. Rowland (2016) argues that these kinds of experiential activities prompt learners to become more aware of things in the external environment, as well as inside themselves – enhanced self-awareness – concurrently. She also suggests that duplicating the precise contexts in which participants lead is most effective, which mirrors what Kneebone (2005) and Getha-Taylor and Morse (2013) have said regarding simulations.

The above figures also echo the need to re-evaluate the process of leadership development programme design in terms of which developmental activities best suit the purposes of a given intervention (Leskiw & Singh, 2007). In an outcomes-based programme, this approach would involve first selecting the desired outcomes and then deciding which components would best address those targets and meet the needs of the participants. Ideally, the selected developmental activities would be a combination of formal training, experiential and work-based learning, and structured support, such as coaching and mentorship (Edmonstone, 2013; Frich et al., 2014; Hernez-Broome & Hughes, 2004; Leskiw & Singh, 2007; McGurk, 2010; Van Aerde, 2013; Zenger & Folkman, 2003). The in-depth analysis of this stage of the thesis research revealed that each developmental activity has certain strengths, weaknesses, and most importantly, key *functions*, and are most effective when utilised according to these key functions. There are of course issues of cost and feasibility that can influence decisions around the best package for each organisation and situation. MacPhail et al. (2014) suggest that programme costs and limited workforce resources restrict the number of staff who are able to attend, generating increased pressure to optimise the experience for those who are able to participate. Regardless of the activity, adult learning is maximised when activities are experiential and include an application component.

The following points regarding the capital functions of development activities and their relation to each other derive from the conclusions of the best available evidence and the preparation of the first conclusion explored. In addition to being important considerations for practice for those designing new or refining existing programmes, these points could represent the beginnings of a theoretical model:

- **Experiential components** such as simulations can effectively form the core of shorter interventions, as can action learning for medium-length and longer programmes. This premise is supported by the best available evidence and immediately addresses the application focus of leadership development. Simulations with repetition and peer and expert feedback are particularly effective for brief interventions targeted at specific skills, particularly observable tasks and behaviours (Getha-Taylor & Morse, 2013; Ten Have et al., 2013), but are not limited to that function. Action learning naturally fits better with medium and longer programmes and can be instrumental in producing results at the team, organisational, and clinical/benefit to clients levels. Samani and Thomas (2016) and others suggest that methods like action learning are far superior to traditional didactic leadership development programmes. Galli and Muller-Stewens (2012) add that action learning leads to developing social capital. With both simulations and action learning, participants have a chance to apply knowledge acquired through other sources, learn and practice skills, develop relationships and teamwork skills, and augment their self-awareness (Bearman et al., 2012; Rowland, 2016). Another benefit of these two activities is that participants develop their problem-solving and *adaptability*, which are vital leadership skills (Heifetz, 1994), in response to changing conditions, while being supported by other programme components. Finally, action learning requires that participants apply their learning as part of the intervention, which enables them to benefit from the programme resources and personnel while they have access to them. This can include troubleshooting when difficulties are being experienced or extending one's goals if progress is good.

- Experiential activities are most effective when **supported** by expert and peer feedback, mentoring, coaching, and networking during and after programmes. These fulcrum mechanisms ensure that learning is maximised and constructive and increase the likelihood that action learning projects will be successfully implemented.
- Experiential components can be effectively **preceded** by psychometric tests and 360s to identify strengths and areas of improvement which can be incorporated as part of the intervention goals.
- **Lectures and workshops** can provide theoretical and conceptual models, as well as practical information and details about the organisation, its protocols, or its situation, internally or externally. This information can be then *applied* in simulations, role plays, and action learning projects.
- **Case study analysis** can provide an opportunity to consider how theoretical and conceptual principles, as well as related practical examples, apply to participants' leadership and organisational situation. This process can also help develop strategic thinking and problem-solving skills, among others.
- **Throughout** the intervention, instances that allow for **discussion and structured reflection** can be implemented so that participants can contemplate the learning attained through didactic, developmental, and support structure means, consider its application to their workplace, and re-evaluate their goals.
- These combinations of activities are most effective when linked to specific **outcomes** and **measurements**. Regardless of an intervention's length, Level 4a and 4b outcomes should be included. As suggested earlier, these measurements can derive from data routinely collected by organisations on workplace satisfaction and human resource statistics, as well as regular business or clinical statistics, such as those used in hospital performance evaluations.

Once again, these points are intended as *principles* of adult learning applied to leadership development, not as a single prescription. Many programmes in the included studies failed to state the role and goals of developmental activities, alone and in relation to each other, leaving one to wonder whether or not the role and goals were selected intentionally and communicated to participants before or during the intervention. As evidenced in the MacPhail et al. (2015) study and elsewhere, when the role and goals of each component of the programme are stated, their purpose and relevance is enhanced. Thus, the above progression allows for infinite variation among combinations of interventions, developmental activities, outcomes, and metrics.

Table 6.1

Summary of Conclusion Explored One

Conclusion Explored 1: Principles of Adult Learning	
Pre	Motivation to learn
1	Self-directed
2	Participants' experience as the basis
3	Content that is practical and relevant to participants' careers
4	Outcomes-based learning
5	Measurement
6	Experiential and application-centred

The table above depicts Knowles’s (1984) principles of adult learning applied to leadership development, with two new principles added based on the SEA analysis.

6.5 Conclusion Explored Two: Factors that Affect the Successful Application of Leadership

As mentioned previously, in addition to being a finding from a moderate evidence study, one of the impetuses for this conclusion explored was the realisation during earlier stages of analysis that many of the studies that reported failure attributed this underperformance to similar factors. Gilpin-Jackson and Bushe (2007), for example, cite numerous statistics and examples of failed or poor training transfer results, with sometimes as low as five per cent of participants reporting that they had successfully applied skills at work and others who alleged that this transfer was lost over time. When examples of poor outcomes were combined with findings from the best available evidence and included studies’ reports of best practice, a set of factors emerged to facilitate the successful application of leadership following interventions.

Allen and Hartman (2008) caution that if these kinds of factors are not heeded, programmes will be set up to fail. Some of these factors are worded constructively, as counters to factors that were said to inhibit the application of leadership. Many others agree that these seemingly peripheral variables impact, for better or for worse, directly on programme outcomes (Peters, Baum, & Stephens, 2011). The compilation of this set of factors adds breadth to the study of optimal leadership development by expanding the scope beyond just the intervention components to key, surrounding factors. As an example of the relevance of these concepts, Gilpin-Jackson and Bushe (2007) describe a link between organisational culture, which is featured in points three and nine below, and outcomes. The authors report a direct correlation between the confidence generated by a supportive organisational culture, which they suggest is one of the keys to training transfer, and the actual application of leadership capabilities in the workplace, confirmed by colleague raters. Managers who worked in organisations with supportive cultures also claimed that they valued training more than those in environments that were not conducive to learning transfer (Gilpin-Jackson & Bushe, 2007).

Although this section is derived less scientifically than the conclusions from the best available evidence, it seems convincing that following principles of optimal leadership development while ignoring the factors below exposes interventions to a considerable risk of failure. In addition to the combined SLRs studies, Leskiw and Singh's (2007) extensive literature review of leadership development best practices, Van Aerde's (2013) summary of best practices of effective leadership courses, and Gilpin-Jackson and Bushe's (2007) article on leadership training transfer are particularly useful in addressing these interconnected topics. The summary points from these three studies are included in the appendix on page 389. What follows draws from these articles and the other sources mentioned above. While the reasons cited for why programmes have failed often rely on authors' own perceptions and are not reinforced by objective data, the points that follow are noteworthy given the opportunity cost, budgetary consequences, and potential effect on employees, patients, and clients, of failed programmes. Although some of these factors pertain more immediately to in-house than external programmes, all the points can either be applied equally to external interventions or can be considered by organisations when sending delegates on external programmes. This section is divided into three sets of factors that facilitate effective leadership development training transfer: pre-programme, programme, and post-programme.

Each of the points below is worded in a directive manner in order to avoid repeating:

at the beginning of each point. These are offered as considerations for those designing and refining programmes.

6.5.1 Pre-Programme

1. Ensure the organisation's doctrine regarding the understanding of leadership and its capability framework is clear, shared, and pertinent, including for different levels of seniority

Echoing the points made in the section on leadership development ontology and epistemology, it is beneficial for organisations to clarify their underlying leadership philosophy and guiding principles, including a definition of leadership and a related capability framework (Rose, 2015; Zenger & Folkman, 2003). A clear understanding of how this doctrine applies to leadership programmes, which is shared among providers, stakeholders, and participants, is said to improve the effectiveness of interventions and measurements after (Edmonstone, 2013; Zenger & Folkman, 2003). Conversely, Rose (2015) and others also allude to the problems or limitations that ensue when organisations lack a clear and unified understanding of these concepts, including conflicting expectations that can negatively affect outcomes.

When an organisation already has such doctrine in place, programme designers or stakeholders approving employees' participation in leadership development programmes can review whether any aspects of the doctrine should be adapted for specific interventions. This adaptation can be based on a number of considerations, including the role, level of seniority, or range of participants, needs of the organisation, or the organisation's current situation. The former two are in line with the principles of adult learning. For organisations that do not have formal leadership doctrine, the process for creating it can begin with conducting an internal audit, which can contribute to building consensus (Giber, Carter, & Goldsmith, 2000; Van Aerde, 2013). This initiative is enhanced when external information is added to ensure that the organisation's theory and practices benefit from other perspectives and are grounded in research (Beeson, 2004). Leskiw and Singh (2007) add that best-practice organisations also use focus groups and strategic planning sessions to develop their final products and build further consensus. Involving multiple stakeholders in the design of leadership doctrine or development programmes strengthens their trust and backing, widens support for interventions, and improves the quality of the finished product (Blumenthal et al., 2014; Van Aerde, 2013). Therefore, starting with a clear corporate understanding of leadership and its adjoining capabilities, specific to the organisation and its context, is a solid foundation upon which decisions regarding the role and design of leadership development can be made.

2. **Conduct a needs and barriers assessment, select participants accordingly, and align participants' developmental goals with the organisational strategy** (Hartley & Hinksman, 2003; Leskiw & Singh, 2007).

Needs and Barrier Assessment

Once the organisation's understanding of leadership and its capabilities is clear, in terms of leadership development interventions, congruence between the expectations of the participant(s) and their organisation(s) is said to be a key factor in facilitating effective post-programme application. Conversely, Edmonstone (2009, 2011) cites conflicts in this regard as a reason for both of his programmes included in the HEE review not meeting expectations. One way of achieving this alignment is by conducting a needs assessment (D. V. Day & Halpin, 2001; Leskiw & Singh, 2007). In addition to making it more likely that the intervention will accurately address the needs of the organisation, along with the previous point, involving many stakeholders in a needs assessment can serve as a strategy for gaining support for the programme and the application of leadership afterwards. Edmonstone (2013) elaborates that needs assessments can identify perceived deficiencies in existing capabilities in individuals or in the organisation overall, as well as capabilities that staff will need in order to face a current or anticipated situation. They can also pinpoint personnel roles that need to be filled or team or organisational areas for improvement or expansion. Needs assessments were included in McCauley's (2008) themes of best practice. Surprisingly, only ten of the HEE included studies reported conducting a needs assessment prior to the programme, including two of four moderate evidence studies; the other two moderate and both good evidence studies left it unclear whether they did or not.

An important though seldom cited pre-programme initiative is raised by Beer et al. (2016). They suggested that it is helpful if organisations also collect confidential data before programmes begin regarding policies or practices embedded in the organisational culture that could possibly inhibit the transfer of learning, presenting barriers to post-programme implementation. Such initiatives could potentially uncover perceptions of a lack of resources or conflicting organisational priorities or assessment structures. A barriers assessment gives providers and stakeholders time to modify aspects of the culture or put measures in place to overcome them, removing such barriers before they can affect programme outcomes.

While needs and barrier assessments more obviously apply to in-house than external programmes, they can still be applied to individuals attending external programmes, such as choosing an external intervention that can meet the needs of the participants and organisation.

The Staff College example offered earlier regarding team training to address an organisational wicked problem is a good example.

Selection

Part of maximising the effect of leadership development from the point of view of the organisation is by nurturing and developing the kind of leadership talent that is intended to serve its strategic purposes (D. V. Day & Halpin, 2001; Leskiw & Singh, 2007). This involves selecting those employees to develop intentionally, whether for routine strategic training at certain stages of employees' careers, for immediate needs based on turnover or expansion, or for long-term planning. For non-immediate needs, it is beneficial to establish a clear connection between organisational succession needs, high potential employees, and appropriate leadership development initiatives (Ibarra, 2005; Redecker, 2004).

Targeting individuals to develop intentionally is not at odds with Day's (2000) assertion that leaders and developing leadership capacity are needed throughout an organisation, since in most organisations, even though leadership can be demonstrated by anyone, certain formal roles need to be filled. Kesler (2002) describes "high potentials" as those who are thought likely to succeed at higher levels based on an objective evaluation of past accomplishments, along with, rather than based exclusively on, a recommendation from a supervisor. Metrics and criteria to inform promotions are interesting topics, but beyond the scope of this study to discuss in detail. Leskiw and Singh (2007) propose a solution to the debate of "all or a select few", which is to have two sets of leadership programmes: an advanced one for high potentials and another for all or larger numbers of employees. Similarly, McCauley (2008) highlights that best practice organisations give more attention to high-potential employees, while still providing opportunities for employees at all levels of the organisation as part of the organisational strategy. Dalakoura (2010) suggests that an advantage of developing leaders at all levels of the organisation is that they act like owners and take initiative, solve problems, experiment, buy into the corporate vision and language, and accept accountability for meeting goals, more so than if they see themselves merely as employees. In addition to their potential, a final sample consideration is potential participants' motivation to learn, which is significantly correlated with outcomes (DeRue, Nahrgang, Hollenbeck, & Workman, 2012; Hassan, Fuwad, & Rauf, 2010; Suutari & Viitala, 2008). Along with identifying the individuals to undergo formal development, organisations must balance leadership needs with budgetary and scheduling restrictions when deciding how to fund and prioritise leadership development.

Finally, it is helpful if organisations treat leadership development as an integral part of the organisational strategy (D. V. Day & Halpin, 2001; Giganti, 2003; Jeon et al., 2013;

McCauley, 2008; Van Aerde, 2013). Montesino (2002) suggests that aligning the two is correlated with high levels of self-reported training transfer and Gilpin-Jackson and Bushe (2007) state that most scholars agree that this alignment is necessary to demonstrate the value and ROI of training. When leadership development is seen as a key priority in the overall organisational strategy, resources and support are likely to follow in a more robust and constructive way than if it is considered extraneous. McCauley (2008) adds that learning experiences have a greater impact if they are connected intentionally to other experiences as part of an organisation's ongoing, through-career leadership development system. Thus, just as it was explained in the principles of adult learning that it is helpful if the role and goals of individual developmental activities are explained in the context of an intervention as a whole, the same is true of explicating the relevance of various programmes in the context of employees' career development.

Therefore, when organisations engage stakeholders in a needs and a barriers assessment, select participants with specific purposes in mind, and align leadership development goals with the organisation's strategy before planning or choosing a programme, interventions can directly address fundamental organisational priorities and are more likely to be successful in doing so.

3. **Generate organisational support, including from the senior management, and involve stakeholders in the design and, at times, the delivery of programmes** (Edmonstone, 2011; Kuo et al., 2010; McCauley, 2008).

Many authors have advocated gaining upper management support before a programme is designed or launched to increase the likelihood that they will share ownership of contributing to its success (Beer et al., 2016; D. V. Day & Halpin, 2001; Gilpin-Jackson & Bushe, 2007; Simmonds & Tsui, 2010; Van Aerde, 2013). Blumenthal et al. (2014), for example, suggest that involving stakeholders in the design of their programme solicited their and wider institutional support. They add that this measure also increases the likelihood that the organisational culture will be receptive to participants applying leadership after the intervention and that they will be more motivated to do so, which is echoed by Gilpin-Jackson and Bushe (2007). Likewise, MacPail et al. (2015) purport that executive and line manager support was a significant factor in contributing to the feasibility of the programme they studied. They suggest that this also imbued the programme with a heightened sense of credibility and demonstrated that the intervention and its participants were valued by the organisation. Conversely, Gilpin-Jackson and Bushe (2007) state that the lack of supervisor support has been referred to as the bane of training transfer. As an example of both instances, Edmonstone

(2009) asserts that the programme he studied was successful when there was an executive-level champion during and after the programme; conversely, when the programme lacked such a person and the ensuing support, the outcomes suffered. McCauley (2008) suggests that best practice organisations make managers accountable for the development of their direct reports. Similarly, Steinert et al. (2012) conclude in their review that institutional support was reported to be critical to the success of many interventions and the lack thereof was the primary obstacle to the successful implementation of leadership learning after. The importance of post-intervention institutional support will be described later; however, several authors suggest that engineering the necessary backing begins at the planning stage.

Involving stakeholders in the delivery, as well as the design, of a programme also carries with it several advantages. First, many authors note that involving senior management as facilitators demonstrates their commitment as organisational champions of leadership development (Blumenthal et al., 2014; Edmonstone, 2013; Van Aerde, 2013). For example, Simmonds and Tsui (2010) report that the involvement of senior executives was most effective in achieving commitment to organisational values and strategies, imparting leadership skills, and encouraging the implementation of learning. Second, the perceived relevance and practical nature of the content for participants in terms of their organisational context is enhanced when internal leaders present the material, since these facilitators have first-hand experience of it. Third, senior colleague facilitators can address intricacies of the organisational underlying norms and culture, including its protocols, challenges, and resources, as they affect leadership (Leskiw & Singh, 2007) in a way that external facilitators cannot. Finally, using in-house staff gives the participants access to internal leaders that they might not otherwise have had, which can be a valuable networking opportunity that can continue after the programme. This is not to discount the novel perspectives and expertise that external facilitators can provider, nor ignore situations of mixed faculty; but rather to highlight some advantages of incorporating internal leaders.

Therefore, whether through the design, delivery, or both, it is crucial to have stakeholder buy-in to optimise the impact of leadership development programmes.

4. Ensure that there is a common understanding of the programme purpose, goals, content, outcomes, and measurements among the provider, participants, and the organisation

(Edmonstone, 2011; Hartley & Hinksman, 2003; Steinert et al., 2012).

The next step in designing or refining leadership development programmes could be ensuring that the educational objectives of a given programme are aligned with the organisation's leadership doctrine, needs assessment, and strategy (Giganti, 2003; McCauley, 2008; Montesino, 2002). Surprisingly, in the HEE review, eight of the studies (32%) left it unclear whether there were explicit goals for the interventions or not, which can lead to confusion or conflicting expectations as to the programme's intent and objectives. As described in the first conclusion explored, based on the purpose of the programme, the design or refining process in an outcomes-based approach begins with the desired outcomes, which translate into or inform the programme objectives. It is beneficial if these objectives are also tailored to the participants' role and career stage and their organisational situation. There are ways of achieving this even in cases of mixed samples, including those of varying levels of seniority or from different organisations. These methods include devising personalised outcomes and goals, as well as administering reflections about how the programme content relates to one's organisational situation. Watkins, Lyso, and deMarrais (2011) suggest that when participants' supervisors have input into the objectives, particularly when they are accountable to the participants' development, the supervisors are more likely to carefully consider what outcomes one can reasonably expect from a particular intervention. McCauley (2008) concurs, asserting that in this situation, outcomes improve as well. It is advantageous if this step is followed by ensuring that the providers, stakeholders, and participants have a mutual understanding of the purpose, goals, content, outcomes, and measurements so the expectations, accountability, and resources made available are harmonious. Edmonstone (2009) adds that problems arise and outcomes suffer when there are unclear or conflicting expectations. Likewise, programme outcomes seem likely to underachieve when there are *no* expectations, given the advantages of goal setting mentioned earlier and accountability, as will be described later in this section.

Furthermore, informing others in the organisation of the factors mentioned above can generate healthy and appropriate expectations and accountability. This can increase participants' motivation to achieve their targets and confirm that the organisation is prepared to offer the necessary resources and support. For example, Coloma, Gibson, and Packard (2012) purport that involving participants' supervisors in assessing and supporting the application of learning following leadership development programmes is a key success factor. Finally, McCauley (2008) proposes that it is helpful when stakeholders and participants understand how the developmental goals for individual programmes relate to other forms of leadership development, performance measurements, and career progression in the organisation. Alignment among the provider, participants, and organisation in terms of the programme's goals, purpose and measurements is therefore said to be key to maximising the application and impact of leadership learning following interventions.

5. Ensure that there is a connection among the outcome goals, content, activities, and programme evaluation and that measurements are collected pre, during, post, and post-post intervention

(Fernandez et al., 2016; Kuo et al., 2010; MacPhail et al., 2015; McCauley, 2008).

Echoing the fourth and fifth points from the adult learning conclusion explored, it is beneficial to draw clear and explicit connections among curricular goals, content, relevant programme activities, and outcome metrics. MacPhail et al. (2015) provide a chart that illustrates an example of how this can look in practice (see Figure 6.3 below). As outlined in the measurement section of the findings, since long-term application is a key goal of leadership development, tracking outcome metrics on several levels before, at the conclusion of, and at various points after programmes is important (Sanfey et al., 2011; Steinert et al., 2012). Only nine of the combined included studies collected data at these intervals. The importance of evaluation during the programme has already been covered, as has the need to build time into the itinerary to collect this data. As suggested earlier, it can be beneficial to have participants and their supervisors involved in selecting the outcomes to be evaluated. McCauley (2008) adds in her review that programme designs should be based on a "theory of change" that specifically outlines the process by which leadership learning can be applied successfully to produce results across the organisation. Mapping out a theory of change also helps identify necessary resources to make the results possible and encourages providers or those sponsoring participants to ensure that those resources are available. As mentioned previously, when supervisors are held accountable to participants' outcomes, they are more likely to offer the support and resources needed to ensure that they are successfully met. Therefore, outcomes-based interventions are maximised when there are explicit links among the outcomes, goals, content, activities, measurements at several points, and a theory of change that explicates how the intervention can lead to results.

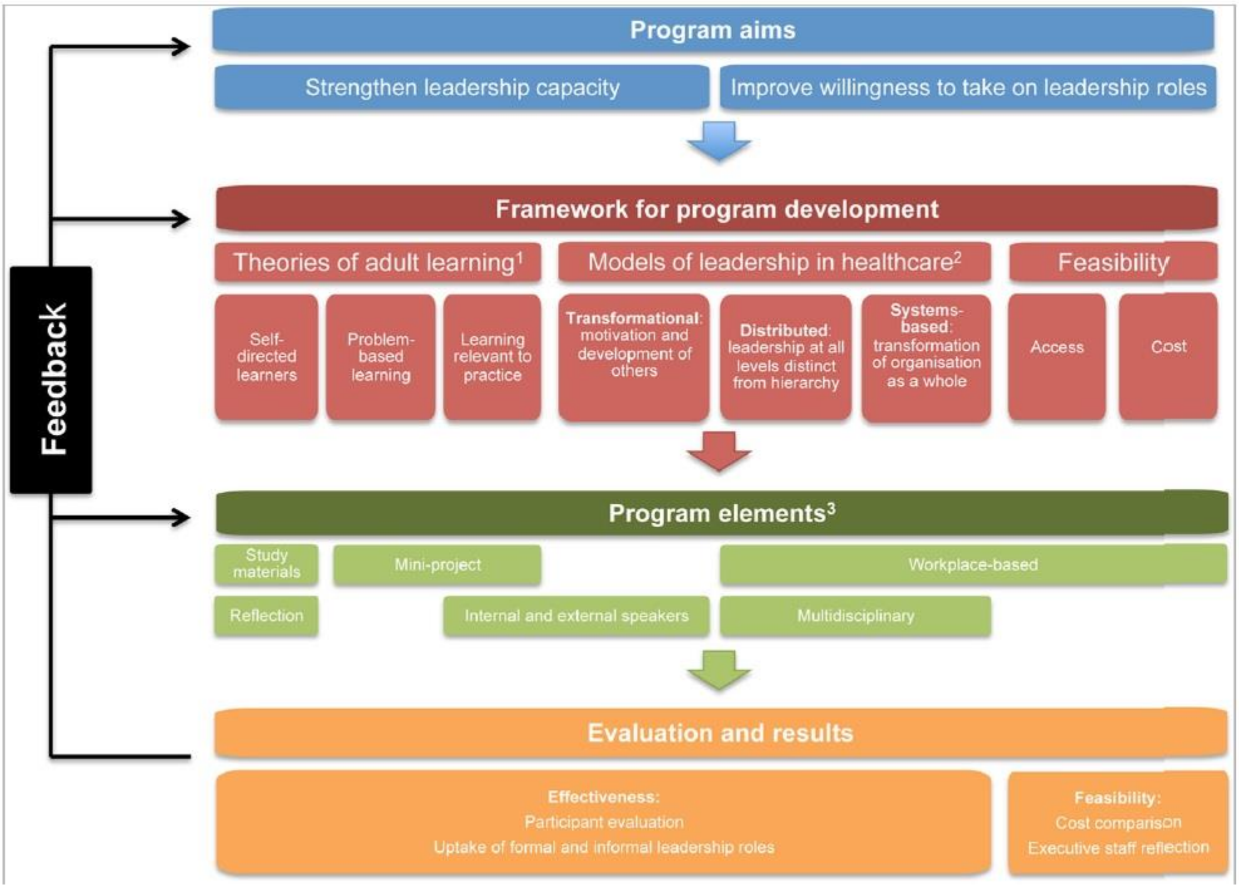

Feedback
Program aims
Strengthen leadership capacity
Improve willingness to take on leadership roles
Framework for program development
Theories of adult learning1
Models of leadership in healthcare2
Feasibility
Self-directed learners
Problem-based learning
Learning relevant to practice
Transformational: motivation and development of others
Distributed: leadership at all levels distinct from hierarchy
Systems-based: transformation of organisation as a whole
Access
Cost
Program elements3
Study materials
Mini-project
Workplace-based
Reflection
Internal and external speakers
Multidisciplinary
Evaluation and results
Effectiveness:
Participant evaluation
Uptake of formal and informal leadership roles
Feasibility:
Cost comparison
Executive staff reflection

Figure 6.3 MacPhail et al.'s Programme Structure.

Above is a depiction from MacPhail et al. (2015) that demonstrates the links among the programme goals, theoretical underpinnings, developmental activities, and evaluation metrics.

6. Incorporate the principles of adult learning, including personalising the development and measurement as much as possible

Many authors cite the value of incorporating the principles of adult learning mentioned in the previous section into the design of leadership programmes, which they believe contributes to positive programme outcomes (Blumenthal et al., 2014; MacPhail et al., 2015; Steinert et al., 2012; Ten Have et al., 2013). One principle to highlight again is personalised content and measurement. It has been said that this measure accommodates different learning preferences and enables participants to understand the relevance of the interventions to their organisational situation, which increases the likelihood that they will apply their learning (Blumenthal et al., 2014; Edmonstone, 2009; Leskiw & Singh, 2007; Van Aerde, 2013).

7. Acknowledge that certain developmental activities have stronger effects on particular learning outcomes than others and that variety is key

(Hartley & Hinksman, 2003; Leskiw & Singh, 2007; McCauley, 2008)

The next point follows from the second additional principle of adult learning, which is that leadership development programmes are maximised when the choice of developmental activities is tailored to suit the participants and the intervention objectives (Leskiw & Singh, 2007). Packages of developmental activities accommodate participants' different learning preferences and are most effective when combined according to their various primary functions. Miller et al. (2007) provide succinct examples of how certain activities are better suited to meet different types of goals. For example, they argue that action learning is useful for applying skills and forming collaborations; skill development seminars can facilitate the development of conceptual understanding, strategies, and techniques; and assessment tools and personalised coaching develop leadership style self-awareness and specific strategies to use strengths and counteract weaknesses (Miller et al., 2007). These assertions reinforce more detailed points made in conclusions from the best available evidence section, which also outline the benefits and application of various activities and how they can complement each other. The amalgamation of these methods is ideally a precise combination of formal training, experiential and work-based learning, and structured support, such as coaching and mentorship (Edmonstone, 2013; Frich et al., 2014; Hernez-Broome & Hughes, 2004; Leskiw & Singh, 2007; McGurk, 2010; Van Aerde, 2013; Zenger & Folkman, 2003). It is additionally helpful if the facilitators themselves are cognisant of the purpose and goals of their individual sessions in the context of the larger programme to provide enhanced continuity and relevance.

Likewise, many authors assert that integrating multiple learning methods is key to participant learning, given the diversity of learning preferences (Bergman et al., 2009; Blumenthal et al., 2014; Edmonstone, 2009; Leskiw & Singh, 2007; McCauley, 2008; McGurk, 2010; Miller et al., 2007; Ten Have et al., 2013). Frich et al. (2014) affirm that combining multiple sources of learning is likely to have the largest impact on leadership development programmes. Finally, offering developmental activities in isolation, such a series of lectures with no follow up, is not only suboptimal but doing so can potentially have negative outcomes. For example, negative feedback as part of a 360-degree report without coaching or mentoring afterward could leave participants feeling bitter or dejected, or they may choose to ignore it and shirk accountability for their development altogether. Therefore, to maximise the effectiveness of interventions, offering a variety of developmental activities is crucial, as is being aware of which goals each activity is best suited to address.

6.5.2 During Programmes

8. Ensure that participants can commit fully to the programme

The second set of factors contributing to effective leadership development consists of two factors that apply during programmes. The first is that participants must be able to commit fully to the programme or the outcomes suffer (MacPhail et al., 2015). In addition to missing content and the experience of activities, participants' attention and commitment can deteriorate if they skip sessions or cannot contribute fully. Group sessions can be undermined if numbers are uncertain and group morale can drop if participants are frequently absent. As one example of a common challenge in leadership development, clinicians in the MacPhail et al. (2015) study cited time pressures as the main barrier to participation, reporting having to juggle clinical, teaching, and research responsibilities and receiving no time off to pursue courses. Unlike military officers who have time allotted specifically for leadership development training, physicians and many other executives have several competing priorities that can impact their ability to participate fully in programmes (Korschun, Redding, Teal, & Johns, 2007). To address this concern, participants in the Hemmer et al. (2007) study had protected time and those in the Nakanjako et al. (2015) study had up to 80 per cent of their time to devote to their action learning projects and other programme components. Participants in the Korschun (2007) study, however, were not given protected time but still maintained good attendance and results. Finally, Satiani et al. (2014) mandated 75 per cent minimum attendance for successful completion. However this is orchestrated, it is important that programmes are structured so that participants benefit from the full experience with as close to full attendance as possible.

9. Include the process of goal setting, activity/experience, measurement, discussion/feedback, reflection, review and revision of goals, support, and repeat

The analysis of the combined included studies revealed a collection of actions, that when combined in a sequential, iterative process, has the potential to maximise the effectiveness of adult leadership development experiences. The progression that follows includes elements of Kolb's (1984) experiential learning cycle and Van Aerde's (2013) process for maximising the effectiveness of interventions,[4] as well as additional elements. This process is intended to apply to activities that are part of formal interventions, as well as informal development over the course of one's career.

[4] Kolb's experiential learning cycle is: experience, observation, abstraction, and experimentation. Van Aerde's process for maximising effectiveness of interventions is: experience, reflection, feedback, and further reflection.

For the sake of focus, motivation, and commitment, this progression begins with **goal setting**, referring to the goals of the whole intervention as well as the goals of individual developmental activities. It is helpful if the participants (and facilitators) are aware of how each activity contributes to the overall purpose and goals of the programme. The importance of goal setting and its link to improved outcomes has already been mentioned (Latham & Locke, 1983). Vague goals are rarely effective and it is commonly held that goals should be SMART: specific, measurable, attainable, results-based, and time-bound (MacLeod, 2012).

Once the goals are identified, participants engage in the **activity** or **experience**.

Following, and at times, during, the experience are pre-determined **measurements**, which can be conducted by a combination of participants themselves, peers, facilitators, or objective statistics, depending on the activity. For example, participants may rate how effectively they learned theoretical concepts from a lecture, facilitators may recount the accuracy with which target behaviours were exhibited during a role or simulation, or the metric may be represented by quantitative performance outcomes following an action learning project. While it may seem excessive to measure after each lecture, for example, the value of experiences with no reflection and application afterward seems to be decreased.

Following measurement, there is ideally structured **discussion** among participants, often led by a facilitator, and often in small groups. This step can also involve **feedback** from peers and most often from a facilitator based on the discussion or the activity itself. These opportunities can consolidate and expand participants' learning, solidifying their grasp of how it applies to their role or organisational context. Constructive feedback also can augment participants' self-awareness, assuage their worries regarding performance, and decrease stress associated with challenging assignments (DeRue & Wellman, 2009).

Participants can then be afforded time to **reflect** on what they have learned and its relevance to their professional roles and situation. They can then re-evaluate their goals, either to extend them further if good progress has been made or revise them to make them more manageable.

One key element that is useful throughout is **support** by way of coaching and mentoring. In addition to personalised guidance, coaches and mentors can ensure that this process of activities is a constructive and not a destructive one. Coaching is perhaps better suited for this purpose than mentoring, since the latter tends to be less formal and frequent; however, given the functional overlap between the two, they are both included.

Finally, the process can be **repeated** for subsequent activities with evolving personal and professional development.

Further considerations relate to the differences among participants. As mentioned earlier, Vygotsky’s (1978) Zone of Proximal Development refers to the space between one’s present developmental stage and one’s capacity limit, which can be exceeded with help from expert facilitators by way of guidance or scaffolding. The term “scaffolding” refers creating connections between a teacher’s knowledge and a learner’s existing experience and knowledge, allowing the teacher to facilitate improvement beyond the learner’s initial capacity. Vygotsky’s theory is important for two reasons: the first is that it places teachers in the role of supporting active learning rather than dispensing knowledge to passive learners, which is also in line with Freire’s educational theory and the principles of adult learning. The second reason for its importance is that it contends that different students have different developmental starting points and maximums. This is an assertion supported by findings from DeRue et al. (2012) regarding participants’ developmental challenge, Hassan, Fuwad, and Rauf (2010) regarding training motivation, and Suutari and Viitala (2008) regarding learning goal orientations, all of which suggest that individuals develop at different rates and have unequal ranges of peak development. Programme facilitators, coaches, and mentors should consider these findings when supporting the learning process in order to maximise the experience for participants without either underwhelming them or pushing them far enough beyond their limits that it becomes a negative experience. Another important factor raised by Ardts, van der Velde, and Maurer (2010) is that participants’ perceived control was found to be the most influential factor in determining positive organisational outcomes, which reflects points raised in the principles of adult learning section. Thus, while the facilitators, coaches, and mentors may guide the process, it is beneficial if participants function as co-leaders of their own development.

Following the sequence of actions described above, while recognising differences in participants’ rates and limits of development has potential to optimise the effectiveness of leadership development and activities.

6.5.3 Following Programmes

10. Ensure that there is proper organisational support and resources available after interventions

(Edmonstone, 2009, 2011; Fernandez et al., 2016; Gilpin-Jackson & Bushe, 2007; Jeon et al., 2013; H. W. Korschun et al., 2007; Leskiw & Singh, 2007)

Short Term, Unsustained Success

Despite the many factors that can contribute to making leadership development a transformational experience, it is what happens following programmes that often determines the extent to which learning can be successfully applied and have an effect on team and organisational outcomes.

As explained in the measurements section regarding the importance of post-post outcomes, this indicates that leadership behaviours can take time to be noticeable. Abrell et al. (2011) report that improved leadership behaviours in their study were not reported until six months after the intervention for supervisees and nine months for supervisors. Likewise, Day and Halpin (2001) suggest that organisational leadership needs to understand that dividends may take time to be realised. As well, post-post measures are important because positive outcomes reported at the conclusion of a programme are not necessarily sustained in the medium or long term. For example, Sanfey et al. (2011) noted a considerable decrease in participant self-reports over time in terms of their perceived ability to take on leadership roles (93 per cent short-term versus 69 per cent long-term) and their leader self-identity (89 per cent dropping to 71 per cent). The authors attribute this diminished confidence to difficulties applying the skills acquired during the intervention in the long-term and postulate that perhaps these skills required further nurturing or reinforcement to be sustained. Fernandez et al. (2016) surmise that a similar decrease in self-ratings from the post to the post-post test in their study was for the same reason. Similarly, Kwamie et al. (2014) reported that four of their five teams achieved their quantifiable clinical targets in the short-term, but these targets were not sustained in the medium term and the system returned to its prior equilibrium. Beer et al. (2016) suggest that one reason for these disappointing outcomes is that participants typically revert back to their old ways of doing things following programmes. Without a solid commitment to change in the form of a clear plan, outcome measures, and accountability, it is likely that previous habits will overcome efforts towards new, optimistic behaviours. For example, Santos and Stuart (2003) identified in their study that 64 per cent of managers reverted to their previous work styles after training. Thus, while the transfer of learning/application of leadership to the workplace is the ultimate goal of leadership development, it is also the classic challenge, particularly in the long-term.

Individual Motivation

On an individual level, there are a series of factors related to organisational culture that can increase participants' motivation to apply their skills. As mentioned previously, participants in several studies reported feeling impelled to apply leadership following programmes when their supervisors recommended them (Kwamie et al., 2014), when they believed that their organisation had invested in them (Gilpin-Jackson & Bushe, 2007), and when their colleagues had previously participated in the same programme (Gilpin-Jackson &

Bushe, 2007). Gilpin-Jackson and Bushe (2007) state that when senior leaders first participate in an intervention, their example afterward generates a supportive organisational culture towards leadership development.

For similar reasons, training transfer can also be increased when teams participate in leadership development together (Gilpin-Jackson & Bushe, 2007; Husebø & Akerjordet, 2016; Rosenman et al., 2014), which is in line with Day's (2000) distinction between leadership and individual leader training. For example, Dannels et al. (2009) stated that the reports of institutional impact were higher at organisations from which three or more participants attended than those of organisations from which fewer attended the programme they studied. This finding is perhaps because participants can develop their teamwork skills together, which could more naturally translate to the workplace. The relevance of their work context would likely be enhanced because they operate in the same organisational structure and culture, face the same challenges, use the same language, and have the same resources at their disposal. Thus, the discussions could more readily pertain directly to their professional context and the solutions and conclusions would be actual, not just abstract. Programme graduates from the same institution could encourage each other to implement their learning afterwards and hold each other accountable. When successful, Rowland (2016) argues, the development experience can serve as a vehicle that can positively influence the systemic dynamics of the organisation.

Lack of Organisational Support

Even when intervention graduates are enthused and committed to applying their leadership afterward, the most common explanation that studies cite for why this effort fails is an organisational culture that stifles participants' efforts (Gilpin-Jackson & Bushe, 2007; Malling et al., 2009; Rowland, 2016).

Brinkerhoff and Gill (1994) summarise the poison of negative culture well: "The workplace can untrain people far more efficiently than even the best training department can train people" (p. 9), quoted in Gilpin-Jackson and Bushe (2007). Beer et al. (2016) add that organisations need "fertile soil" in place before any "seeds" of developmental programmes can grow. The authors suggest that even following the most outstanding programmes, if the organisational culture is not conducive to change, the interventions will have little to no effect on organisational outcomes. Malling et al. (2009) speculate that the intervention that they studied failed due to a lack of organisational support. An often overlooked factor that can suppress change is the policies and practices created by top management (Beer et al., 2016), a factor which could be identified during the pre-programme barrier assessments discussed earlier. This is also another advantage of action learning: participants are given the opportunity to face possible barriers to implementation while the programme is under way and experiment with strategies to overcome them, incorporating the support offered by the programme while it is taking place.

Ways to Fix It: Cultural Attitude

Creating an organisational climate that is conducive to learning transfer can be facilitated by generating a cultural attitude that encourages and expects leadership innovation, particularly from those who have undergone formal leadership training (Dalakoura, 2010a; Gilpin-Jackson & Bushe, 2007). Following from the previous points, this collective support ideally comes from everyone in the organisation (Dalakoura, 2010a; Peters et al., 2011), particularly one's supervisor and one's peers. Peters, Baum, and Stevens (2011) assert that the direct supervisor is the key to the ROI of a leadership programme by actively encouraging and endorsing new initiatives and removing obstacles that might frustrate this effort (Leskiw & Singh, 2007). Without this support, the authors contend, little real evidence of training transfer is likely to occur. Gilpin-Jackson and Bushe (2007) report that participants in their study claimed that the most influential factor that facilitated leadership application was the need to believe that their actions are supported by others. McCauley (2008) takes this a step further, concluding that in addition to the importance of the support of senior leaders, training transfer is most effective when ownership of outcomes is shared and CEOs, senior leaders, and direct supervisors are accountable for the development of their supervisees (Ready & Conger, 2003). Gilpin-Jackson and Bushe (2007) clarify that a fertile workplace environment also means that one's colleagues must be willing to change themselves. They add that proximity to colleagues who are also applying leadership learning increases people's motivation to utilise new knowledge and skills, fosters an open and safe environment with a common language to discuss leadership ideas, and offers peer support and mentoring. It is helpful when these characteristics are reinforced by organisational practices, programmes, and policies that support individuals exercising leadership and launching new initiatives (DeRue et al., 2012). These factors can liberate participants to experiment more confidently and welcome accountability for the success of new initiatives.

Make Resources Available

The second component of organisational culture that is said to be necessary for facilitating training transfer is a collection of organisational resources that are made available as a clear indication that continuous leadership development is an integral aspect of the corporate strategy (Gilpin-Jackson & Bushe, 2007; Jeon et al., 2013; Leskiw & Singh, 2007). Organisations demonstrate this commitment by devoting monetary, technological, and personnel resources to supporting the application of learning and further training (Dalakoura, 2010a). Providing stipends, covering course costs, or offering alternate funding for time commitments devoted to leadership development are some such gestures (Van Aerde, 2013). Organisations can also evince this by allowing employees time to innovate, since one of the major reasons that participants claim that they do not implement learning is that they are too busy with their jobs and fall back into old habits (Gilpin-Jackson & Bushe, 2007). For example, the managers in Santos and Stuart's (2003) study cited time as the primary explanation for low transfer, as did those in the MacPhail et al. (2015) study. Likewise, in the Gilpin-Jackson and Bushe (2007) study, less than half of respondents claimed to have the time to apply their training. The authors question the extent to which this is an inescapable inhibitor or an excuse for not making a concerted effort to try to practice new leadership in the workplace. Further resources can include additional investments in formal leadership development, stretch assignments and job rotations, and ongoing coaching and mentoring. Thus, in addition to the social and professional support needed for leaders to apply their learning, a commitment of organisational resources to sustain and extend leadership development over time is necessary.

One of the most important elements for the application of leadership development to be successful is therefore an organisational culture that encourages participants to experiment with ways to attain their performance outcomes and career objectives and provides the necessary resources to assist this effort.

11. Evaluate effectiveness, hold participants and teams accountable, reward successes, and support improvement

(Leskiw & Singh, 2007; Peters et al., 2011).

As discussed in the first conclusion explored, to maximise the effects of leadership development, evaluating performance outcomes at the team, organisational, and clinical/benefit to clients levels is essential (McCauley, 2008; McGurk, 2010). These evaluations can take several forms and draw from multiple sources, but should ideally include pre-arranged quantitative measures and connect, as has been stated throughout, the organisational doctrine and needs assessment to the leadership development intervention and longer-term career projections. Again, it is beneficial if the post-programme assessments are built into the design of programmes and are announced to participants, their supervisors, and supervisees ahead of time so that the process, expectations, and measures are known at the outset (Edmonstone, 2013; Van Aerde, 2013). This evaluation can also be extended to apply to ongoing, development over people's careers. While this may seem like an obvious measure, Suutari and Viitala (2008) report that of the nearly 900 senior managers they surveyed, only 39 per cent had regular performance evaluations.

The next aspect that can encourage the transfer of learning following programmes is holding participants, teams, and others accountable to improve outcomes and following up accordingly (Coloma et al., 2012; Gilpin-Jackson & Bushe, 2007; Rose, 2015; K. E. Watkins et al., 2011). This practice signals that the outcomes expectations are being taken seriously (Edmonstone, 2013), as is continuous, through-career development (Leskiw & Singh, 2007; Zenger & Folkman, 2003). Being held accountable also makes it more likely that participants will request additional resources and support in order to meet their goals. Participants' supervisors might be more likely to provide such resources and support if they too are held accountable for their participants' performance. Accountability not only validates the importance of the outcomes, it demonstrates a confidence and investment in the participants, through the expectation that they will succeed and grow professionally. Gilpin-Jackson and Bushe (2007) suggest that goal-setting and feedback mechanisms are useful strategies to support the long-term maintenance of learned capabilities and avoid relapse (Richman-Hirsh, 2001). Thus, it is said that shared accountability among colleagues for improved outcomes increases the likelihood of achievement.

Another related factor that can contribute to maximising the outcomes of leadership development is a corresponding system of rewarding successes and improving on underperformance (Dalakoura, 2010a; Gilpin-Jackson & Bushe, 2007; Peters et al., 2011). The former includes public recognition for having completed leadership development programmes (M. E. Green, 2002) and meeting individual, team, and organisational targets. Although they are beyond the scope of this thesis to discuss at length, the advantages and drawbacks of individual versus team reward structures, as well as intrinsic versus extrinsic motivation, are interesting topics for organisations to consider. In addition to rewards, having systems in place to provide sufficient remediation when expectations are not being met is also valuable (Peters et al., 2011), a practice that again demonstrates a human resource and organisational commitment to employees. Follow-up can involve further training, expert and/or peer coaching, and mentoring.

Therefore, having structured evaluation following programmes, holding participants and colleagues accountable for outcomes, and providing rewards and support ensures that the transfer of learning is an expectation and is more likely to be successful.

12. Provide formal and informal follow-up opportunities that continue after programmes

(Edmonstone, 2009; Gilpin-Jackson & Bushe, 2007; Sanfey et al., 2011; Satiani et al., 2014).

Building on the previous point, a factor that can undermine the effects of leadership development is a dearth of follow-up opportunities after interventions (Beer et al., 2016; Leslie et al., 2005; Steinert et al., 2012). For example, although Bergman et al. (2009) list many benefits of their one-week intervention for first-line managers, they concede that one may question the sustainability of the effect of such a short programme, especially without follow-up opportunities. Similarly, despite many reports in the Korschun et al. (2007) study of increased aspirations to lead, engagement, commitment, and skills, some fellows allegedly felt disappointment at the lack of opportunities for advancement, encouragement, and support following the intervention. Participants in the Edmonstone (2009) study mentioned that a drawback of the programme was the lack of measures in place to continue to apply their learning after the intervention. Furthermore, only 23 per cent of respondents in the D'Netto, Bakas, and Bordia (2008) study indicated that there were post-programme development options in the workplace. To avoid this pitfall, in-house providers and stakeholders, along with participants and their supervisors, can consider how the application of leadership development can endure and increase beyond the programme (Edmonstone, 2009; McCauley, 2008).

Follow-Up Methods

There are several follow-up methods that can cultivate the long-term application of new leadership knowledge and capabilities. One mechanism meant to connect interventions to post-programme opportunities formally, as mentioned previously, is a Personal Development Plan (PDP) (Fernandez et al., 2016; Korschun et al., 2007; Steinert et al., 2012). As an example, the leadership programme described in the Gilpin-Jackson and Bushe (2007) study devoted a session at the end to planning application strategies in the workplace and related goal setting, the latter of which is said to have proven to lead to successful training transfer (S. K. Johnson et al., 2012). Another follow-up idea raised by managers in the Gilpin-Jackson and Bushe (2007) study is to host formal refreshers for programme graduates to reignite interest in leadership application and offer networking opportunities and further development. Ongoing goal setting, 360s and formal feedback, and coaching and mentoring can also contribute to sustained leadership following development (Gilpin-Jackson & Bushe, 2007; Ladyshewsky, 2007; Richman-Hirsh, 2001). As mentioned previously, Bowles et al. (2007) assert that participants who received coaching in their study demonstrated significantly higher quota achievement than those who relied on on-the-job learning with no coaching. In addition to making further opportunities available, organisations do well to also ensure that promising leaders have the *time* to undertake and profit from them. For example, the fellows in the study claimed that they were so busy with their day-to-day work that they found it difficult to pursue additional learning opportunities to further develop their leadership skills (Fernandez et al., 2016). Therefore, goal setting, supported by 360-degree feedback, coaching, and mentoring are keys to furthering the application of leadership learning.

Rose (2015) advocates extending the scope of leadership development to the span of employees' careers as an organisational priority. Implementing this can include elements from the previous paragraph, as well as other formal and informal means. The first way this can happen is by encouraging leadership programme graduates to pursue more advanced leadership development programmes and even, as Rose (2015) suggests, formal qualifications, so that all leaders have similar experience and training across the healthcare system. Leaders can take part in further multi-source feedback (MSF), specific to their stage of career, supported by coaching and mentoring (McCauley, 2008). As mentioned earlier, formal evaluation is an integral part of this support that can keep participants motivated and aware of the progress they are making, as well as of what assistance that they need to improve. Regular MSF feedback and performance measurements are also said to strengthen a sense of shared accountability and continuous improvement (Leskiw & Singh, 2007; Zenger & Folkman, 2003). McCauley (2008) adds that high-potentials benefit from specific feedback and programming that is respectful of their status. Additional post-programme stretch assignments or challenging projects, particularly when tied to performance improvement measurements, are thought to further enhance training transfer (Gilpin-Jackson & Bushe, 2007; McCauley, 2008; Peters et al., 2011). Therefore, more advanced leadership development interventions, MSF, and stretch assignments, supported by evaluation, coaching, and mentoring are effective ways to expand leadership development following individual interventions.

Alongside formal leadership development programmes and follow-up measures, informal opportunities can positively contribute to leaders' development and the best organisations are said to arrange both. Proponents of this idea include Noel Tichy, professor of Organisational Behaviour at the University of Michigan Business School, who says that 80 per cent of an executive's development can be attributed to on-the-job and life experiences, while formal training can affect only 20 per cent (Kesner, 2003). This suggests that formal and informal opportunities can be used effectively in a complementary way. Gilpin-Jackson and Bushe (2007) profess that in a favourable organisational environment, many participants are likely to create their own opportunities to apply their learning. Participants in several studies reported networking benefits, having developed collaborative relationships, and mentoring opportunities as valuable outcomes that can serve as ongoing resources and support long after programmes finish. Steinert et al. (2012) add peer coaching as an additional resource that can be introduced in an informal capacity. Finally, in the Korschun et al. (2007) study, participants praised the networking benefits of the programme, attesting that they found it easier to seek advice or establish collaborations with peer leaders in other parts of the organisation following the intervention. These informal means of continued leadership development can complement formal interventions, often without a cost to the organisation.

Therefore, in line with optimising the programmes themselves, to maximise the benefit of leadership development interventions, it is essential that an organisational culture is nurtured that encourages and supports the application of leadership after programmes. Ideally, each leadership programme has a distinct role as part of the larger organisational strategy. This process begins with a needs assessment, which is followed by the providers selecting appropriate outcomes and metrics that emerge from the needs of the organisation. Finally, resources should be made available during and after the interventions to support the ongoing application of leadership. Table 6.2 below summarises the points from the second conclusion explored.

Table 6.2

Summary of Conclusion Explored Two

Conclusion Explored 2: Factors that Facilitate the Successful Application of Leadership	
Pre	
1	Ensure the organisation's doctrine regarding the understanding of leadership and its capability framework is clear, shared, and pertinent, including for different levels of seniority
2	Conduct a needs and barriers assessment, select participants accordingly, and align participants' developmental goals with the organisational strategy
3	Generate organisational support, including from the senior management, and involve stakeholders in the design and, at times, the delivery of programmes
4	Ensure that there is a common understanding of the programme purpose, goals, content, outcomes, and measurements among the provider, participants, and the organisation
5	Ensure that there is a connection between the outcome goals, content, activities, and programme evaluation and that measurements are collected pre, during, post, and post-post intervention
6	Incorporate the principles of adult learning, including personalising the development and measurement as much as possible
7	Acknowledge that certain developmental activities have stronger effects on particular learning outcomes than others and that a variety is key
During	
8	Ensure that participants can commit fully to the programme
9	Include the process of goal setting, activity/experience, measurement, discussion/feedback, reflection, review and revision of goals, support, and repeat
Following	
10	Ensure that there is proper organisational support and resources available after interventions
11	Evaluate effectiveness, hold participants and teams accountable, reward successes, and support improvement
12	Provide formal and informal follow-up opportunities that continue after programmes

Above is a table summarising the second conclusion explored, which describes the factors before, during, and after leadership interventions that facilitate the application of learning.

6.6 A Model of Leadership Programme Design and Evaluation

This section describes a prototype theoretical model of leadership development programme design and evaluation that, as mentioned previously, is the product of an in-depth exploration as part of the systematic evidence analysis. Many of the principles from the conclusions of the best available evidence and the conclusions explored appeared to fit together in a cohesive, sequential structure that could inform the process of design, delivery, and evaluation and potentially enhance the impact of programmes. While each of the points in the outcomes-based model derives from research, the model itself has not been tested per se. It was thought that in addition to principles of best practice and evaluation, as well as pitfalls to avoid, offering one unified model could be useful to practitioners and academics alike. Given the massive global investment in leadership development and the frequent claims, such as those by D'Netto et al. (2008), that the quality or yield of programmes is low, there is a clear need for better guidance when designing programmes and better evidence that programmes are having a positive, tangible impact. Since the currency of success in leadership development is application, which is measured most explicitly by outcomes, it is appropriate that they form the heart of the model.

There are three sections below: the first is a description of a suggested sequence of steps involved in preparing and implementing an outcomes-based leadership development programme. The second section lists the advantages of the approach. The third section discusses further considerations.

6.6.1 Outcomes-Based Stages of Implementation:

1) The first step in designing an outcomes-based leadership intervention is determining the post-programme **desired outcomes** at the individual, team, organisational, benefits to patients, and possibly economic levels (Dale, 1969; Fayolle & Gailly, 2008; Nabi et al., 2017). These can be categorised according to the Kirkpatrick model and are most effective when they go beyond Levels 1 – 3 (individual outcomes) to Levels 4a and 4b (organisational and benefit to clients outcomes). As mentioned previously, this can involve input from various stakeholders, including the participants themselves, in line with the self-direction principle of adult learning. These can also reflect the results of an organisational

needs assessment, the organisational leadership doctrine, and can be aligned with the overall organisational strategy.

2) These final outcomes then either become, or majorly inform the choice of, the programme **goals** and **objectives**.
3) From there, the content, logistical **components** (such as length), developmental activities, faculty, and forms of evaluation are decided according to how they contribute to achieving the final outcomes, individually and in combination. As outlined in the sixth point of the first conclusion explored, different developmental activities have different key and complementary functions, and design choices should reflect these factors. Assembling programmes intentionally is likely to maximise their impact. Finally, the **role** and **goals** of each component can be determined, on their own, as well as how each contributes to the overall programme and its outcomes.
4) As described in the second conclusion explored, leadership development and evaluation are enhanced when **organisational culture** is taken into account, given its potential to significantly affect outcomes. One key factor to repeat is that having participants' supervisors accept accountability for the former's performance makes it more likely that they will agree to reasonable goals and provide the necessary support and resources to make attaining them successful. Conducting a barriers assessment can be helpful to troubleshoot issues of implementation before the programme begins.
5) The post-programme outcomes and all other intervention components can then be **announced** to stakeholders, facilitators, and participants alike before the programme begins. This information makes it clear how each aspect contributes to the intervention goals, making the role and relevance of each component explicit.
6) **Evaluation** should begin in certain forms at **baseline** as a point of comparison. This can involve self-ratings of confidence, knowledge, and skills, personal goals, 360-degree feedback reports, and organisational, clinical, or statistical data which will be collected again later. **Self-ratings** are helpful for establishing ROI and are useful for providers, participants, and researchers; their usefulness in research is however limited unless they are reinforced by objective data. Self-ratings are also helpful as tools to facilitate personal development activities, such as coaching, where the coach and participant can discuss the scores as the basis of their work together. **Programme-wide outcome metrics** are useful for researchers for comparative

purposes and demonstrating ROI in a more objective manner. Thus, it is helpful to have a combination of individually-selected goals, which are useful for individuals but restrict research comparisons given their heterogeneity, and standardised metrics, which are less personal but useful for comparative purposes. Finally, objective outcome metrics, statistical data, and individually-selected goals are useful for participants' own sake in terms of motivation. These goals could become part of each participants' PDP, which can map out their development and carry forward after the intervention has finished.

7) Evaluation ideally takes place as the intervention **progresses** to chart progress and identify when further support, resources, or goal modification might be necessary. Participants can share their results with facilitators, coaches, and peers to enhance their development while those resources are available to them. Ideally, participants would be granted structured **reflection** time to complete these assessments and to consider feedback they have received.
8) Evaluation occurs again immediately **following** the intervention and a minimum of six to nine months afterwards. This practice allows time for the application of learning to occur and to ensure that changes are sustained. Application is maximised when accompanied by goal setting, coaching and mentoring, and the availability of resources to support continuous development.
9) Providers can then use the information collected to **refine** their programmes and generate a pool of outcome measures that can be offered to subsequent participant samples for their consideration.

The **forms** that this evaluation can take in practice is a variety of the following:

- short **quality control** surveys after each day or module soliciting feedback on the quality and relevance of each component in terms of meeting its intended goals,
- **structured reflection** following each module on how activities relate to participants' own development and situation, as well as how learning can contribute to achieving their outcomes. For example, participants who attend a workshop on high-pressured decision making can be granted the opportunity to reflect on which workshop principles or discussion points resonated with them and how they could implement them in their own future decision making.
- **formative reviews or assessments** of their progress with facilitators, coaches, or peers. This can involve discussing the results of 360 reports or feedback from peers

or experts, participants' perceived developmental progress according to their goals, or statistical outcomes. The discourse can involve participants' disclosing feelings about their results and progress, celebrating successes, postulating reasons for underperformance, discussing strategies for improvement, and possibly revising their individual goals.

- **summative assessments** that tie together the results of the outcomes at all three levels. Organisations and participants can discuss together what should go in this assessment from among the various data options, including personal development based on comparative self-assessment from baseline to post-post, reports from coaches and facilitators, skill development as evidenced by self-reports and possibly 360s, and the success of action learning projects. Part of this assessment could also be the impact at the organisational and benefit to clients levels, reinforced by objective and statistical data. A final component of this assessment could be further goals at all three levels, as well as opportunities to undertake as part of an evolving PDP.

6.6.2 Support for the Model

It has been stated throughout that application is the goal of leadership development, which implies that outcomes are implicitly and necessarily intended. The definition of leadership presented in chapter two depicts leaders and team members working towards outcomes together, which implies that the impact of leadership development extends beyond the leader to others and ideally to the organisation. The outcomes-based model presented here makes this fundamental purpose explicit and ensures that each aspect of the programme design and assessment clearly points back to it symbiotically.

More specifically, there are eight **reasons** why the outcomes-based model has the potential to be an optimal approach to leadership development:

1) It gives the intervention **purpose** and **direction** and focuses participants' and facilitators' attention on attaining the outcomes
2) Transparently focusing on outcome attainment reinforces **accountability** of participants and to some extent facilitators and participants' supervisors as well
3) Evaluating programme components is useful for **quality control** in terms of developmental activities, facilitators, and logistics (location, length etc) to ensure that the programme is meeting its intended goals and at a high standard

4) Evaluating programmes as they progress enables providers the opportunity to **modify** and **adapt** them according to participants' and facilitators' feedback to maximise their impact before it is too late to do so
5) There is evidence that structured reflection time reinforces the essential **application** aspect of leadership development by giving participants an opportunity to contemplate how each activity relates to the relevant goals and outcomes, as well as to their individual leadership situation and organisational context
6) Tracking outcomes provides evidence of whether **progress** is being made towards these outcomes or if further support or goal revision is needed. This is most helpful when it is an ongoing process and begins before the end of the programme while facilitator support is still available
7) The previous steps make it more likely that the outcomes will be achieved than if they are unspecified or if participants are not held accountable. Achieving outcomes is evidence of the **return on investment**, which is of interest to stakeholders, particularly those funding the endeavour
8) This form of structured evaluation and support can effectively enable progressive **through-career development** is based on an organised data set with clear trajectories.

6.6.3 Further Considerations

This final section outlines a series of other considerations based on the points raised earlier in the chapter. The first is that it is beneficial if participants are involved in selecting or personalising some of the final outcomes, specific to their leadership and organisational situation. The less abstract and the more tailored to the people and their immediate needs the outcomes are, the more likely they are to be successful. The second is that ideally these outcomes extend beyond individual outcomes to organisational and benefit to patients levels. Third, collecting objective data to measure performance adds credibility to the operation. Fourth, goals are most effective when they are specific, measurable, attainable, results-based, and time-bound (SMART). It is important to build time for participants to complete evaluations during the intervention, rather than leaving them to do so in their free time, which can often detrimentally affect the response rates. It is also helpful to explain the purpose of each form of evaluation and how it will be used. Once the data is collected and analysed, the process is enhanced when providers share the results with the facilitators and participants, as well as modifications that are being made based on the feedback. This demonstrates to participants that there is value in the process and that the providers are committed to maximising the experience and its impact.

The prototype outcomes-based theoretical model described is based on the analysis of the best available evidence and offers a sequence of principles of effective design, delivery, and evaluation of leadership development which has potential to optimise leadership development programs. Attention in the following chapter turns to the implications for research and practice, as well as the discussion.

7 Chapter Seven: Implications for Practice and Research, Discussion, Limitations, and Strengths

This chapter begins by describing the implications for practice and research, which are informed by each stage of the systematic evidence analysis. The implications for research derive from the close analysis of elements of effective and ineffective study design characteristics identified in the extant reviews and MULTI and HEE included studies. They also highlight areas of need for further research. This is followed by the description of a feature article, which is the only randomised control study in the combined included studies data set and includes three valuable elements of effective research not seen in other studies. The final part of this section is a critique of MERSQI, which identifies potential revisions to optimise its usefulness. The discussion follows, beginning with speculations regarding why the leadership development evidence base is so limited and what can be done to ameliorate the situation. The section then presents a summary of the overall thesis findings and conclusions, organised according to the research sub-questions. The discussion concludes with the limitations and strengths of the current study. With this, the attention now turns to the implications for practice.

7.1 Implications for Practice

The goal in selecting the systematic evidence analysis (SEA) methodology was to provide a transparent and credible analysis and isolate "what we know" concerning optimal leadership development for professionals, as well as "based on what evidence." The intention was to enable readers to decide for themselves how they value the findings and conclusions, which is the approach recommended by the PRISMA guidelines (Liberati et al., 2009). The two conclusions explored that the SEA methodology uncovered are interwoven into the points below. Applying this knowledge is said to enhance the outcomes of leadership development and, in some cases, predict and avoid situations that can contribute to programme failure. The nuances and suggestions for implementing the principles described in the conclusions explored provide further insights into the phenomenon. Likewise, the prototype of the theoretical model married the theoretical and empirical findings in a way that those designing programmes can use to plan or refine their interventions. The implications for practice below are the culmination and application of these three sections, the details of and references for which have been provided in the preceding chapter. These points appear to be equally relatable to through-career leadership development as they are to individual interventions, as well as to both in-house and external programmes. The following implications for practice are not meant as prescriptions; they are presented as considerations for practitioners based on the findings of the best available evidence.

Each bullet point below finishes the following sentence:

The academic literature suggests that the impact of leadership programmes is enhanced and outcomes are improved when designers and/or organisations

7.1.1 Background

- Consider the pre-programme factors from the second conclusion explored that facilitate the successful application of leadership
- Involve stakeholders in a needs and barriers assessment and the design of the programme, as well as the delivery at times, and embed leadership development as an integral part of the organisational strategy
- Consider the purpose of the intervention and the need it is intended to fulfil as a starting point
- Appreciate how much organisational culture affects the application of leadership and find ways to create a supportive, nurturing environment for participants. This ideally begins before interventions start by engineering stakeholder buy-in and involvement, as well as by having teams from the same workplace.

7.1.2 Design

- Begin with the desired outcomes at multiple levels (individual, organisational, economic, and benefit to clients), which then translate into or inform the programme objectives
- Outline clear and explicit connections among the desired outcomes, curricular goals, content, programme activities, and measurements
- Personalise the aspects of programmes mentioned in the previous point by enabling participants to be involved in selecting or adapting them
- Cater to different learning preferences by offering multiple developmental activities
- Are cognisant that certain developmental activities serve different functions and are better suited to address different types of goals
- Incorporate the principles of adult learning by valuing participants' experience as a developmental resource, respectfully addressing biases, providing practical and relevant content, and embedding application and experiential components
- Ensure that participants can commit fully to the programme.

7.1.3 Developmental Activities

- Make use of workshops followed by videotaped simulations and expert feedback to improve observable behaviour in task-specific skills. They can also develop problem-solving, adaptability, teamwork, and leadership skills, such as communication, as well as enhancing self-awareness

- Include action learning projects in medium and long programmes to enable participants to apply learning as the intervention progresses and to target outcomes at the organisational and benefits to patients/clients levels. These are best when accompanied by coaching and mentoring, the quality of which is said to depend on the quality of the coaches and mentors
- Add 360s, coaching, mentoring, and other support systems to facilitate project completion, skill development, and increase self-awareness
- Incorporate lectures for theoretical, conceptual, or practical information regarding the organisation, its protocols, or its situation
- Include case study analysis to enable participants to consider how theoretical, conceptual, or practical information applies to their leadership situation, which develops problem-solving and systems thinking skills, among others
- Include workshops for specific skill development
- Enable participants to apply learning from didactic sessions in simulations, role plays, and action learning projects
- Follow didactic and experiential activities with discussion and structured reflection to reinforce the relevance to the workplace
- Structure activities according to the process of goal setting, activity/experience, measurement, discussion/feedback, reflection, review and revise goals, support, and repeat
- Enable participants to increase their self-awareness, regardless of their previous experience or level of seniority, which can be addressed through videotaping simulations, providing peer and expert feedback, personality tests, coaching, mentoring, and time for structured reflection.

7.1.4 Sample

- Consider the impact of motivation as a precursor to adult learning and find ways to ensure that participants are interested in developing
- Select participants for leadership development intentionally, whether based on their role, attitude, or potential
- Are aware that medical residency, interdisciplinary, and profession-specific leadership programmes can be effective, but perhaps in different ways. Providers can consider

experimenting with different variations of these to ascertain what works best for them and in what circumstances

- Consider having teams attend together
- Consider adding domain-, profession-, or level of seniority-specific breakout or syndicate sessions to increase the perceived relevance of the discussions when programmes involve a mixed population.

7.1.5 Measurement

- Set outcome metrics at the organisational (Level 4a) and benefit to patients/clients (Level 4b) levels with participants' involvement, in addition to goals at the individual-level (Kirkpatrick Levels 2a – 3b)
- Include other raters, objective outcome measures, and statistics to add more credibility than self-reports alone, as well as collecting qualitative data to explore the nuances of leadership development
- Ensure that measurements begin at baseline, extend a minimum of six months after the completion of the intervention, and are accompanied by feedback and support for participants
- Collect data during the programme that can serve three functions: a) prompting participants to reflect on how each programme component relates to their organisational and leadership situation, as well as assessing their own developmental progress, b) by evaluating or receiving feedback on their progress, participants can use this data to facilitate coaching and mentoring relationships. This can be helpful to extend one's skills when progress is good, as well as offer appropriate support when development is falling short of expectations, and c) by using anonymous feedback or PPE data on quality control, providers can modify aspects of the programme as it happens, which is enhanced when participants are made aware of how this information is being used to adapt the programme
- Hold participants, teams, and supervisors accountable, reward success, and remediate underperformance
- Consider making outcome results public, particularly those at the team, organisational, and benefit to clients levels, as long as doing so is seen as constructive.

7.1.6 Follow-up

- Support participants, hold them accountable to achieve their goals, and continue to expand and enhance their leadership following programmes. This can be done through ongoing performance measurement and pursuing further developmental opportunities, both formal and informal

- Consider providing monetary, technological, and personnel resources to support further development, as well as allowing participants *time* to experiment and develop
- Use PPE data to refine programmes so that they are constantly evolving
- Align leadership programmes with organisational strategy, elucidate the connections among interventions, and subsume individual interventions within a larger through-career developmental context.

The preceding points are practical applications of principles of optimal leadership development that are derived from the empirical evidence. These relate to designing, delivering, and evaluating programmes, as well as related considerations regarding principles of adult learning and organisational culture that can affect the transfer of leadership learning to the workplace. The information regarding the cardinal functions of developmental activities and how they can complement each other is further intended to aid providers in designing programmes intentionally to meet certain goals by selecting the tools that are best suited to making this successful. The suggestions regarding effective ways of measuring leadership, particularly following programmes, are included because they appear to be intimately connected to design and practice that leads to enhanced outcomes. No definitive evidence emerged regarding optimal choices for other programme components, such as size, length, structure, or location of programmes, faculty characteristics, or ideal sample constitution. Discussion of the advantages and examples of the application of each of these factors has been offered throughout when information allowed.

Thus, the implications for practice are a fusion of the best available evidence, tailored and extended theories of adult education, the collection of considerations that are reported to significantly affect the success of leadership development, and the outcomes-based model of leadership development.

7.2 Conclusions: Research

Throughout the analysis of the findings of the included studies listed in chapter five and those of the analysis of the extant reviews, sets of identifiable methodological characteristics emerged that increased or decreased the credibility of studies and the usefulness of their conclusions for the reader. These have been incorporated into the conclusions and implications for research below.

7.2.1 Included Study Flaws or Limitations

The following set of characteristics represents flaws in studies and reviews that lessened their credibility or limited the usefulness of their findings and conclusions.

For individual studies:

- Leaving out important details of the study, data collection instruments, faculty, sample, or intervention, which limits readers' ability to judge the generalisability of the case and the quality of the findings
- Not reporting the frequency or representative nature of qualitative responses, leaving the reader to wonder the extent to which those reports were shared or representative
- Reporting only favourable highlights or majority opinions (Malling et al., 2009) and not including outlying responses (Alvesson & Spicer, 2012), which could offer valuable nuances
- Failing to include a control group or comparable statistics, such as national averages, to contrast participants' reported outcomes, limiting the comparative magnitude of improvements
- Relying exclusively on self-reports and Kirkpatrick levels 1 – 3a outcomes without any objective data to substantiate them (Malling et al., 2009). The negative correlations between MERSQI groupings and qualitative data only and Kirkpatrick Levels 1 – 3a reinforce why this is insufficient. For example, 73 per cent of the HEE studies that were in one of the two low MERSQI groupings omitted Kirkpatrick Level 3b (objective) outcomes. The limitations of self-reported data and restricting evaluation to the individual-level has been described
- Failing to differentiate between programme-wide outcome metrics and individually reported outcomes and benefits, which calls into question the extent to which the benefits were shared or representative
- Focusing exclusively on the individual-level of measurement (Levels 1 – 3b) without attempting to measure outcomes at the organisational (Level 4a) or benefits patient/clients levels (Level 4b)
- Relying exclusively on a post or post-post measurement with no baseline to track net change, a limitation which the examples of studies in which post self-ratings actually dropped when compared to identical baseline ones make clear
- Not including a post-post measurement to assess the application of learning or whether it was sustained, which is problematic as evinced by studies described in chapter five (Abrell et al., 2011; Beer et al., 2016; Kwamie et al., 2014)

For literature reviews:

- Neglecting to conduct a review of extant literature reviews as background and to incorporate the results of the analysis into the findings, which strengthens the design of the review and joins the review's findings with existing knowledge
- Omitting details, such as of the sample, which limits layers of analysis
- Failing to publish a table of key details of the included studies, which allows readers to cross-check information and decide the extent to which the studies relate to other contexts
- Not distinguishing between subjective and objective outcome data
- Not analysing the relationships among variables, including programme components and developmental activities, which is necessary given the complex nature of leadership development
- Not closely critiquing individual studies' programme evaluation, the result of which can affect the clarity and strength of the review's conclusions
- Not rating or ranking the credibility of included studies, which fails to offer a sense of the quality of evidence supporting each study's conclusions, and those of the review, by extension
- Providing only the final score for the critique of each study, not the score for each aspect, which limits the transparency of the assessments
- Not critiquing the included studies' reported outcomes, which takes their findings and conclusions at face value, without judging the quality of the results
- Providing only highlights of reported outcomes, rather than the full data set, including negative or outlying reports. The former increases transparency and credibility and the latter unveils helpful nuances
- Grouping all included studies' conclusions together, rather than tiering them according to the calibre of the evidence

Many of the flaws in the extant reviews cited above are related to those of the included studies, suggesting that issues in the current state of the literature are common at the systemic and individual levels. This also indicates that the solutions offered below have the potential to influence both concomitantly.

7.2.2 Study Characteristics that Enhanced Studies' Credibility and Usefulness

It has already been mentioned that the fact that only three HEE studies received a score of 13 or higher out of 18 on the MERSQI scale, coupled with the mean score of 9.94, which places it in the anecdotal calibre category, suggests that there is a definite lack of strong evidence in the field and that the quality of leadership development research needs to improve. To this end, a set of methodological characteristics emerged that enhance the credibility of the included studies and reviews and increase the usefulness of their findings and conclusions for readers. When MULTI or EMD reviews raised similar points, citations have been added.

For individual studies:

- **Full reporting** of the details of the study, sample, intervention, and findings to increase transparency and the ability for readers to judge its generalisability
- Comparing participants' performance outcomes to a **control group,** national averages, or relevant statistics (Steinert et al., 2012) to better illuminate the magnitude of the improvement
- **Using multiple iterations and sites** to enhance the generalisability
- **Collecting quantitative, objective data** to show evidence of outcomes, which was found to be statistically significantly correlated to the most credible MERSQI groupings. Adding other raters can also be valuable (Malling et al., 2009; Steinert et al., 2012), both of which can add credibility to the reported outcomes
- **Collecting qualitative data** to examine the nuances of to what extent, in what circumstances, and for whom leadership development and its components can be effective (Kwamie et al., 2014; Steinert et al., 2012; Straus et al., 2013)
- **Using multiple data collection methods** to allow for data triangulation (Steinert et al., 2012)
- Collecting measurements at **all of the Kirkpatrick levels**, including 3b (objective behaviour change), and particularly 4a (**organisational**) and 4b (**benefit to patient/client**) **outcomes** (Rosenman et al., 2014; Steinert et al., 2012)
- **Combining a pre or baseline measurement with a post-post measure** (Steinert et al., 2012), the latter of which is ideally collected six to nine months minimum following an intervention. This allows for evidence of the application of learning, as well as the extent to which it is sustained over time, both of which are more effective when compared to an initial rating
- **Collecting measurements during programmes** to add valuable data for providers, participants, and facilitators alike, as discussed in the previous chapter.

For literature reviews:

- Conducting and describing a literature review of extant reviews to inform the design of the review and place the findings and conclusions in the context of existing knowledge
- Employing multiple researchers to enhance credibility and minimise bias
- Searching multiple databases to account for the multidisciplinary nature of leadership development
- Including all details of each article's study, sample, programme, and evaluation to increase layers of analysis
- Presenting all the above information in a table to increase transparency and enable readers to assess the generalisability
- Analysing the relationship among variables, such as programme components and developmental activities
- Applying a validated instrument to critique the credibility of each study and publishing the full results to increase transparency
- Separating types of outcomes levels (perhaps according to the Kirkpatrick model), such as individual, organisational, and benefit to clients
- Separating subjective from objective outcomes
- Critiquing the studies' reported outcomes to clarify the evidence supporting each of their findings and conclusions
- Providing the full data sets, including negative or outlying responses, not only highlights. Full sets increase transparency and give an indication of the representative nature of results, as well as helpful nuances provided by outlying opinions
- Providing tiered overall conclusions based on the calibre of evidence.

Before moving on, it should be stressed that although some of the points above are not astonishingly novel, they are not being applied consistently at the review and the individual study level, which is significantly affecting the calibre of the research in the field.

7.2.3 Feature Study: An Example of Three Key Components

During this research, one article in particular stood out as having the potential to fill three important gaps in the research findings. One reason for this is that it was the only randomised controlled trial in the included study combined data set. It is described here, following the characteristics that enhanced studies' credibility, because of how effectively it applied these characteristics, in some ways more effectively than any of the other studies. As mentioned earlier, experiments with a balanced and representative control group can help isolate variables and establish causal relationships between the programmes and outcomes. The challenges associated with securing a control group for professional leaders, as well as with isolating one particular aspect in a complex, multi-faceted intervention like leadership development to assess its impact have been explained (Edmonstone, 2013). Jeon et al. (2013), however, chose an interesting methodology – a cluster randomised controlled trial – and given its usefulness, particularly in healthcare research, it is surprising that only one of the combined included studies employed this design.

The programme they analysed featured a year-long, evidence-based leadership intervention involving action learning projects, 360-degree feedback, case study analysis, and one-on-one interactions with a programme facilitator. The goal of this programme was to develop managers' leadership capabilities and support the delivery of improved quality healthcare. The study confirmed 24 experiment and control sites in total (12 of each), thus the cluster site was the unit of randomisation. Members of the experiment group participated in the intervention at their own workplace, while the control group participants did not receive any leadership training during the time of the study but agreed to provide the usual care. The treatment allocation was not disclosed to the assessors or staff at any of the sites, though restricted randomisation was employed to guarantee equal numbers of sizes and location (urban/rural).

Outcome measures were selected at all four Kirkpatrick levels, including Level 1: participants' PPEs, Level 2a: increased goal orientation, Levels 2b and 3a: increased leadership knowledge and skills measured by the Multi-factor Leadership Questionnaire (MLQ), Level 4a: staff job satisfaction, perceived access to technology, equipment, training, and career progression opportunities, stress levels, intention to stay or leave, turnover, and staff absenteeism based on two validated instruments (the Work Environment Scale (WES-R) and the Workforce Dynamics Questionnaire (WDQ)), and Level 4b: staff perceptions of care quality, measured by a validated clinical quality indicator called the P-CAT, as well as statistical data on the number of unplanned hospital admissions, falls with injury, and new urinary tract infections over a two year period. There were also two outcomes that were a combination of Levels 4a and 4b, which were changes to practices and procedures and sustainability of change.

Finally, unlike any other of the combined included studies, Jeon et al. included a set of economic outcomes, culminating in a comparison between the cost of delivering the programme measured against costs saved by reduced absenteeism and turnover.

Data collection involved questionnaires at baseline, immediately after the programme, and nine and 18 months following the programme using a combination of Likert scale and open-ended questions. Although no external raters were involved, the article cited helpful statistics to reinforce participants' reports. The data analysis intended to investigate the pre and post questionnaire responses for both the experiment and the control groups, as well as the aforementioned economic outcome assessment.[5]

The cluster randomised controlled trial methodology is an excellent choice for leadership development research for two reasons. First, as healthcare is most often delivered in teams, analysing at the team or cluster level is appropriate and moves the outcomes automatically beyond the individual-level to the organisational and benefit to patients/clients. Second, the cluster approach can identify control groups in a balanced and representative way without inconveniencing the time or resources of the control group, beyond sharing pre-existing or routinely collected data. Another strength of this study is using human resource data, such as measures of workplace satisfaction and others that are often collected by organisations at regular intervals but have rarely been included as metrics in leadership development research. As explained previously, leadership is said to have the potential to influence these factors, as well as improved organisational outcomes and fewer errors at work, among others. Jeon et al. demonstrated the use of clinical outcomes that are measured by hospitals mandatorily as programme evaluation metrics, an approach that no other included studies used. Not only would improvements in these types of outcomes provide valuable information regarding optimal leadership development, but they could also contribute immediately to demonstrating the ROI of programmes. Use of the design elements demonstrated by Jeon et al. along with other principles mentioned above, could provide much more of the kind of research needed in this field.

7.3 Implications for Further Research

In addition to the suggestions regarding the theory and mechanics of conducting quality research mentioned above, there are many areas yet to be investigated sufficiently in the field.

These include:

- **Testing** whether the **conclusions** from the good and moderate evidence HEE studies are reproducible in other contexts to see if, how, and in what ways the points are generalisable

[5] This is the only study to be included that had not yet published the results of their data collection. It was included nevertheless because of its unique and effective design

- **Testing** the prototype **theoretical model** of leadership development design and measurement described in the previous chapter in terms of its usefulness and applicability to different contexts
- **Conducting much-needed experiments** that can establish causation in terms of:
  - In what ways and to what extent leadership development is effective or not by comparing participants to a control group or relevant statistics or clusters in a longitudinal study, and
  - By comparing programmes and isolating one key variable, determining which elements and combinations of elements of leadership development are most effective (Edmonstone, 2013; McCauley, 2008; Steinert et al., 2012)
- The latter point can be used to analyse the **optimal effectiveness** of developmental activities and combinations of them (McCauley, 2008), as well as programme components such as content, location, size, structure, length, location (in-house versus external), faculty, and participant characteristics (Husebø & Akerjordet, 2016). It would also be useful to investigate how and to what extent these factors differ, if at all, in different domains and at different stages of a person's career (McCauley, 2008). As mentioned previously, questions concerning profession-specific (such as physician-only) versus interdisciplinary programmes are also worthy of further investigation. Once optimal manifestations of these components are identified, one could analyse how to best implement them in various contexts
- Whether through linear regression analyses or alternative means, analysing the **relationships** among variables credibly and transparently, since leadership is a complex phenomenon
- Incorporating **organisational, economic, and clinical/benefit to client measurements**, such as: workplace satisfaction reports, retention of staff, staff absenteeism, meeting or exceeding organisational goals, economic benefits, such as the money saved by decreased absenteeism, clinical and client outcomes, and policy change. As suggested earlier, much data of this nature is already collected by organisations, but it is seldom incorporated into the evaluation component of leadership development
- Continuing to identify tangible and effective outcome metrics to evaluate leadership, particularly with examples at the **Kirkpatrick 4a and 4b level** (Steinert et al., 2012). This is one of if not the central priority for leadership development, certainly in terms of determining the ROI

- Similarly, devising effective ways of measuring **less tangible** leadership skills, such as cognitive abilities including broadening one's vision, softer skills such as communication, and more personal outcomes such as increased confidence (Watkins et al., 2011) and self-efficacy, the latter of which is tied to better performance
- Investigating the philosophical and practical implications of different approaches to andragogy (adult education), including distinguishing the terms "train," "teach," "coach," "educate," "develop," and "enable" in terms of contributing to the attainment of different outcomes
- Pinpointing optimal sample **sizes** of interventions for different purposes
- Determining the optimal requirements for intervention **facilitators**, in addition to appearing knowledgeable and credible to participants, as mentioned previously
- Performing a deeper analysis of which aspects of **organisational culture** create conditions that facilitate or inhibit the transfer of leadership application before, during, and following interventions and identifying best-practice examples of how this can be achieved successfully (McCauley, 2008)
- Conducting research in **other professions and domains**, as well as in **other parts of the world** (eg Africa, Asia, South America) to examine to what extent principles of optimal leadership development and measurement apply across contexts and cultures (McCauley, 2008; Steinert et al., 2012) and highlighting significant differences
- Assessing what **informal** opportunities can impact leadership development and to what extent, whether as substitutes for or complements to formal leadership development interventions
- Examining how leadership development should differ for those in **formal roles**, such as CEOs, versus those who exercise leadership without formal positions
- In addition to critiquing and consolidating information that is already available, as has been done in this study, it would be interesting to explore **totally innovative** forms of leadership development
- Investigating how **technology** can be used effectively in leadership development (McCauley, 2008), through approaches such as online learning and 3D simulators
- Examining leadership development on a deeper **psychological** level to investigate why certain principles are optimal and what they reflect about the nature of humans. For example, does the good evidence surrounding videotaped simulations and peer and expert feedback suggest that humans typically lack self-awareness?

- Undertaking an analysis of **whether MERSQI needs to be revised to improve it** for critiquing leadership development literature. Suggestions to this effect will follow further on in this chapter
- **Reviewing MERSQI score category groupings (strong, good, moderate etc)** used in this study to decide whether they should be revised
- **Analysing how MERSQI** in its current iteration or a revised version **can be adapted** for other domains, especially including an adaptation of Level 4b outcomes for non-healthcare domains. Given that there is an equally strong need for a **tiered approach to the evidence using a validated instrument in other domains**, the urgency of this step is high.

Husebø and Akerjordet (2016) suggest that the risk of selection, performance, and detection bias threaten causal inferences and validity in the design of leadership studies (‘Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0’, 2011). Facing such challenges, assuming one does not give up on pursuing credible empirical evidence altogether, an interesting question relevant to this study is: what qualifies as evidence in leadership development research? Although experiments are generally considered supremely credible because of their scientific nature, some outcomes are difficult to prove using an experiment, especially when analysing a complex phenomenon like leadership (Steinert et al., 2012). It is also challenging to use experiments to assess outcomes such as the less tangible skills described previously (Watkins et al., 2011). Self-reports on their own seem insufficient to demonstrate some leadership outcomes convincingly, such as organisational benefits; however, for others, such as self-efficacy, which is tied to better leadership performance (Fernandez et al., 2016; McCormick, 2001), self-reports are quite valid. Likewise, quantifying colleagues’ softer leadership skills is thought to be challenging (K. E. Watkins et al., 2011), which may cause some to question the appropriateness of doing so. These challenges however should not preclude efforts to measure any aspects of leadership development formally and in an empirical way.

Straus et al. (2013) assert in the implications for practice in their review that better quality studies will allow sharing of best practice, but best practice on its own is insufficient for the kind of evidence that is needed. As this thesis has demonstrated, the field of leadership development is not limited to case studies and best practice, which have value in their own right. As suggested earlier, comparing to control groups and isolating variables are strategies that can sharpen the results, enhance the generalisability, and provide insight into the extent to which and in what circumstances individual aspects of leadership development are more effective than others. Furthermore, while credible scientific evidence is needed, the second conclusion explored highlights the importance of some less scientific wisdom and demonstrates that ignoring those insights can significantly jeopardise the success of programmes. It is clear that the social scientific nature of the field means that the best available evidence in leadership development needs to reflect high quality research that includes both quantitative and qualitative data.

Summary of Key Research Components

Given the methodological weaknesses unearthed in the current literature, the need for better research in the field, and the host of methodologies at researchers' disposal, each of which has its own strengths and limitations, one point bears repeating. Regardless of which design or methodology one chooses, there are certain key elements of research design that can enhance studies' credibility and usefulness, which should be included as often as possible.

They are:

- collecting objective outcome data,
- using a control group,
- adding qualitative data that explores the nuances and relationships among variables,
- including open-ended questions to allow for unexpected insights or critiques to surface,
- involving multiple iterations and sites,
- incorporating pre, post, and post-post measures, and
- adding Kirkpatrick 3b, 4a, and 4b outcome metrics.

Embedding these characteristics into research design is equally relevant for researchers as for editors in terms of requiring higher calibre work, or at least of making the quality of articles more transparent. It can also be useful for those designing and delivering programmes to build more robust forms of evaluation into their programmes for the various reasons mentioned throughout this dissertation.

7.4 Discussion

Having now outlined the conclusions from the best available evidence on leadership development for doctors, extended it in many ways, and included data from other fields, as well as concomitantly having critiqued the manner in which research is being carried out in the field, the attention now returns to the research sub-questions. The first sub-question relates to the current state of the literature. Given the importance of leadership in organisations and the prevalence of programmes outlined in the introduction, the main initial question at this point concerns why the evidence base in the field is so thin when many aspects of what is needed to improve its quality seem quite clear. What follows is a summary of the responses to this question, but it should be noted that many of the discussion points are embedded throughout this thesis in the findings, conclusions, and conclusions explored. That is to say that what follows is a summary, not an exhaustive list.

7.4.1 Why Is the Evidence Base So Thin?

Insufficient Calibre Work

There are several interrelated possible reasons for the low quality of evidence in the field. The first is that the majority of the published work is of insufficient calibre. The fact that 19 of 25 HEE included studies were of limited or anecdotal value, as was the mean MERSQI score of the included studies, makes this clear. This finding is reinforced by the number of studies that left out key information or failed to collect a post-post measurement to track the application of learning. Although some of the methodological elements that would improve the calibre of research can be challenging and time-consuming, these challenges are not insurmountable. For example, while obtaining a control group for experiments of this type can be difficult, especially at senior levels, using clusters or accessing existing, regularly collected data sets, such as the clinical outcomes mentioned previously or national averages, are feasible solutions to the control group challenge. Including multiple raters and collecting data at different points requires additional time, as does processing and analysing more data; however, these measures have been explained as adding significant value to the quality of research findings. Again, what is most surprising is how few studies attempted to collect even self-reports on benefits to the organisation or benefit to patients/clients, never mind more robust, objective data at these levels. This omission is most surprising given that outcomes at these levels arguably best represent the ultimate purpose of leadership development.

Furthermore, the number of studies that did not evaluate the intervention at all to attempt to investigate what about it works, along with the relevant nuances, is unacceptable given the many benefits of doing so. Similarly, whether or not linear regressions are optimal for leadership development reviews, systematic exploration of the relationship among variables such as programme goals, content, faculty, samples, developmental activities, outcomes, and measurements is necessary; and yet, many studies failed to explore these altogether. It is possible that useful information regarding optimal leadership development is available in sources other than peer-reviewed journals, but given that this medium is meant to contain the most credible evidence, the possible excuses mentioned before for their quality are unacceptable. Similarly, the previous acknowledgement of the real space limitations for publications does not fully account for the reporting issues identified earlier, nor for other gaps in study designs and analysis, an assessment which is supported by the PRISMA guidelines. Therefore, despite the typical challenges associated with producing higher calibre research, it would be beneficial if researchers considered the aspects of credible studies listed in the implications for research worthwhile and prioritised them.

Lack of Programme Evaluation

The second reason for the thin evidence base is that many organisations are simply not evaluating their programmes at all. There are many speculations about why this is, as suggested in the introduction. To review, formal evaluation may appear to be overly time-consuming or costly, when coupled with a common conjecture of complacency that things appear to be going well or that the current design is the way the organisation has always done it. Prevailing fatigue or frustration with evaluation is another common factor, especially if past evaluations have been lengthy and the benefits were not clearly demonstrated. Alternatively, organisations or providers may fear negative feedback and elect not to pursue formal assessments, especially when considerable money has been invested in the programme or it has received a lot of corporate or internal media attention. Some may fear that critical appraisals or failing to meet expectations may result in funding cuts, personal embarrassment, or adverse professional ramifications. It is also possible that a real, clandestine purpose of some programmes may be the latent benefits of rewarding high performers or serving as a form of corporate branding. An attitude in this scenario could be that star performers seem to be progressing naturally, so the main purpose of the intervention could be the subliminal or approbatory benefits, rather than an expectation that it will result in any demonstrable change. Finally, it has been suggested that some organisations do not evaluate their programmes because leadership development is considered a source of competitive advantage and there is a concern that this edge could be diminished if details of the intervention were published. Challenges to that logic are that it would not be as applicable to not-for-profit institutions and that many organisations are not evaluating their programmes even for their own sake without sharing the results. Powerful counterarguments to this are the collection of benefits afforded by measurement and outcomes-based development programmes in terms of improved quality and the advantages for the participants elucidated in the previous chapters.

The Role of Editors and a Proposed Solution

A third contributing factor to the poor evidence base is that editors are allowing articles to be published without requiring them to include key details, as mentioned before, and without maintaining a certain calibre. It merits restatement that in order to advance the field, future research needs to be **clear and transparent** about the credibility reinforcing authors' citations of previous studies, as well as their own findings and conclusions (Steinert et al., 2012). As has been demonstrated throughout this thesis, this is currently far from the norm. One potential solution is for editors to require authors to include a breakdown of the MERSQI (or revised version of it) score for their own studies with journal submissions. Peer reviewers and editors could verify this assessment and include those results in the publication. This would not only make the credibility of the conclusions immediately clear for readers, but it could potentially encourage those publishing to do better research, knowing how transparent the quality score and composite ratings would be. For example, if a study experienced an initial questionnaire response rate of 60 per cent, the researchers may make an extra effort to get to 75 per cent in order to increase their MERSQI score. It might also inspire researchers, who might otherwise have not, to include a Level 4a or 4b outcome metric to receive the highest score for this category. Standardising publications in this way would also provide greater consistency and make it easier for readers to quickly compare studies' credibility. This approach could have an identical effect to the PRISMA statement authors' assertion that if their guidelines were endorsed and adhered to in journals, evidence of improved calibre systematic reviews would ensue (Liberati et al., 2009). Stevens et al. (2014) reiterate that editors are chiefly responsible for ensuring published articles are clear, complete, transparent, and as free as possible from bias (World Medical Association, 2013). Given the manifest need to raise the standard of publications, journal editors should be at the forefront, enforcing such a progression.

Lack of a Common Assessment Metric and Outcome Metrics

The lack of a universal common metric to evaluate the impact of leadership development programmes, as well as individual outcome metrics, is also a challenge. The former issue can lead to inconsistency and unclarity in the calibre of research, as demonstrated throughout this thesis. This issue allows details of the research design, analysis, and results, as well as the intervention, its faculty, and participants themselves, to be overlooked or omitted. This lack of consistency in measurement makes it difficult to judge the value of the findings, detracting from the generalisability, and can also have consequences in terms of perpetuating unsubstantiated findings. This can also potentially lead to inappropriate allocations of funds and resources. Finally, the lack of a common metric also limits comparisons among studies by practitioners and researchers alike. This is not to discount the value of individual goals, which has been described, despite their heterogeneity, which can limit comparisons across the sample. Likewise, as outlined in chapter six, objective outcome data, particularly at the organisational and benefit to clients levels, can be highly credible, especially when compared to a control group, clusters, or national averages. Thus, there are many ways to improve the empirical knowledge base, and a common metric (or set of metrics) should be considered among them.

To review, the first step of the solution proposed here is to use a standardised instrument for evaluating the studies' calibre (by assessing its component parts), rather than relying exclusively on a single, generic outcome metric. As described in the chapter two, MERSQI is considered the most appropriate instrument for leadership development research and is thought to have the potential to be used more widely as a standardised tool. It likely needs to be revised and validated again for the purposes of assimilating the evaluation of leadership development research. The updated form of MERSQI should add numerical weight to prioritise the most important aspects of study design and make the criteria as objective as possible to minimise researcher bias. As mentioned previously, it is important to consider whether benefit to clients is the ultimate outcome for leadership development in other professional domains, comparable to benefit to patients in healthcare, or if another one would suit better. The absence of an updated instrument does not excuse researchers from using an instrument of this nature altogether, nor from failing to include a measure of study calibre in their published work so that readers can judge for themselves what credibility to attach to the findings. Therefore, a standardised instrument to judge the credibility of studies could contribute to increased calibre, consistency, and generalisability of research.

A second step towards greater consistency in leadership development research could involve utilising common intervention outcome metrics, particularly at the 3b, 4a, and 4b levels. When programmes have miscellaneous outcome metrics, cross-context comparisons and the ensuing generalisability of findings are limited. Suggestions for such metrics include examples offered in the feature study of Level 4a outcomes such as absenteeism, retention, or workplace satisfaction. There are practical advantages to incorporating these outcomes as they are already routinely collected by organisations and therefore require no additional time and are immediately comparable to other sites or national averages. Although there are expected differences among and within domains, using common outcome metrics in different contexts serves two functions. The first is that it addresses the question of the extent to which the nature of leadership development is generic or contextual. Second, given that there is evidence that interdisciplinary programmes can be effective, suggesting that leadership is to some extent generic, using common metrics in different contexts can broaden the evidence base by combining learning from variant environments.

Despite the aforementioned challenges, for the sake of all stakeholders, including funders, the organisations involved, participants, and all the employees and patients/clients who stand to benefit from better leadership, it is essential that researchers and providers alike generate and publish the best evidence possible to ensure that leadership development is optimal.

7.4.2 The Next Questions

The central question driving this study is what is known about the effectiveness of leadership development and how to make it optimal? It is no longer necessary to ask *if* it works; there is now enough evidence that it works on some levels. It is also important to stop lamenting that very little is known and implying that nothing valuable is known at all, as Kellerman alludes. Finally, although there were accounts of programmes that failed to meet expectations, Pfeffer's controversial assertion that leadership development is having a largely negative impact is not supported by this study. This review has outlined very carefully what is known, along with the evidence that reinforces it from within the academic literature. It has also addressed the sub-questions that arose within this larger question and has generated a set of further considerations in the implications for research. The focus now turns to more specific, overarching questions, including: in what ways does development work? (Kwamie et al., 2014) That is, which outcomes and benefits can be effectively achieved, with what consistency, for and by whom, and in what circumstances? What works best in terms of combinations of programme components and developmental activities and in what situations? The latter refers to the extent to which principles of optimal leadership development apply directly at different stages of one's careers, and in different organisations, organisational situations, such as a merger or start-up, professional domains, and countries. If there are significant differences among these variables, how should programme designs change to accommodate them? How, when, and for whom are interventions more effective than informal versions of development, such as stretch assignments, mentoring, or than hiring proven performers from outside the organisation? Which forms of measuring leadership effectively described in this study are best?

As mentioned in the implications for research, many of the answers to these questions would benefit by being reinforced by experiments, comparisons to control groups, and objective data.

7.4.3 The Next Step: How Is Leadership Development Made Optimal? Specifics and Implementation

As more evidence is collected regarding what works effectively in leadership development and in what circumstances, a need arises concerning information regarding effective implementation and logistics. This relates to optimal ways of incorporating individual activities, designing interventions, and assembling development strategies over the course of employees' careers. The answers depend on a variety of factors and potentially involve adaptations and nuances based on differing contexts and situations. The Rose Report pinpointed the urgent need for better leadership and development and this thesis has clarified the evidence of principles of optimal design, delivery, and evaluation. The next step is concomitant with addressing the outstanding research questions mentioned above and involves identifying best-practice examples of proven principles, as well as untested examples of alleged best-practices that can be subject to empirical verification. While information of this kind could help optimise programme, individual, team, and organisational outcomes, one must also take into account the significant effect that organisational structures and culture can have on learning transfer.

For example, Rose (2015) mentions that to truly maximise the fruits of leadership development, there needs to be a common vision and ethos across the healthcare system to which all branches of the system align, creatively and intelligently. This echoes the first point of the second conclusion explored, which describes clarifying organisational leadership doctrine. Without this clarity, Rose argues, the fruits of leadership development will likely never fully be realised. This is not to say that all leadership development should stop in every organisation until such a common vision and ethos are in place; rather that individual interventions cannot serve as panaceas for larger systemic dysfunction. Furthermore, leadership development can potentially equip leaders in the organisation with the skills and support required to *create* a shared vision and put structures in place to resolve this kind of institutional gap. Therefore, the knowledge base regarding leadership development can be extended by investigating the intricacies of implementation of optimal principles across contexts, as well as gathering best-practice examples, testing such cases empirically, and scrutinising how organisational culture can nurture leadership development.

The series of questions that follow in the discussion below are practical versions of those in the implications for research. In order to avoid redundancy, they are worded as if the answers are simple, but these answers are rarely simple. Each of the points to follow should be considered in terms of the additional question:

What are the strengths and limitations of each of the following variables and in what circumstances is each preferable or optimal compared to the other alternative(s)?

- Along with the points made in the second conclusion explored, which specific strategies can be incorporated to create an ideal **organisational culture** to facilitate optimal application of leadership?
- How specifically can the **principles of adult learning** be included optimally in the design, delivery, and evaluation of programmes?
- In terms of **selection**, should leadership programmes be offered to all employees, only those in specific roles, or those considered high potentials or performers?
- What are the best ways to **motivate** employees to want to develop and to maximise the impact of programmes?
- Should interventions be offered before promotions as preparation for certain roles and as an incentive, or after promotions as preparation for the new role and as a reward?
- When is it beneficial to have programmes that are highly specific to the participants in terms of their stage of career or level of seniority, profession (eg physicians), institution, professional domain, versus a mixture? In which situations are certain combinations optimal?
- Are there predictable **stages of people's careers** that are pivotal times to undertake leadership development programmes? If so, what are they and what type of programmes are most effective at each stage?
- Which **outcomes** at the various Kirkpatrick levels can be effectively achieved through leadership development and in which circumstances? (This question implies that these examples are backed by good evidence). It would be helpful to have a broader list of examples for each level.
- In terms of **location**, when are in-house, external, or mixed programmes most effective? The same question applies to outdoor education and destination training experiences, which were not mentioned specifically in any of the included studies.
- What is the optimal **length** of programmes? For which participants and outcomes?
- In terms of the **faculty**, when is it preferable to have internal, external, or combined personnel?
- What are the key qualities of effective facilitators? What training do they need? What qualifications or experience should they have?

- What more can be said about the chief functions of **developmental activities** and which outcomes are they best suited to enable? Which combinations of these activities are optimal and in what circumstances?
- Similarly, which **topics, content, or skills** are most essential in different circumstances, such as at different career stages?
- What are the best ways to implement the principles of **measuring** the impact of leadership development optimally that are mentioned in this study?
- What other strategies of measuring leadership are most effective?
- To what extent are principles of optimal leadership development universal versus contextual?
- How can **informal** forms of leadership development complement or perhaps substitute for formal interventions?
- Which **totally innovative** forms of leadership development are found to be effective?

It would also be interesting to explore *why* formal leadership development is effective compared to informal forms, including from a psychological perspective.

Therefore, many aspects of the what, how, for whom, when, where, and why leadership development are optimal still need to be explored and reinforced with high quality evidence, as well as practical examples of successful implementation.

Finally, both in terms of improving the quality of research and offering the most effective programmes, in addition to what is optimal is the question of what is *feasible*. This refers to the balance between what is optimal and the amount of time, money, and resources that researchers and organisations are willing and able to devote to the enterprises of research, delivery, evaluation, and refining programmes. Better quality research and evaluation can lead to more optimal programmes; more optimal programmes can lead to better leadership; and evidence shows that better leadership is linked to a variety of positive outcomes at the individual, team, organisational, and most importantly, clinical/benefit to clients levels. Hopefully this encourages researchers and providers to contribute to improving the quality and breadth of the evidence base for such a valuable venture. A further recommendation is investigating the economic benefits of leadership development, which could add further insight into the impact of programmes and provide convincing ROI support for the endeavour.

7.4.4 Ways of Improving MERSQI

Although MERSQI was selected as the most appropriate instrument for this study, the analysis has unveiled aspects of this tool that are worthy of reconsideration and possibly modification. In this vein, there are two main comments to make. The first is that there were several items that are not included in the instrument that could be added to make it more valuable for evaluating the credibility of leadership development studies. They are structured sequentially below according to the MERSQI order.

Modifications Part One: Potential Additions

The first point that would be useful to add concerns whether studies' **reporting** is complete in terms of the study design, sample, faculty, programme, data collection instruments, process of data collection, measurements, analysis, and the connection between the previous items and the studies' conclusions. The description of the criterion could be whether the reporting was complete or incomplete and perhaps be weighted at two for complete, one for partial, and zero for not at all complete.

Study design: there is no mention of post-post measurements, which have been described throughout as an essential component of leadership development evaluation, particularly when they involve comparisons to a baseline measure and a control group. Consideration should be given to including a minimum of six or nine months given the reports of the advantages of this timeframe.

Sampling: in addition to points for more than one institution, multiple iterations could be added for increased generalisability.

Type of data: although objective data is already included, one component that could be added is combining both quantitative and qualitative data. The scoring could perhaps be 2.5 points for objective data only and three for mixed data. Another consideration is whether there should be different tiers of objective data. For example, one supervisor's testimonial of behaviour change would be deemed objective data, as would statistical performance data contrasted to national averages, but these two are clearly of an unequal weight. A more general wording is whether external raters' appraisals should be on par with statistics or factual outcomes, such as policy change.

Another item not addressed in the instrument section that could be added is the completeness of data reporting in terms of the representative nature of qualitative data, as well as outlying opinions.

Validity of evaluation instruments' scores: though there is significant value in validated instruments for the sake of generalisability and cross-study comparisons, the benefits of personalised data collection for individual programmes and participants has been elucidated. It would be interesting to discuss how the latter can be included in MERSQI, since not doing so may implicitly discourage researchers from including them.

Outcomes: it would be beneficial if organisational outcomes (Level 4a) were added and given high prominence. It has already been mentioned that an ultimate outcome is needed for non-medical domains, whether this is benefit to clients or another. In its present form, there is no distinction between self-reports and objective Level 4a and 4b outcomes, which could be remedied even though there is a separate category of subjective versus objective data. The reason for this is that a study that collected objective data on individual-level outcomes and self-reports of organisational-level outcomes would receive the same outcome score as a study that provided objective evidence of organisational-level improvements. Finally, because it has become evident that **economic benefits** can effectively demonstrate the ROI of leadership development, another consideration is whether this category outcome should be added.

Modifications Part Two: Subjective Elements

The second set of observations about MERSQI is that there are items that lend themselves to subjective researcher assessments, which would be more credible and would minimise bias more if they were more objective. The following points concern whether objective criteria could be identified for:

Study design: characteristics of **control groups**, such as those that are balanced and representative.

Sampling: whether the sample is representative or not, which is one of the criteria in the Rowan and Huston (1997) analytic approach. This also relies on determining what qualifies as a representative sample.

Validity of evaluation instruments' scores: the internal structure of the instruments.

Data analysis: the appropriateness of data analysis. This is an issue because the analysis of self-reported-only data could potentially be appropriate given the data set, but the data set may not be appropriate to satisfy the study purposes because of the limitations of relying on self-reports exclusively. Second, the sophistication of data analysis also relates to whether the relationship among variables is analysed and whether this is done simply by descriptive analysis or in a more comprehensive way.

As an example of how the existing version of MERSQI can fall short, there could be very strong evidence that something minimal has been achieved (eg a simple task). This could be demonstrated using a small sample by way of an RCT with 100 per cent response rate and self-reports of Level 4b outcomes (not verified by objective data), which would receive a near-perfect MERSQI score. By contrast, another study that targeted broader skills and systemic organisational change with a large sample could have a lower tally. For example, the highest-rated study in the Rosenmen et al. (2014) review was a half an hour-long intervention related to CPR training. The authors tested performance according to the frequency of "leadership utterances," which is quite a weak metric. This goes to show that even with a reliable evaluative instrument, attention needs to be paid to the details of the study.

7.4.5 Review of the Research Sub-Questions

Finally, this section reviews how the information presented in the findings, conclusions from the best available evidence, and conclusions explored relate to the research sub-questions, which are presented sequentially.

1) What is the **current state** of the leadership development literature regarding available information relating to professionals what is its calibre?

The current state of the literature is that the evidence base is thin and that the overall calibre of its conclusions is low and muddled by unclear reporting. Before commenting on the calibre of studies, a summary of the findings, including mention of those from the two good and four moderate evidence HEE studies, is presented below to illustrate the work being done in the field.

Summary of Findings

In terms of **designs**, the majority of 72 included studies are case studies (between 74% and 79%) with a small group of quasi-experiments (6%) and experiments (6%). Only one randomised control study (RCT) appeared in this review, that by Jeon et al. (2013). This dearth of confirmatory designs, with only 14 per cent of included studies testing hypotheses, limits the strength of causal relationships that studies can claim. Although the extant reviews included many more RCTs, they tended to be very short, team-based interventions that were task-centric and directive. Of the two good evidence studies, one was an experiment and one was a quasi-experiment. Only 51 per cent of the included studies included both quantitative and qualitative data, and although both are needed to advance the field, both of the good and two of the moderate evidence studies collected *only* quantitative data. This restricts helpful nuances regarding in what ways, for whom, and in what circumstances leadership development is effective. Nearly 80 per cent of studies included questionnaires, with almost a third (n = 21) relying exclusively on that form of data collection. The only experiment was featured in a good evidence study and the only two studies to include statistical analysis were moderate evidence ones.

In terms of **samples**, nearly 60 per cent of the sample participants were men; however, in the HEE review, only 34 per cent were male. Nearly 80 per cent of the studies were single-profession (MULTI) or physician-only (HEE) samples, compared to interdisciplinary ones, though 40 per cent of the HEE studies were interdisciplinary. This is similar to the 64 per cent of the interventions in the Steinert et al. (2012) review that were physician-only. One of the good and all four moderate evidence HEE studies were physician-only. Nearly a third of studies were for senior or executive-level leaders, with mid-level and junior leaders being featured in 15 and 12 studies respectively. This is a stark contrast to the Frich et al. (2014) review, in which not one of the 45 included studies featured senior-level participants. Only two studies were mixed levels of seniority, though 32 studies left this category unspecified. Among the best calibre studies, the distribution of sample levels of seniority is two senior, two junior, one mid-level, and one unspecified. MULTI reveals that the most common professional domains featured in studies are healthcare (36%) and business (30%). The selection criteria for nearly half of the studies' samples is unclear; however, the most common were those who were nominated or volunteered (17% for each). In only two studies were participants required to attend, though one of these is a good evidence study. The other good and two moderate evidence studies' samples volunteered for the programme and participants in two moderate calibre studies applied and were selected.

In terms of the **interventions** themselves, most of the work is being done in Western countries (88%), with only three studies being from Africa and one from Asia. 51 per cent were in-house programmes and the length was most often four to six months (14%) and eight to 11.5 months (14%). Interestingly, in HEE, there were three one to two-day interventions, compared to none in MULTI, and 40 per cent of HEE interventions were a year or longer, compared to nine per cent in MULTI. More than half of the latter were healthcare leadership programmes. One of the good evidence studies featured a one to two-day intervention, with the other being unclear about the length, while three of the four moderate evidence studies were a year-long or longer. The included studies' interventions most often featured workshops (44%), 360s (39%), coaching (38%), lectures (33%), and action learning (32%). This challenges claims that the majority of physician leadership programmes are lecture-based (Frich et al., 2014; Rosenman et al., 2014) and indicates that experiential methods are becoming more of a focus (Steinert et al., 2012; Suutari & Viitala, 2008; K. E. Watkins et al., 2011). Only one of the best calibre studies included lectures, though three utilised simulations or role plays with facilitator feedback.

In terms of **measurements**, the most common, as categorised by the Kirkpatrick model, were increased knowledge and skills (71%), changes in attitude or perception (57%), and subjective behaviour changes (53%). All six of the best calibre studies included a Level 3b objective behaviour change outcome, though only 35 per cent of the total sample did. The number of included studies that measured at a Level 4b (benefit to patients/clients) increased by 26 per cent from 2005 – 2012 to 2013 – 2016, which could indicate that it is beginning to become considered more important. Surprisingly, they were not even considered in the Straus et al. (2013) and Steinert et al. (2012) reviews. More than half of the studies relied on single ratings, 44 per cent of which were self-ratings. Four of the six best calibre studies included multiple raters. 48 studies (67%) were limited to subjective data only, though both good and two moderate evidence HEE studies included objective data. In terms of the focus of the evaluation, 31 per cent concerned participant outcomes and benefits, 15 per cent assessed only the programme, and just over half (56%) did both. The most common time to measure was post-post (49%), followed by post (44%) and baseline (42%). Five of the six best calibre studies collected data at baseline and post or post-post.

In terms of **reported outcomes and benefits**, the most common were subjective increased behaviours (36%), knowledge (35%), and skills (33%). Increased self-awareness, confidence, and increased leadership capability/competence/capacity/effectiveness/self-efficacy were also mentioned frequently.

Having summarised the details of the literature's raw data, the focus now shifts to the assessment of the current state of the literature.

The Current State of the Literature

One key goal of this study was to assess the calibre of the literature overall. Though there are many claims that the quality of research is low, and some, like Kellerman, imply that there is hardly anything redeeming about its yield, an in-depth study was required to assess it systematically and transparently. This was undertaken by applying MERSQI, a validated instrument, to the HEE included studies. The fact that 19 of 25 studies qualified as having limited or anecdotal evidence and that the overall mean was 9.94, which situates it in the anecdotal category, reinforces the thinness of the evidence base. This is further echoed by similar findings in the extant reviews, such as Rosenman et al. (2014), who reported a MERSQI mean score of 11.4 for their 52 included studies. Likewise, the current state of the literature reviews in the field is also sub-optimal, with many common issues that detract from the usefulness of their findings and conclusions. This has been discussed at length in chapter two and in the implications for research.

The thinness of the knowledge base is broadly attributable to four factors, the first of which is insufficient quality research. This is partly due to poor reporting, which is a pervasive common error. The fact that 38 per cent of the included studies are unclear about how long the intervention is one example of many that reflects the flaws in authors' reporting, which detract from the usefulness and credibility of their studies. Compounded with poor reporting is a host of elided key methodological aspects. For example, in terms of the data collected, a third of the included studies collected only qualitative data and only roughly the same number relied exclusively on subjective descriptions of outcomes, which are quite weak measures of outcomes. This substantiates assertions that much of the data in leadership development research relies on self-reported data (Blumenthal et al., 2014; Malling et al., 2009; Straus et al., 2013). Surprisingly, only 33 per cent of included studies gathered objective data and only two used statistics to reinforce their claims, which echoes Frich et al.'s (2014) assessment regarding the paucity of quantitative outcomes overall. Although both types of data are needed for complete analysis, only 51 per cent of studies collected qualitative and quantitative data and more than a third restricted themselves to single data collection measures, precluding triangulation. Only ten per cent featured a control group, despite its usefulness in reinforcing comparative improvement. Similarly, only 14 per cent featured multiple iterations of sample populations. Only 25 per cent reported outcomes at the organisational level (4a) and even fewer, 11 per cent, did so at the clinical/benefit to clients level (4b). As has been mentioned, these levels are essential for maximising the impact of leadership development. More than half neglected to include a post-post measure, despite the necessity of tracking long-term application of leadership to the workplace, as well as whether short-term success is sustained over time. Similarly, less than ten per cent collected baseline, post, and post-post data, which has been explained as being the ideal combination for various reasons. Finally, only one HEE study collected data during the programme, which means that the others missed an opportunity to add the many benefits of doing so. As suggested earlier, many of these shortcomings could be easily remedied with little or no additional time or resources.

The second cause of the poor evidence base is that the editors are approving publications that lack important components, such as those mentioned above, and are unclear about the substantiation behind studies' findings and conclusions. A third reason is that many programmes are evaluated in an unsatisfactory manner and many are not evaluated at all, the latter of which Avolio (2005) estimates is more than 90 per cent. This limits their benefit to providers and the research community alike. It is argued that this is sub-optimal for participants as well, given the benefits for them of measuring, which have already been described. Lastly, the lack of a universal metric to evaluate the success of programmes, as well as common individual outcome metrics, challenge cross-study comparisons. The combination of these factors results in a body of knowledge that is both incommensurate relative to the prevalence of programmes and generally highly ambiguous regarding what exactly is known and based on what evidence.

A related challenge is that the state of literature reviews is also quite low, as was demonstrated in chapter two. Authors frequently leave out key details of the included studies, an omission they themselves condemn in the studies that they analysed, and few used a validated instrument to assess the calibre of each study. Not one extant review attempted to investigate the relationship among variables in a systematic way, which would be useful given the complex nature of leadership development. Finally, a major challenge is that not one of the EMD reviews tiered their findings or conclusions based on the quality of the evidence, which naturally perpetuates the pervasive uncertainty regarding the strength of the evidence in the field.

The solution proposed in this study is to review and possibly revise MERSQI and then deploy it as the standard quality instrument for assessing leadership development research. Authors of individual studies could evaluate their own work according to this instrument and could embed the results in their submission, which peer reviewers and editors could then verify. The final score and its component parts could be included in the actual publication to give readers an immediate sense of the credibility reinforcing the study. Similarly, review and meta-analysis authors can publish the scores for each of the included studies, as well as the composite parts, to produce the same effect. The hope is that in addition to clarity, this level of transparency and consistency would ameliorate the calibre of work that is submitted and published.

That said, the state of the literature is not as deplorable as Kellerman and others have described. This dissertation demonstrates that there is a growing body of research that includes strong and moderate evidence, which suggests that aspects of leadership development can contribute successfully to achieving outcomes at the individual, organisational, and benefit to patients levels. There is also evidence surrounding key functions of developmental activities, such as action learning's role in forming the core of medium and long-length interventions to target Level 3b, 4a, and 4b outcomes and enhance self-awareness. An array of effective research and programme design components also emerged, such as effective approaches to evaluating leadership development programmes. Furthermore, there are several useful models that can inform the study and design of leadership development, especially when modified or adapted, as has been demonstrated in this study. These include Kirkpatrick's (2006) training evaluation model, MERSQI (Reed et al., 2007), Knowles's (1984) principles of adult learning, and Dale's (1969) Cone of Experience. This thesis has also introduced a prototype theoretical model of leadership programme design and evaluation, as well as the beginnings of a related model regarding the cardinal and interconnected functions of developmental activities. Both of these models draw on the empirical evidence uncovered by the HEE review, as well as from the deeper levels of analysis afforded by the SEA methodology. Finally, an empirically-supported set of factors that can facilitate the successful application of leadership has been described in the second conclusion explored. This collection contains findings from the best available evidence, which are reinforced by a host of examples and authorial recommendations that are intended to enhance their usefulness. Given the number of studies that purportedly failed or did not meet expectations and the potential ensuing consequences, the factors in this conclusion explored that are designed to prevent such failures carry extra weight.

The process of answering this first sub-question unveiled a set of points regarding effective ways of conducting systematic literature reviews, as are described in detail above in the implications for research. The process of designing the research protocol for the HEE review, information from the guiding tools, including PRISMA, and the analysis of the extant reviews revealed key methodological strengths that could enhance the credibility of reviews' findings. The analysis of the designs of the MULTI and HEE included studies also brought to light several elements of conducting effective research at the individual study level, which are listed above as well.

Therefore, although the current state of leadership development literature is overall quite thin, many principles, reinforced by empirical evidence, are presented in this dissertation, as are concrete elements of conducting credible research that can help improve the situation and expand the knowledge base. Thus, this has provided clarity on a much-discussed topic in the field, as well as straightforward and feasible recommendations for ameliorating the current state of literature.

2) What evidence is available regarding **optimal leadership development** for professionals? This refers to programme components, such as length, developmental activities, such as lectures, facilitators, such as internal versus external, and professional characteristics of the participants.

As mentioned previously, despite the gaps in the research, there is a growing knowledge base regarding principles of optimal leadership development. The systematic evidence analysis afforded analysis on many levels, which revealed conclusions of three main types: empirical evidence, theoretical principles, and useful anecdotal information.

The first level of conclusions derived is **empirical evidence** surrounding which outcomes can be improved through leadership development, including increased confidence, knowledge, skills, and behaviours, as well as having a positive impact on participants' career progression. Among other benefits, this is helpful given the link between self-efficacy and leadership behaviours (McCormick, 2001; McCormick et al., 2002), as well as leadership effectiveness (Seibert et al., 2017). Implementation of action learning projects is a key outcome that is linked to demonstrating an impact at the organisational and benefits to patients levels. This was reinforced by the findings of the statistical analysis, which showed statistically significant correlations among action learning, coaching, longer programmes, and outcomes at Levels 4a and 4b. The effectiveness of workshops, videotaped simulations, peer and expert feedback, coaching, and 360s is also evident in terms of increasing leadership behaviours. Despite these findings and claims that leadership development generally is moving from a more didactic to experiential focus, the aforementioned developmental activities, as well as self-reflection, are under-utilised, as reflected in the relatively low numbers of programmes overall that incorporated them. Self-awareness was a commonly-mentioned benefit (in 26% of included studies), despite its scant use as an outcome metric (Frich et al., 2014), appearing in only two HEE studies. Facilitating self-awareness fits with Parker Palmer's (1998) notion of leadership and is supported by others (Bergman et al., 2009; Rowland, 2016). It is also surprising that only two studies used 360-degree feedback pre and post as outcome measures. Equally, many programmes are still based on a lecture-style model, rather than considering the role of lectures in reference to specific learning objectives. Medical residency, interdisciplinary, physician-only, and outcomes-based designs have been found to effectively enable leadership development. Knowles's principles of adult learning have been successfully incorporated into leadership development designs, which authors of a good evidence study and others claim enhanced the results. There is reliable information available regarding the benefit of goal setting (Latham & Locke, 1983), as well as the individual nature of people's range of development, as explained by Vygotsky (1978). The latter varies person-to-person and there is a correlation between the outcomes people can achieve and their developmental challenge (DeRue et al., 2012), training motivation (Hassan, Fuwad, et al., 2010), and learning goal orientations (Suutari & Viitala, 2008). There is evidence from credible studies that leadership interventions can underperform or fail, many of which the authors attribute to aspects of the organisational culture. This gave rise to the collection of points that made up the second conclusion explored. Finally, there is moderate evidence that leadership development can be linked to organisational and clinical outcomes, particularly as a result of implementing action learning projects.

There are also **theoretical principles** for three concepts, reinforced by elements of the empirical data, examples from included studies, and authorial recommendations. These are the principles of adult learning into programme design, the key functions of developmental activities, and the model of outcomes-based design and evaluation. First, a modified and extended version of the principles of adult learning, which includes pre-programme motivation, self-direction, valuing participants' experience, relevant and practical content, an outcomes-based design, measurement, and an experiential and application focus ensure that leadership development is tailored specifically for adult learners to maximise its impact. As just one example, the MacPhail et al. (2015) study demonstrates how incorporating these principles into the design of the programme can improve outcomes and avoid failure. Second, the beginnings of the model of the key functions of the developmental activities elucidate how they can be best implemented to meet specific programme objectives and how they are optimised in concert with each other. Finally, the outcomes-based theoretical model provides a sequence of programme design, delivery, and evaluation approaches centred on the ultimate purpose of leadership development, which is application. It recommends selecting specific outcomes at various levels of the Kirkpatrick model and holding participants and ideally facilitators and supervisors accountable to achieving those outcomes. These three concepts are thought to influence meaningfully participants' experience of leadership development and most importantly, its impact on outcomes. They provide the potential to serve as guides for designing programmes for specific purposes, organisational needs, and participant populations. Finally, these three concepts also appear to be equally relevant to long-term career leadership development, as they are to individual interventions.

Finally, as has been mentioned previously, there are **anecdotal** accounts of best-practice and failed programmes that contain information that reinforces, nuances, and offers examples of implementation of the previous two categories of evidence, the empirical and the theoretical. For example, although there is moderate evidence that 360s are an effective tool for enhancing self-awareness, the anecdotal studies nuance this finding by suggesting without adequate follow-up support they can be ineffective or destructive. The set of factors outlined in the second conclusion explored that influence the transfer of leadership learning, particularly following interventions, provides insights into considerations when planning or revising programmes. Not all the information in this section derives from anecdotal studies, so it should not be restricted in this way; however, many examples from anecdotal studies provide helpful details or nuances to the factors that made up this conclusion explored. These points relate to strategies before, during, and following interventions, such as creating a receptive organisational culture, implementing a cycle of experiential learning, holding participants accountable, rewarding successes, and supporting improvement. Therefore, as has been explained already, there is useful information in anecdotal studies, which can be added to findings from better calibre studies. It is important to be clear about which is which; however, and that is a clarification that is not made consistently enough in the field.

Therefore, although more evidence and nuances are needed, there is a good deal of useful information, including those presented in this thesis, to guide the design and refining of leadership development to optimise their effectiveness.

3) What evidence is available in terms of effective ways of **measuring** leadership, particularly following interventions. This refers equally to effective approaches measurement, as well as to which post-programme outcomes are achievable.

The analysis of the included studies in the MULTI and HEE reviews revealed several key elements of effective leadership development measurement. The application of this as it pertains to researchers has been explained above in implications for research. What follows below describes recommendations for effective leadership measurement in practice by providers.

Measuring leadership is most effective when it is an integral part of the programme design, as outlined in the adapted and expanded principles of adult learning and the outcomes-based theoretical model. Leadership development outcomes are effectively categorised using the modified Kirkpatrick model (Level 1, 2a, 2b, 3a, 3b, 4a, 4b outcomes) and it is beneficial when providers are intentional and explicit about which post-programme outcomes a given intervention is addressing. It is also useful to devise outcomes according to SMART goals (specific, measurable, attainable, results-based, and time-bound). For the sake of comparison, measurement should begin at baseline and also include collection points during and following the intervention, as well as six to nine months following it. This timeline allows for the application of leadership to the workplace, as well as testing whether early successes are sustained. The forms that these measurements can take are Likert scale self-ratings of outcomes such as confidence in one's ability to lead (self-efficacy), 360-degree feedback reports, performance appraisals, self-selected goals for improved development, organisational data such as workplace satisfaction, objective outcomes such as launching a new initiative, and statistics, such as lowering the number of preventable hospital deaths. Different outcome metrics can address different goals. For example, 360-degree feedback reports pre and post can demonstrate increased self-awareness (Level 2a), knowledge and skills (Level 2b), and behaviour change (3b), but are less equipped than the kind of outcomes mentioned above to address organisational change (4a) or benefits to patients (4b). Therefore, outcomes that are SMART, at various levels of the Kirkpatrick model, selected intentionally for the goals of the programme, and administered at baseline, during the intervention, post, and six to nine months following are likely to yield the best outcomes.

Deciding on outcome metrics is most effective when participants are involved in creating or selecting them so that the metrics are relevant to their own individual and organisational situation. This process is further enhanced when participants' supervisors are held accountable for the former group's development so that they can ensure the goals are reasonable and can provide the necessary support and resources to make attaining them successful. Another way in which measurement can improve the impact of leadership development is when providers establish a clear link among the post-programme desired outcomes, programme goals and objectives, content, developmental activities, and measurements. This strategy is the basis of the theoretical model presented in the previous chapter. When this blueprint is announced to all facilitators and participants, the expectations and the role and relevance of each component is made clear. Incorporating measurements throughout the programme gives participants the chance to reflect on their own development, to extend their goals and learning if success is being had, and to benefit from the support of facilitators while it is still available if difficulties are encountered. It also enables providers to adapt aspects of the intervention to participants' preferences as it progresses to further maximise the experience. At the conclusion of a programme and following it, qualitative feedback can be solicited from facilitators and participants. This is useful for quality control purposes and to refine programmes by having participants evaluate the programme and its components in terms of quality and success in meeting their respective goals. Participants can also be invited by way of open-ended questions to volunteer outcomes and benefits that were not originally expected. While this may not be comparable in the same way as programme-wide metrics, this data can provide valuable information about unexpected or outlying benefits of the intervention. A further step could be to institute commonly-reported benefits from previous years as universally-applied outcome metrics in succeeding years, or they can be offered as suggestions to future participants when selecting their own pre-programme outcomes.

The benefits to participants and organisations of measuring leadership development effectively have been described as well. For participants, having measurable goals can focus their learning, motivate them to strive to achieve them, and function as an additional developmental tool by encouraging them to reflect on the extent to which they are progressing towards their goals or not. The latter assessment can be used formally as a catalyst for coaching sessions, or participants can approach peers or mentors for advice and support. Re-evaluating goals during an intervention can provide an opportunity to extend goals which have been met, or modify goals or ask for remedial support when challenges present themselves. There is further evidence that individuals who have specific, challenging goals perform better than those who do not. Structured reflection, which can come in the form of evaluation, has also been said to enhance programme impact by prompting participants to consider how the learning from each session can be applied to their own organisational and leadership context. For organisations, measuring leadership development outcomes can equally indicate when successes are being had or when further support is required. Measuring outcomes is also a way of demonstrating the ROI of programmes, as well as how and in which ways it can contribute to the organisational strategy. Finally, the consequences of *not* measuring the impact of leadership development is said to range from a danger of stagnation, to the perpetual use of suboptimal means, to interventions that fail to meet expectations or fail altogether. Once again, effective measuring can help improve the impact of leadership development at the individual, organisational, and benefit to clients levels, as well as demonstrating the return on such an important and sizeable investment.

Therefore, as with optimal programme design and delivery choices, more examples of effective measurement would be beneficial. That said, there are many aspects of measurement that have been shown to improve the credibility of research and the impact of programmes; and yet, as evinced in this study, many are not being implemented consistently in both the academic and the practical forums.

4) What insights can be drawn regarding the nature of leadership in terms of it being **generic versus contextual**? This refers to nuances of the extent to which leadership development transfers naturally among different countries, professions, organisations, teams, roles, and levels of seniority of participants.

As mentioned previously, no study analysed this question specifically; however, some preliminary comments can be made nonetheless.

Based on this study's sample of 72 unique studies, no conclusions can be made regarding the generic or contextual nature of leadership or development concerning nations, since nearly all the literature comes from Western countries. There was also no credible evidence identified regarding specific characteristics of leadership development for different roles, levels of seniority, or professions.

There are isolated or anecdotal claims that role and profession-specific programmes are beneficial, such as Blumenthal et al.'s (2014) endorsement for the intervention tailored specially for medical residents and Vimr and Dickens' (2013) contention that physician-only programmes are optimal; but there are also numerous counter contentions that mixed or interdisciplinary programmes are effective. It was also discovered that there is moderate evidence that within healthcare, interdisciplinary and physician-only programmes can work well. As mentioned earlier, further research is required to ascertain in what circumstances the one might be preferable to the other.

A further observation is that it seems clear that many of the key leadership capabilities, such as self-awareness, decision-making, the ability to manage resources and inspire colleagues, among others, are to a large extent generic. The *application* of them, the organisational culture including the language, protocols, and procedures, and the potential outcome metrics, to name a few, may differ, but many of the core skills and behaviours are remarkably similar. This is in line with Taylor's (2010) statement, mentioned in chapter three: that leaders and their behaviours are largely the same, but the organisational climate, structure, and cultures differ. As a nuance, it should be repeated that Goodall (2011) and colleagues demonstrate with strong evidence that senior leaders who are exceptional performers in the core business of the profession, such as being an outstanding surgeon, produce better organisational outcomes than those who are mediocre or non-technical business managers. This finding supporting an aspect of context-specific leadership has interesting implications for selection of participants to develop, as well as for the choice of interdisciplinary programmes, adding weight to a profession-specific aspect of leadership development programmes or syndicates.

Similar to the overarching point made above, it follows that many of the major conclusions from this study regarding leadership development are also largely generic. For example, simulations are used commonly in multiple domains, such as healthcare, business, sports, and the military. As described earlier, these are commonly followed by peer and expert feedback and discussion, which the U.S. army calls After Action Reviews, as one example. This position is supported by Pinnington's (2011) survey that found no difference in the perceived effectiveness of leadership development practices in private versus public/not-for-profit sectors, which seems to indicate that there is some overlap in different domains. Other key points, such as the role of 360-degree feedback in contributing to self-awareness, coaching, mentoring, the manifold benefits of action learning, the importance of respecting organisational culture, and the benefits of an outcomes-based approach do not appear to be context-specific. Likewise, the points relating to evaluating leadership development also likely transfer to many contexts. This is not to conclude definitively that leadership development is entirely generic; the advantages of adapting interventions based on the participants and their leadership situation have been described at length in the two conclusions explored. Rather, it seems that the core principles of optimal leadership development and measurement are largely universal, but can manifest themselves in numerous ways, which can be modified to suit different situations.

Second, amidst the debate of specialised versus mixed or open programmes, there is a hybrid option. The first way to accommodate this is by enabling participants to personalise their own goals according to their role and organisational needs based on the groundwork recommended in the second conclusion explored. This can be reinforced throughout the intervention by individual support by facilitators, mentors, or coaches. The second way is by ensuring that larger open activities such as lectures or workshops are followed by profession- or role-specific syndicate sessions. The advantage is that larger sessions can sometimes accommodate higher profile faculty or resources, such as a world-class keynote speaker, as well as providing a diversity of perspectives and experiences. To this can be added the benefits of specific sessions that cater to each sub-group's common language, experiences, challenges, and strategies of application. Finally, this can be further enhanced by giving participants time for structured reflection to consider how each activity or module relates to their own situation.

More work needs to be done regarding investigating the specific nuances that determine the generic versus contextual nature of optimal leadership development principles and its application. Although it appears that many of the core leadership capabilities and leadership development and measurement practices are to some extent generic, this does not indicate that all forms of interventions are equally effective in all cases. To use the example of the physician-only versus interdisciplinary healthcare professionals samples, there are advocates on both sides, which suggests that the answer is not binary (ie that one is conclusively better than the other). The work that needs to be done is to uncover the advantages and drawbacks of each to determine in what ways, for whom, and in what circumstances one might have a greater impact that the other.

7.5 Limitations and Strengths

7.5.1 Limitations

The findings and conclusions of this study should be considered along with several limitations. First, there were several choices that limited the scope of the return of the literature search, such as doing systematic instead of non-systematic reviews for MULTI and HEE, restricting the search to peer-reviewed articles without including any nonindexed or open access journals or unpublished studies, and including only articles published in English. The latter is a choice all other extant reviews made as well. Cook and West (2012) use a helpful analogy, describing a systematic review as a lighthouse shining over the ocean, illuminating a small space and leaving the rest dark. The choices listed above were made for a few reasons. First, they aligned with the overall thesis goals of isolating the best available evidence, highlighting its strengths and weaknesses (D. A. Cook & West, 2012), making the presentation clear and transparent, and enhancing the replicability of the findings. These measures, along with limiting the sample to professionals and excluding nurse and education leadership programmes, were felt to be necessary to keep the sample size manageable enough to analyse in an in-depth way and to maintain a professional commonality among sample participants. Furthermore, Cook and West (2012) suggest that in-depth studies are crucial, since the authors assert that the degree to which reviewers explore the strengths, weaknesses, heterogeneity, and gaps in the evidence determines in large part the value of the review. A second limitation is that the high level of heterogeneity among the included studies' designs, reporting, interventions, and assessments made some aspects of the analysis, such as comparing samples of various levels of seniority or comparing common outcome metrics, challenging. These, as well as the small HEE sample size, also precluded conducting a meta-analysis. Likewise, for feasibility reasons, topics and content were excluded as a variable in the statistical analysis, even though the connection between them and post-programme outcomes is an interesting consideration for further exploration.

A third limitation, as with the analysis of the extant reviews, is that the findings and the depth of analysis depended to some extent on the quality of the included studies (Liberati et al., 2009), which, as was demonstrated, was quite low overall. Once again, although it is acknowledged that some of the reporting issues identified in the included studies may be related to the space limitations authors face when publishing, the need for greater clarity and transparency merits reiterating. Fourth, most of the studies originated in Western countries, which is a common challenge in the field, thus one should not assume that this is representative of leadership development for professionals globally.

Similarly, the statistical analysis that was used to reinforce the careful descriptive and illustrative analysis is admittedly based on a small sample size, which limits the strength of the conclusions that can be derived from this effort alone. This was considered worthwhile because the interrelated nature of aspects of leadership development is a highly important concept and yet none of the extant review authors attempted to analyse these relationships systematically and many did not investigate them at all. As mentioned previously, there is support for the value of using of small sample sizes and thus, it was decided to include this measure to enhance the credibility of the analysis, with expectations of largely non-significant results. Surprisingly, there were noteworthy correlations, which might suggest trends that one can imagine with a larger sample and hopefully can inspire further investigation into the strength of their relationship. Thus, they might be useful for hypothesis generating purposes.

A sixth limitation concerns aspects of the MERSQI instrument, which have been described in detail above. Other instruments were considered for assessing study quality, including the Cochrane Assessment Tool for Nonrandomised Studies of Interventions that analyses risk of bias, but MERSQI was selected because of its appropriateness for leadership development programmes, specificity of each item, applicability to both quantitative and qualitative studies, and numerical scoring. The latter was particularly useful, given the goals of providing a transparent analysis, so readers can judge for themselves with what weight they consider the conclusions. In this vein, unlike the approach used in some of the extant reviews, no studies were excluded based on a low MERSQI score. This was additionally helpful in terms of the information included in the conclusions explored, some of which derived from limited and anecdotal studies. Likewise, excluding all limited and anecdotal calibre studies would have lessened the sample from 25 to six, which would have majorly restricted the ability to generate much of the information provided in chapter six. Maintaining the best available evidence as the core of this dissertation's conclusions allowed information and examples from the lower calibre studies, as well as from uncertain calibre studies in the case of the MULTI studies, since they were not assessed using a validated instrument, to be included usefully. These points function as elaborations, nuances, and further examples that enhance the conclusions, conclusions explored, and the theoretical model.

Another limitation is the exhaustive nature of having the findings from the three reviews (MULTI, EMD, and HEE) and results of the linear regressions presented in full; however, the PRISMA guidelines emphasise that the benefit to readers of being able to critically appraise a clear, complete, and transparent systematic review report outweighs the possible increase in length of the report (Liberati et al., 2009). To explore the replicability of the findings, this methodology could potentially be repeated for other professions.

Two final limitations and their justification are the most significant of this thesis. The first is electing to use a novel methodology, the systematic evidence analysis (SEA), instead of a typical case study, as nearly 80 per cent of the included studies did. Access to leadership development programmes had been gained by this thesis's author and the benefit of adding new empirical data in the traditional sense was considered. It was decided that given the tremendous unclarity regarding the state of the literature, adding a small amount of new data would leave the extent to which it supported, expanded upon, challenged, or contradicted existing knowledge in the field unclear. What seemed of pre-eminence was establishing in a clear and transparent manner what is known and based on what evidence, which other methodologies, such as case studies, could not accomplish. Once that had been completed in the form of the conclusions from the best available evidence, the SEA's iterative, multi-layered analytical approach enabled in-depth studies of the two crucial topics to be investigated in the conclusions explored. It also facilitated the treatment of another key topic, which formed the beginnings of a theoretical model of the cardinal and complementary features of developmental activities, which also would not have been as extensive through a case study, for example. For the same reason, the formulation of the theoretical model was made possible. Lastly, the SEA allowed an extensive examination of the way research is being done in the field and provided suggestions on how to improve it in both individual studies and literature reviews. Part of this information came by virtue of comparing aspects of the three reviews, which in a standard PhD structure with one literature review, would not have been possible. Each of these steps provides knowledge that is more comprehensive, detailed, and novel than currently exists. For example, although others have mentioned applying Knowles's (1984) principles of adult learning to leadership development, the set included in the previous chapter were extended by adding two and are more detailed than any others in the included studies. Therefore, the SEA methodology enabled the analysis to go beyond the scope of traditional methodologies to investigate several central topics in leadership development in a more extensive and detailed way that currently exists, especially that of clarifying what is known in the field and based on what evidence.

The final limitation surrounds comparing MULTI and HEE, given their methodological differences. The use of MERSQI and the linear regression analysis and the ensuing lack of tiered conclusions of MULTI precluded a direct comparison between the two and prevented assessments of the best available evidence of the MULTI. The analysis of MULTI and the extant literature review inspired the methodological choices for HEE and has two further implications. The first is that this progression demonstrates this thesis author's development as a researcher from MULTI to HEE. The second is much more significant, which is that these two systematic reviews represent a microcosmic example of the calibre of research that is needed (HEE) versus a representation of the current state of the literature (MULTI). Without tiered conclusions and clear and transparent findings, cross-review and cross-study comparisons are limited or impossible. The ensuing challenges and dangers of this confusion have been mentioned throughout. The awareness of the many limitations above sheds further light on gaps in the current state of evidence and highlights areas for improvement.

7.5.2 Strengths

There were also several strengths of this study that enhanced its validity and reliability.

The former refers to a piece of research being considered an accurate representation of the phenomenon using a plausible and credible study (Smith, 2004). This is determined by the quality of the questions asked and the data collected in terms of its detail, accuracy, and ability to answer the research questions driving the study (Denscombe, 2010). The latter refers to the quality of methods of data collection in terms of their consistency in producing similar results under certain conditions at different times, all other things being equal (Creswell, 2008).

The SEA methodology and its various composite parts include many strengths of this nature. The first is the comprehensive research protocol and the elements of doing a systematic review, along with the chosen methodology for the HEE review. The former was designed with a team of healthcare professionals and academics to enhance its quality and generalisability (D. A. Cook & West, 2012). Another strength was basing the search strategy on the guidance of two specialist librarians to ensure that all relevant materials were collected. This also involved searching in seven different databases and leaving the sample aspect of the search open so as to identify as many relevant articles as possible. Although this resulted in a predictably large initial sample, it was felt necessary to ensure that no pertinent articles were missed. Beginning the work by examining and critiquing six extant systematic literature reviews enhanced the study's design, building on their strengths and shortcomings, and broadening the findings by combining theirs with those of this study to make for a more robust data set. No other extant reviews did this even though Cook and West (2012) suggest that a key component of an SLR is establishing how it supports, contradicts, or extends the findings of previous relevant reviews.

To make this study's analysis and conclusions more credible and minimise bias, two researchers worked independently at each stage of the HEE review process. Similarly, what was deemed to be the most appropriate validated instrument to assess the quality of the included studies was applied and, unlike other reviews, the scores for each item of each article were published, along with the codes for all the included studies. This was intended to enable readers to verify the analysis presented above themselves. It is also believed that this is also the only review to devise groupings based on study credibility and most importantly, to base the analysis and conclusions on these groupings. As mentioned previously, this is perhaps a unique and one of the most significant methodological features of the HEE review in terms of its credibility. To further enhance the credibility of this study's analysis, in addition to traditional descriptive analysis, the linear regression analyses were added. Negative and outlying results were also sought out and paid close attention to, since they were felt to reveal important points about the intricacies of effective leadership development and its application, as well as identifying factors that could contribute to programmes failing. The conclusions explored outline two vitally important aspects of optimal leadership development and their various nuances in a way that is more comprehensive than what was previously available. The original Kirkpatrick model and the principles of adult learning were expanded in light of this study's analysis, which lends itself to theoretical generalisability for its relevance to theory development (Yin, 2003). The set of factors in programme design, delivery, and evaluation discussed in the second conclusion explored not only provide evidence-based guidance on facilitating the transfer of learning to the workplace, but also present factors to avoid that can set interventions up to fail. A theoretical model of leadership programme design and evaluation was generated by fusing the best available evidence with the various nuances of the phenomenon unveiled in the analysis. This model could potentially inform further research, programme design, and practice. Another strength is the juxtaposition of MULTI and HEE, which highlights the gap between the calibre of research that is needed and the current state of the literature and the repercussions of the latter, as described in the limitations section. Finally, as stated throughout, a major strength of this study is that it has attempted to be clear and transparent about what is known and most importantly, based on what evidence, so readers can judge the relevance and credibility for themselves in a field where this information is so often muddled.

7.6 Conclusion

In conclusion, given that the annual spending on leadership development worldwide is nearly half of that spent on cancer treatment, there is heightened pressure to demonstrate that this enormous and costly investment is justified. The confidence that many have in the importance of leadership and the research linking good leadership to superior performance at many levels elevates the demand for effective programmes. In leadership contexts in many professional domains, people's lives are at stake, which further reinforces the need to ensure that the development of those leaders is optimal and that the ultimate outcome of application to the workplace is being achieved. In the case of healthcare, this ideally translates directly to improved outcomes that benefit the organisation and the patient. And yet, the gaps in the research and its equivocal state, as well as the evidence of sub-optimal and even ineffective programmes, can jeopardise the lives of the patients. As mentioned in chapter one, this is a situation that simply cannot be afforded.

Thus, it was decided that this dissertation's starting point should follow Klimoski and Amos's (2012) suggestion that it is important to begin by clarifying what is known regarding optimal leadership development and what evidence substantiates this. The systematic evidence analysis was designed with this purpose at the forefront and was able to present, in a clear and transparent manner, an assessment of the current state of the literature. Not only did it demonstrate that the overall mean of study calibre, even in peer-reviewed journals, was of anecdotal quality, but it pinpointed what was missing, identified common errors, and provided suggestions for improvement.

This methodology has brought new knowledge to light, tied different pieces of available information together in novel ways, provided examples from the included studies to illustrate each point, and expanded on existing theory and knowledge. Even though some of the conclusions in this study feature practices that are not totally innovative, this thesis has demonstrated that many key research and programme design, delivery, and evaluation approaches are not being implemented consistently, or indeed, very rarely. Failing to strive for organisational or benefit to clients level-outcomes is just one example and individual studies and literature review authors neglecting to tier their findings and conclusions according to the calibre of studies is another such flaw.

For practitioners, this dissertation has offered conclusions on four levels. This begins with the conclusions from the best available evidence, which are made more robust by nuances and examples derived from other included studies. They also offer applications and extensions of perhaps the most common educational theory that is featured in leadership development design: the principles of adult learning. This, along with the twelve factors to consider before, during, and after programmes that facilitate the application of leadership to the workplace, can have potential benefits for all stakeholders involved in planning, sponsoring, or delivering interventions. There are also descriptions of factors in the second conclusion explored that are said to be responsible for programmes failing, which stakeholders would be keen to avoid. Lastly, the beginnings of a theoretical model of cardinal and complementary functions of developmental activities and the prototype outcomes-based model of leadership development design and evaluation also have the potential to be of use to those preparing or refining interventions.

For researchers, there are step-by-step recommendations to improve the quality of individual studies and literature reviews alike. Once again, there are many methodological elements that are consistently absent from studies, which affects the credibility and usefulness of the findings and conclusions. The discussion of possible revisions to MERSQI to improve its utility in assessing studies, as well as an overarching proposal to promote better quality research in the field, also have the potential to influence future research.

Finally, this methodology, and comparing the MULTI to HEE reviews in particular, symbolises the gap between the current state of the literature and its limitations (MULTI) and the kind of research that is needed to advance the field and inform policy and practice in the future (HEE). In alignment with the PRISMA guidelines and other resources, this dissertation has attempted to be as clear and transparent as possible, so readers can judge for themselves how they view these results. It is hoped that this study has established a solid evidence base, as well as generating useful suggestions and tools with which to move forward, so that the investment in leadership development can truly be returned, as evinced in improvements in the lives of the clients and patients leaders mean to serve.

8 WORKS CITED

Abrell, C., Rowold, J., Weibler, J., & Moenninghoff, M. (2011). Evaluation of a Long-term Transformational Leadership Development Program. *Zeitschrift Für Personalforschung*, *25*(3), 205–224.

Ackerly, D. C., Sangvai, D. G., Udayakumar, K., Shah, B. R., Kalman, N. S., Cho, A. H., … Dzau, V. J. (2011). Training the next generation of physician-executives: An innovative residency pathway in management and leadership. *Academic Medicine*. https://doi.org/10.1097/ACM.0b013e318212e51b

Alavi, M., & Leidner, D. E. (2001). Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, *25*(1), 107–136.

Allen, S. J., & Hartman, N. S. (2008). Leadership Development: An Exploration of Sources of Learning. *Sam Advanced Management Journal*, *26*(Summer), 75–87.

Allio, R. J. (2005). Leadership Development: Teaching Versus Learning. *Management Decision*, *43*(7/8), 1071–1077.

Alvesson, M., & Spicer, A. (2012). Critical leadership studies: The case for critical performativity. *Human Relations*, *65*(3), 367–390.

Amagoh, F. (2009). Leadership development and leadership effectiveness. *Management Decision*, *47*(6), 989–999.

Anderson, H. M. (n.d.). Dale’s Cone of Experience. Retrieved from http://www.queensu.ca/teachingandlearning/modules/active/documents/Dales_Cone_of_Experience_summary.pdf

Ardts, J. C. A., Velde, M. E. G. van der, & Maurer, T. J. (2010). The Influence of Perceived Characteristics of Management Development Programs on Employee Outcomes. *Human Resource Development Quarterly*, *21*(4), 411–434.

Artz, B., Goodall, A. H., & Oswald, A. J. (2016). Boss competence and worker well-being. *Industrial and Labor Relations Review, published online*.

Arvey, R. D., Rotundo, M., Johnson, W., Zhang, Z., & McGue, M. (2006). The determinants of leadership role occupancy: genetic and personality factors. *Leadership Quarterly, 17*(1), 1–20. https://doi.org/10.1016/j.leaqua.2005.10.009

Arvey, R. D., Zhang, Z., Avolio, B. J., & Krueger, R. F. (2007). Developmental and genetic determinants of leadership role occupancy among women. *The Journal of Applied Psychology*, *92*(3), 693–706. https://doi.org/10.1037/0021-9010.92.3.693

Avolio, B. J. (2005). *Leadership Development in Balance: Made/Born*. Oxford: Psychology Press.

Avolio, B. J., Avey, J. B., & Quisenberry, D. (2010). Estimating return on leadership development investment. *The Leadership Quarterly, 21*(4), 633–644.

Baker, G. (n.d.). The roles of leaders in high-performing healthcare systems. Retrieved 11 February 2016, from www.kingsfund.org.uk/leadershipcommission

Bandura, A. (1997). *Self-Efficacy: The exercise of control*. New York: W. H. Freeman.

Bartram, D. (2005). The Great Eight Competencies: A Criterion-Centric Approach to Validation. *Journal of Applied Psychology*, *90*(6), 1185–1203.

Bearman, M., O'Brien, R., Anthony, A., Civil, I., Flanagan, B., Jolly, B., … Nestel, D. (2012). Learning Surgical Communication, Leadership and Teamwork through Simulation. *Journal of Surgical Education*, *69*(2), 201–207.

Beer, M., Finnström, M., & Schrader, D. (2016). Why leadership training fails - and what to do about it. *Harvard Business Review*.

Beeson, J. (2004). Building bench strength: a tool kit for executive development. *Business Horizons*, *47*(6), 3–10.

Berg, M. E., & Karlsen, J. T. (2012). An evaluation of management training and coaching. *Journal of Workplace Learning*, *24*(3), 177–199.

Bergman, D., Fransson-Sellgren, S., Wahlstrom, R., & Sandahl, C. (2009). Impact of short-term intensive and long-term less intensive training programmes. *Leadership in Health Services*, *22*(2), 161–75.

Berman, E. M., Bowman, J. S., West, J. P., & Wart, M. V. (2010). *Human resource management in public service: Paradoxes, processes, and problems* (3rd ed.). Thousand Oaks, CA: Sage.

Blaikie, N. (1993). *Approaches to social enquiry*. Oxford: Polity Press.

Bloom, N., Sadun, R., & Reenen, J. V. (2014). Does Management Matter in Healthcare? *London School of Economics Working Paper,* 1–30.

Blume, B. D., Ford, J. K., Bladwin, T. T., & Huang, J. L. (2010). Training transfer: A meta-analytic review. *Journal of Management*, *38*(4), 1065–1105.

Blumenthal, D. M., Bernard, K., Bohnen, J., & Bohmer, R. (2012). Addressing the leadership gap in medicine: residents' need for systematic leadership development training. *Academic Medicine*, *87*(4), 513–522.

Blumenthal, D. M., Bernard, K., Fraser, T. N., Bohnen, J., Zeidman, J., & Stone, V. E. (2014). Implementing a pilot leadership course for internal medicine residents: design considerations, participant impressions, and lessons learned. *BMC Medical Education*, *14*(257), 1–11.

Boaden, R. J. (2006). Leadership development: Does it make a difference? *Leadership and Organization Development Journal*, *27*(1), 5–27. https://doi.org/10.1108/01437730610641331

Bohmer, R. (2012). *The Instrumental Value of Medical Leadership* (pp. 1–31). The King's Fund.

Bolden, R. (2005). *What is Leadership Development: Purpose and practice. Leadership South West Research Report* (No. 2) (pp. 1–60). Exeter: Centre for Leadership

Studies. Retrieved from http://business-school.exeter.ac.uk/documents/discussion_papers/cls/LSWreport2.pdf

Bowles, S., Cunningham, C. J. L., Rosa, G. M. D. L., & Picano, J. (2007). Coaching leaders in middle and executive management: goals, performance, buy-in. *Leadership & Organization Development Journal*, *28*(5), 388–408.

Brinkerhoff, R. O., & Gill, S. J. (1994). *The Learning Alliance: System Thinking in Human Resource Development*. San Francisco: Jossey-Bass.

British Educational Research Association (BERA). (2011). Ethical Guidelines for Educational Research. British Educational Research Association (BERA).

Bryson, A., Forth, J., & Stokes, L. (2017). Does employees' subjective well-being affect workplace performance? *Human Relations*, *70*(8), 1017–1037.

Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology, 71*(2), 232–245. https://doi.org/10.1037/0021-9010.71.2.232

Burnes, B., & O'Donnell, H. (2011). What can business leaders learn from sport? *Sport, Business and Management*, *1*(1), 12–27.

BusinessWeek/Hay Group. (2010). *Selected Results from Best Companies for Leadership Survey*. Retrieved from http://www.businessweek.com/careers/special_reports/20100216best_companies_for_leadership.htm

Butler, D., Forbes, B., & Johnson, L. (2008). An Examination of a Skills-Based Leadership Coaching Course in an MBA Program. *Journal of Education for Business,* (March/April), 227–232.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24,* 409–429.

Candace, I., & Giordano, R. W. (2009). Doctors as leaders. *British Medical Journal, 338*, b1555.

Carlyle, T. (2007). *On Heroes, Hero-Worship, and The Heroic in History*. London: Echo Library.

Castro, P., Dorgan, S., & Richardson, B. (2008). *The McKinsey Quarterly: A healthier health care system for the United Kingdom*. McKinsey and the London School of Economics. Retrieved from http://www.washburn.edu/faculty/rweigand/McKinsey/McKinsey-Healthier-Care-In-UK.pdf

Catlin, A., Cowan, C., Heffler, S., & Washington, B. (2007). National health spending in 2005: The slowdown continues. *Health Affairs*, *26*(1), 142–153.

CCACE Research Group. (2013). Systematic reviews and meta-analyses: a step-by-step guide. Retrieved from http://www.ccace.ed.ac.uk/research/software-resources/systematic-reviews-and-meta-analyses

Centre for Reviews and Dissemination. (2016, February 11). Centre for Reviews and Dissemination Systematic Reviews CRD's Guidance for Undertaking Reviews in Health Care. Retrieved from http://www.york. ac.uk/media/crd/Systematic_Reviews.pdf

Cherry, R., Davis, D., & Thorndyke, L. (2010). Transforming Culture through Physician Leadership. *Physician Executive*, *36*(3), 38–44.

Chochard, Y., & Davoine, E. (2011). Variables influencing the return on investment in management training programs: a utility analysis of 10 Swiss cases: Management training programs. *International Journal of Training and Development*, *15*(3), 225–243. https://doi.org/10.1111/j.1468-2419.2011.00379.x

Clark, J., & Armit, K. (2010). Leadership Competency for Doctors: A Framework. *Leadership in Health Services*, *23*(2), 115–129.

Clarke, N. (2012). Evaluating Leadership Training and Development: A Levels-of-Analysis Perspective. *Human Resource Development Quarterly*, *23*(4), 441–460. https://doi.org/10.1002/hrdq.21146

Clay-Williams, R., Nosrati, H., Cunningham, F. C., Hillman, K., & Braithwaite, J. (2014). Do large-scale hospital- and system-wide interventions improve patient outcomes: a systematic review. *BMC Health Services Research*, *14*(369), 1–13. https://doi.org/10.1186/1472-6963-14-369

CMO Clinical Advisor Alumni. (2012). Leadership Development for Early Career Doctors. *Lancet*, *379*, 1847–1849.

Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. (2011, March). The Cochrane Collaboration. Retrieved from http://training.cochrane.org/handbook

Collins, D. B., & Holton III, E. E. (2004). The Effectiveness of Managerial Leadership Development Programs: A Meta-Analysis of Studies from 1982 to 2001. *Human Resource Development Quarterly*, *15*(2), 217–248.

Coloma, J., Gibson, C., & Packard, T. (2012). Participant Outcomes of a Leadership Development Initiative in Eight Human Service Organizations. *Administration in Social Work*, *36*, 4–22.

Combs, J., Liu, Y., Hall, A., & Ketchen, D. (2006). How much do high-performance work practices matter? A meta-analysis of their effects on organizational performance. *Personnel Psychology*, *59*(3), 501–528.

Cook, D. A., & West, C. P. (2012). Conducting systematic reviews in medical education: a stepwise approach. *Medical Education*, *46*, 943–952.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Skokie, IL: Rand McNally.

Cresswell, J. W. (2003). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (Second). Thousand Oaks, CA: Sage.

Critical Appraisal Skills Programme (CASP). (2013, May 31). Qualitative Research Checklist. Retrieved 11 February 2016, from http://media.wix.com/ugd/dded87_29c5b002d99342f788c6ac670e49f274.pdf

Crotty, M. (2003). *The foundations of social research: meaning and perspective in the research process* (2nd ed.). London: Sage.

Cutler, D. M. (2004).

System. New York: Oxford University Press.

Daimler, M. (2016). Why Leadership Development Has to Happen on the Job. *Harvard Business Review*. Retrieved from https://hbr.org/2016/03/why-leadership-development-has-to-happen-on-the-job

Dalakoura, A. (2010a). Differentiating Leader and Leadership Development: A Collective Framework for Leadership Development. *The Journal of Management Development, 29*(5), 432–441.

Dalakoura, A. (2010b). Examining the effects of leadership development on firm performance. *Journal of Leadership Studies, 4*(1), 59–70.

Dale, E. (1969). *Audiovisual Methods in Teaching* (3rd ed.). New York: The Dryden Press.

Dannels, S. A., McLaughlin, J. M., Gleason, K. A., McDade, S. A., Richman, R. C., & Morahan, P. S. (2009). Medical school deans' perceptions of organizational climate: Useful indicators for advancement of women faculty and evaluation of a leadership program's impact. *Academic Medicine, 84*(1), 67–79.

Dannels, S. A., Yamagata, H., McDade, S. A., Chuang, Y.-C., Gleason, K. A., McLaughlin, J. M., … Morahan, P. S. (2008). Evaluating a Leadership Program: A Comparative, Longitudinal Study to Assess the Impact of the Executive Leadership in Academic Medicine (ELAM) Program for Women. *Academic Medicine, 83*(5), 488–495.

Darzi, A. (2008). *High Quality Care For All: Next Stage Review Final Report*. London: Department of Health.

Davies, H. T. O., & Harrison, S. (2003). Trends in doctor-manager relationships. *British Medical Journal*, *326*, 646–649.

Day, C. S., Tabrizi, S., Kramer, J., Yule, A., & Ahn, B. (2010). Effectiveness of the AAOS Leadership Fellows Program for Orthopaedic Surgeons. *Journal of Bone and Joint Surgery of America*, *92*(16), 2700–2708.

Day, D. V. (2000). Leadership development: A review in context. *Leadership Quarterly*, *11*(4), 581–613.

Day, D. V. (2004). Leadership Development. In G. R. Goethals, G. J. Sorenson, & J. M. Burns (Eds.), *Encyclopedia of Leadership* (Vol. 2, pp. 840–844). Thousand Oaks, CA: SAGE.

Day, D. V., & Halpin, S. M. (2001). *Leadership development: a review of industry best practices. Review on corporate training.* (Technical Report) (pp. 1–61). Alexandria, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences.

Day, D. V., & O'Connor, P. M. (2003). Leadership development: Understanding the process. In S. E. Murphy & P. Riggio (Eds.), *The future of leadership development* (pp. 11–28). Mahwah, NJ: Lawrence Erlbaum.

Day, D. V., & Sin, H.-P. (2011). Longitudinal Tests of an Integrative Model of Leader Development: Charting and Understanding Developmental Trajectories. *The Leadership Quarterly*, *22*, 545–560.

Day, M. (2007). The Rise of the Doctor-Manager. *British Medical Journal*, *335*(7613), 230–231.

Deaton, A. (2009). *Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development*. Presented at the The Keynes Lecture, British Academy, London.

Denscombe, M. (2010). *The good research guide for small-scale social research projects* (4rth ed.). McGraw Hill: Open University Press.

Denzin, N., & Lincoln, Y. S. (Eds.). (2003). *Collecting and interpreting qualitative materials* (3rd ed.). London: Sage.

DeRue, D. S., Nahrgang, J. D., Hollenbeck, J. R., & Workman, K. (2012). A Quasi-Experimental Study of After-Event Reviews and Leadership Development. *Journal of Applied Psychology*, *97*(5), 997–1015.

DeRue, D. S., & Wellman, N. (2009). Developing Leaders via Experience: The Role of Developmental Challenge, Learning Orientation, and Feedback Availability. *Journal of Applied Psychology*, *94*(4), 859–875.

Dexter, B., & Prince, C. (2007). Evaluating the impact of leadership development: a case study. *Journal of European Industrial Training*, *31*(8), 609–625.

Dickey, C., Dismukes, R., & Topor, D. (2014). Creating Opportunities for Organizational Leadership (COOL): Creating a culture and curriculum that fosters psychiatric leadership development and quality improvement. *Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*. https://doi.org/10.1007/s40596-014-0082-2

Dine, C. J., Kahn, J. M., Abella, B. S., & Shea, J. A. (2011). Key Elements of Clinical Physician Leadership at an Academic Medical Center. *Journal of Graduate Medical Education*, (March), 31–36.

D'Netto, B., Bakas, F., & Bordia, P. (2008). Predictors of management development effectiveness: an Australian perspective. *International Journal of Training and Development*, *12*(1), 2–23.

Doran, D., McCutcheon, A., Evans, M., MacMillan, K., Hall, L., Pringle, D., … Valente, A. (2004). . Ottawa: Canadian Health Services Research Foundation.

Dowrick, A., Wootten, A., Murphy, D., & Costello, A. (2015). We Used a Validated Questionnaire": What Does This Mean and Is It an Accurate Statement in Urologic Research? *Urology*, *85*(6), 1304–1310.

Drew, G. (2009). A ‘360’ degree view for individual leadership development. *Journal of Management Development*, *28*(7), 581–592.

Dugan, J. P. (2011). Pervasive myths in leadership development: Unpacking constraints on leadership learning. *Journal of Leadership Studies*, *5*(2), 79–84.

Dunning, T. (2012). *Natural experiments in the -based approach*. Cambridge, UK: Cambridge University Press.

Dwyer, A. J. (2010). Medical managers in contemporary healthcare organisations: a consideration of the literature. *Australian Health Review*, *34*, 514–522.

Edler, A., Adamshick, M., Fanning, R., & Piro, N. (2010). Leadership lessons from military education for postgraduate medical curricular improvement. *The Clinical Teacher*, *7*(1), 26–31. https://doi.org/10.1111/j.1743-498X.2009.00336.x

Edmans, A. (2011). Does the stock market fully value intangibles? Employee satisfaction and equity prices. *Journal of Financial Economics*, *101*, 621–640.

Edmonstone, J. (2009). Evaluating clinical leadership: A case study. *Leadership in Health Services*, *22*(3), 210–224.

Edmonstone, J. (2011). The development of strategic clinical leaders in the National Health Service in Scotland. *Leadership in Health Services*, *24*(4), 337–353.

Edmonstone, J. (2013). Healthcare leadership: learning from evaluation. *Leadership in Health Services*, *26*(2), 148–158.

Edwards, N., Komacki, M., & Silversin, J. (2002). Unhappy doctors: what are the causes and what can be done? *British Medical Journal*, *324*, 835.

Ely, K., Boyce, L. A., Nelson, J. K., Zaccaro, J., Hernez-Broome, G., & Whyman, W. (2010). Evaluating leadership coaching: a review and integrated framework. *The Leadership Quarterly*, *21*, 585–599.

Ercikan, K., & Roth, W.-M. (2006). What Good Is Polarizing Research Into Qualitative and Quantitative? *Educational Researcher*, *35*(5), 14–23.

Falcone, R. E., & Santiani, B. (2008). Physician as Hospital Chief Executive Officer. *Vascular and Endovascular Surgery*, *42*(1), 88–94.

Fallesen, J. J., Keller-Glaze, H., & Curnow, C. K. (2011). A selective review of leadership studies in the U.S. Army. *Military Psychology*, *23*(5), 462–478. https://doi.org/10.1080/08995605.2011.600181

Fayolle, A., & Gailly, B. (2008). From craft to science: Teaching models and learning processes in entrepreneurship education. *Journal of European Industrial Training*, *32*(7), 569–593.

Fernandez, C. S. P., Noble, C. C., Jensen, E. T., & Chapin, J. (2016). Improving leadership skills in physicians: A 6-month retrospective study. *Journal of Leadership Studies*, *9*(4), 6–19.

Fiedler, K., & Kareev, Y. (2010). Clarifying the Advantage of Small Samples: As It Relates to Statistical Wisdom and Cahan’s (2010) Normative Intuitions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 1039–1043.

Finkelstein, L. M., & Poteet, M. L. (2007). Best practices in workplace formal mentoring programs. In T. D. Allen & L. T. Eby (Eds.), *The Blackwell handbook of mentoring: A multiple perspectives approach* (pp. 345–368). Malden, MA: Blackwell.

Freedman, E. (2012). Building the Talent Pipeline for Future Leaders. *Financial Executive*, *28*(9), 89–91.

Freidson, E. (1983). The Theory of Professions: State of the Art. In R. Dingwall & P. Lewis (Eds.), *The Sociology of the Professions*. London: The MacMillan Press LTD.

Freire, P. (1992). *Pedagogy of Hope*. London: Continuum.

Freire, P. (2007). *Pedagogy of the Oppressed* (30th Anniversary). London: Continuum.

Frich, J. C., Brewster, A. L., Cherlin, E. J., & Bradley, E. H. (2014). Leadership Development Programs for Physicians: A Systematic Review. *Journal of General Internal Medicine*, *30*(5), 656–74.

Gagnon, S., Rithchie, J., Lynch, A., Drouin, S., Cass, V., Rinfret, N., … Valois, M. (2006). *Job Satisfaction and Retention of Nursing Staff: The Impact of Nurse Management Leadership*. Toronto: Canadian Health Services Research Foundation.

Galli, E. B., & Muller-Stewens, G. (2012). How to build social capital with leadership development: Lessons from an explorative case study of a multibusiness firm. *The Leadership Quarterly*, *23*, 176–201.

Geerts, J. (2009). *Catholic Secondary School Leadership: A Definition and Key Skills and Qualities*. University of St. Michael’s College, Toronto, Canada.

Geerts, W. H., Bergqvist, D., Pineo, G. F., Heit, J. A., Samama, C. M., Lassen, M. R., & Colwell, C. W. (2008). Prevention of venous thromboembolism: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy (8th edition). *Chest*, *133*, 381–453.

Geertz, C. (1973). *The interpretation of cultures: Selected essays*. New York: Basic Books.

General Medical Council. (2009, September). Tomorrow’s Doctors: Outcomes and Standards for Undergraduate Medical Education. General Medical Council.

Getha-Taylor, H., & Morse, R. S. (2013). Collaborative leadership development for local government officials: Exploring competencies and program impact. *Public Administration Quarterly*, (Spring), 72–103.

Giber, D., Carter, L., & Goldsmith, M. (2000).

Development Handbook. San Francisco: Jossey-Bass/Pfeiffer.

Giganti, E. (2003). Developing leaders for 2010. *Health Progress*, *84*(1), 11–12.

Gilpin-Jackson, Y., & Bushe, G. R. (2007). Leadership development training transfer: a case study of post-training determinants. *Journal of Management Development, 26*(10), 980–1004.

Goodall, A.H. (2011). Physician-leaders and hospital performance: Is there an association? *Social Science & Medicine, 73*(4), 535–539.

Goodall, A. H. (2009). Highly cited leaders and the performance of research universities. *Research Policy, 38*, 1079–1092.

Goodall, A. H. (2011). Physician-leaders and hospital performance: Is there an association? *Social Science & Medicine, 73*(4), 535–539.

Goodall, A. H., & Baker, A. (2015). A theory exploring how expert leaders influence performance in knowledge-intensive organizations. In I. M. Welpe, J. Wollersheim, S. Ringelhan, & M. Osterloh (Eds.), *Incentives and Performance: Governance of Knowledge-Intensive Organizations* (pp. 49–68). Heidelberg: Springer International Publishing.

Goodall, A. H., Kahn, L. M., & Oswald, A. J. (2011). Why do leaders matter? A study of expert knowledge in a superstar setting. *Journal of Economic Behavior and Organization*, *77*, 265–284.

Goodall, A. H., & Pogrebna, G. (2015). Expert leaders in a fast-moving environment. *Leadership Quarterly, 26*(2), 123–142.

Goodall, A. H., & Stoller, J. K. (2017). The future of clinical leadership: evidence for physician leadership and the educational pathway for new leaders. *BMJ Leader*, *0*, 1–4.

Green, M. E. (2002). Ensuring the organization’s future: a leadership development case study. *Public Personnel Management, 31*(4), 431–439.

Green, S., Higgins, J., Alderson, P., Clarke, M., Mulrow, C., & Oxman, A. (2011). Chapter 1: Introduction. In J. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic*

Reviews of Interventions (Version 5.1.0 (updated March 2011)). The Cochrane Collaboration.

Greene, J. C. (2008). Is Mixed Methods Social Inquiry a Distinctive Methodology? *Journal of Mixed Methods Research*, *2*(1), 7–22.

Gronn, P. (2002). Leader Formation. In K. Keithwood & P. Hallinger (Eds.), *Second International Handbook of Educational Leadership and Administration* (pp. 1031–1070). Great Britain: Kluwer Academic Publishers.

Gronn, P. (2004). Distributing Leadership. In G. R. Goethals, G. J. Sorenson, & J. M. Burns (Eds.), *Encyclopedia of Leadership* (Vol. 1, pp. 352–355). Thousand Oaks, CA: SAGE.

Gronn, P. (2010). Leadership: Its genealogy, configuration, and trajectory. *Journal of Educational Administration and History*, *42*(4), 405–435.

Grotrian-Ryan, S. (2015). Mentoring Functions and their Application to the American Council on Education (ACE) Fellows Leadership Development Program. *International Journal of Evidence Based Coaching and Mentoring*, *13*(1), 87–105.

Groves, K. S. (2007). Integrating leadership development and succession planning best practices. *Journal of Management Development*, *26*(3), 239–260.

Gunderman, R., & Kanter, S. L. (2009). Educating physicians to lead hospitals. *Academic Medicine*, *84*, 1348–1351.

Gurrera, R. J., Dismukes, R., Edwards, M., Feroze, U., Nakshabandi, F., Tanaka, G., & Tang, M. (2014). Preparing residents in training to become health-care leaders: a pilot project. *Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*. https://doi.org/10.1007/s40596-014-0162-3

Guskey, T. R. (2002). What Makes Professional Development Effective? *The Phi Delta Kappan*, *84*(10), 748–750.

Guyatt, G., Oxman, A., Vist, G., Kunz, R., Flack-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, *336*, 924–926.

Hägerland, T. (2015). The Future of Criteria in Historical Jesus Research. *Journal for the Study of the Historical Jesus*, *13*, 43–65.

Halbesleben, J. R., & Rathert, C. (2008). Linking physician burnout and patient outcomes: exploring the dyadic relationship between physicians and patients. *Health Care Management Review*, *33*(1), 29–39.

Halligan, A. (2008). Aidan Halligan on why Darzi needs clinical leadership. *Health Services Journal*.

Ham, C., & Dickenson, H. (2008). *Engaging doctors in leadership: what can we learn from international experience and research evidence?* U.K.: NHS Institute for Innovation and Improvement.

Hamlin, B. (2010). Evidence-based leadership and management development. In J. Gold, R. Thorpe, & A. Mumford (Eds.), *Gower Handbook of Leadership and Management Development* (5th ed., pp. 197–220). Surrey, England: Gower.

Hannum, K. M., & Bartholomew, C. S. (2010). Introduction to special issue on leadership development evaluation. *Leadership Quarterly*, *21*(4), 581–582.

Harden, R. M., Grant, J., Buckley, G., & Hart, I. R. (1999). BEME Guide No. 1: Best Evidence Medical Education. *Medical Teacher*, *21*(6), 553–562.

Hartley, J., & Benington, J. (2010). *Leadership for healthcare*. Bristol: Policy Press.

Hartley, J., & Hinksman, B. (2003). *Leadership development: A systematic literature review. A report for the NHS Leadership Centre* (pp. 1–77). Warwick, UK: Warwick Institute of Governance and Public Management.

Hassan, R. A., A.Fuwad, B., & Rauf, A. I. (2010). Pre-training motivation and the effectiveness of transformational leadership training: An experiment. *Academy of Strategic Management Journal*, *9*(2), 1–8.

Hatch, M. (1993). The dynamics of organizational culture. *Academy of Management Review*, *18*(4), 657–693. https://doi.org/doi:10.5465/AMR.1993.9402210154

Hayes, J. (2007). Evaluating a Leadership Development Program. *Organization Development Journal*, *25*(4), P89–P94.

Haynes, A. B., Weiser, T. G., Berry, W. R., Lipsitz, S. R., Breizat, A.-H. S., Dellinger, E. P., … Gawande, A. A. (2009). A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population. *New England Journal of Medicine*, (360), 491–499.

Heifetz, R. A. (1994). *Leadership without easy answers*. Boston: Harvard University Press.

Hemmer, P. R., Karon, B. S., Hernandez, J. S., Cuthbert, C., Fidler, M. E., & Tazelaar, H. D. (2007). Leadership and Management Training for Residents and Fellows: A Curriculum for Future Medical Directors. *Archives of Pathology & Laboratory Medicine*, *131*, 610–614.

Hernez-Broome, G., & Hughes, R. L. (2004). Leadership development: past, present, and future. *Human Resource Planning*, *27*(1), 24–32.

Hiller, N. J., DeChurch, L. A., Murase, T., & Doty, D. (2011). Searching for Outcomes of Leadership: a 25-Year Review. *Journal of Management*, *37*(4), 1137–1177.

Hirst, G., Mann, L., Bain, P., Pirola-Merlo, A., & Richver, A. (2004). Learning to lead: the development and testing of a model of leadership learning. *Leadership Quarterly*, *15*(3), 311–327.

Hockey, P. M., & Bates, D. W. (2010). Physicians' Identification of Factors Associated with Quality in High- and Low-Performing Hospitals. *The Joint Commission Journal on Quality and Patient Safety*, *36*(5), 217–224.

Horton, R. (2008). The Darzi vision: Quality, engagement, and professionalism. *The Lancet*, *372*, 3–4.

House of Commons Health Committee. (2005). *The Prevention of Venous Thromboembolism in Hospitalised Patients: Second Report of Session 2004 05* (No. HC 99). London: House of Commons London: The Stationery Office Limited.

Hunziker, S., Bühlmann, C., Tschan, F., Balestra, G., Legeret, C., Schumacher, C., … Marsch, S. (2010). Brief leadership instructions improve cardiopulmonary resuscitation in a high-fidelity simulation: A randomized controlled trial. *Critical Care Medicine*, *38*(4), 1086–1091. https://doi.org/doi: 10.1097/CCM.0b013e3181cf7383

Husebø, S. E., & Akerjordet, K. (2016). Quantitative systematic review of multi professional teamwork and leadership training to optimize patient outcomes in acute hospital settings. *Journal of Advanced Nursing*, *72*(12), 2980–3000. https://doi.org/10.1111/jan.13035

Ibarra, P. (2005). Succession planning: an idea whose time has come. *Public Management*, *87*(1), 18–23.

Ireri, S., Walshe, K., Benson, L., & Mwanthi, M. (2011). A Qualitative and Quantitative Study of Medical Leadership and Management: Experiences, Competencies, and Development Needs of Doctor Managers in the United Kingdom. *Jounral of Management and Marketing in Healthcare*, *4*(1), 16–29.

Jackson, W., & Verberg, N. (2007). *Methods: Doing Social Research* (Fourth). Toronto: Pearson Education Inc.

Jeon, Y.-H., Simpson, J. M., Chenoweth, L., Cunich, M., & Kendig, H. (2013). The effectiveness of an aged care specific leadership and management program on workforce, work environment, and care quality outcomes: design of a cluster randomised controlled trial. *Implementation Science*, *8*, 126–136.

Jesuthasan, R., & Holmstrom, M. S. (2016). As Work Changes, Leadership Development Has to Keep Up. *Harvard Business Review*. Retrieved from https://hbr.org/2016/10/as-work-changes-leadership-development-has-to-keep-up?referral=03758&cm_vc=rr_item_page.top_right

Jiang, K., Lepak, D., Hu, J., & Baer, J. (2012). How does human resource management influence organizational outcomes? A meta-analytic investigation of mediating mechanisms. *Academy of Management Journal*, *55*(6), 1264–1294.

Johnson, P., Zualkernan, I., & Garber, S. (1987). Specification of expertise. *International Journal of Man-Machine Studies*, *26*, 161–18.

Johnson, S. K., Garrison, L. L., Hernez-Broome, G., Fleenor, J. W., & Steed, J. L. (2012). Go For the Goal(s): Relationship Between Goal Setting and Transfer of Training Following Leadership Development. *Academy Oí Management Learning & Education*, *11*(4), 555–569.

Jones, S., McCay, L., & Keogh, S. B. (2011). The Importance of Clinical Leadership. In T. Swanwick & J. McKimm (Eds.), *ABC of Clinical Leadership* (pp. 1–3). Chichester, UK: BMJ Publishing Group Ltd.

Jong, N. de, Könings, K. D., & Czabanowska, K. (2014). The Development of Innovative Online Problem-Based Learning: A Leadership Course for Leaders in European Public Health. *Journal of University Teaching & Learning Practice*, *11*(3), 1–9.

Judge, T. A. (2016, April). *Losing Leadership: Problems and Opportunities In Leadership Scholarship*. Judge Business School, University of Cambridge.

Keller, R., Julian, S., & Kedia, B. (n.d.). A multinational study of work climate, job satisfaction, and the productivity of r&d teas. *IEEE Transactions of Engineering Management*, *43*(48–55), 1996.

Kellerman, B. (2012). *The End of Leadership*. New York: Harper Collins.

Kesler, G. C. (2002). Why the leadership bench never gets deeper: Ten insights about executive talent development. *Human Resource Planning*, *25*(1), 32–44.

Kesner, I. (2003). Leadership Development: Perk or Priority. *Harvard Business Review,* (May), 29–38.

Kim, T. H., & Thompson, J. M. (2012). Organizational and Market Factors Associated with Leadership Development Programs in Hospitals: A National Study. *Journal of Healthcare Management*, *57*(2), 113–131.

Kirkpatrick, D., & Kirkpatrick, J. (2006). *Evaluating Training Programs* (3rd ed.). San Francisco: Berrett Koehler Publishers.

Klimoski, R., & Amos, B. (2012). Practicing Evidence-Based Education in Leadership Development. *Academy of Management Learning and Education*, *11*(4), 18.

Kneebone, R. (2005). Evaluating clinical simulations for learning procedural skills: a theory-based approach. *Academic Medicine, 80,* 549–553.

Knowles, M. S. (1981). *The Modern Practice of Adult Education: From Pedagogy to Andragogy* (Revised and Updated). Englewood Cliffs, NJ: Prentice Hall Regents,.

Knowles, M. S. (1984). *Andragogy in Action*. San Francisco: Jossey-Bass.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development.* Englewood Cliffs, N.J.: Prentice Hall.

Komives, S. R., Nance, L., & McMahon, T. R. (1998). *Exploring Leadership: For College Students Who Want To Make a Difference*. San Francisco: Jossey-Bass.

Korschun, H., Redding, D., Teal, G., & Johns, M. (2010). Realizing the Vision of Leadership Development in an Academic Health Center: The Woodruff Leadership Academy. *Academic Medicine, 82*, 264–271.

Korschun, H. W., Redding, D., Teal, G. L., & Johns, M. M. E. (2007). Realizing the vision of leadership development in an academic health center: The Woodruff Leadership

Academy. *Academic Medicine*, *82*(3), 264–271. https://doi.org/10.1097/ACM.0b013e31803078b5

Kunzle, B., Kolbe, M., & Grote, G. (2010). Ensuring patient safety through effective leadership behaviour: a literature review. *Safety Science, 48*, 1–17.

Kuo, A. K., Thyne, S. M., Chen, H. C., West, D. C., & Kamei, R. K. (2010). An Innovative Residency Program Designed to Develop Leaders to Improve the Health of Children. *Academic Medicine*, *85*(10), 1603–1608.

Kwamie, A., Dijk, H. van, & Agyepong, I. A. (2014). Advancing the application of systems thinking in health: realist evaluation of the Leadership Development Programme for district manager decision-making in Ghana. *Health Research Policy and Systems*, *12*(29), 1–12.

Ladyshewsky, R. K. (2007). A strategic approach for integrating theory to practice in leadership development. *Leadership & Organization Development Journal*, *28*(5), 426–443.

Latham, G., & Locke, E. (1983). Goal setting - a motivational technique that works. In J. Hackman, E. Lawlor, & L. Porter (Eds.), *Perspectives on Behavior in Organizations* (pp. 296–304). New York: McGraw Hill.

Lazarus, A. (2009). Professional and career issues in administrative medicine. *Journal of Healthcare Leadership*, *1*, 1–5.

Lee, T. H. (2010). Turning Doctors into Leaders. *Harvard Business Review*, (April), 50–58.

Lefebvre, C., Manheimer, E., & Glanville, J. (2011). Chapter 6: Searching for studies. In J. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (Version 5.1.0 (updated March 2011)). The Cochrane Collaboration.

Leskiw, S., & Singh, P. (2007). Leadership development: learning from best practices. *Leadership & Organization Development Journal*, *28*(5), 444–464. https://doi.org/https://doi.org/10.1108/01437730710761742

Leslie, L. K., Miotto, M. B., Liu, G. C., Ziemnik, S., Cabrera, A. G., Calma, S., … Slaw, K. (2005). Training Young Pediatricians as Leaders for the 21st Century. *Pediatrics*, *115*(3), 765–773.

Lester, W., Freemantle, N., Begaj, I., Ray, D., Wood, J., & Pagano, D. (2013). Fatal venous thromboembolism associated with hospital admission: a cohort study to assess the impact of a national risk assessment target. *Heart*, *99*(23), 1734–1739.

Liberati, A., Altman, D., Tetzlaff, J., Mulrow, C., Gotzche, P., Ioannides, J., … Moher, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. 2009;151:264–269, W64. *Journal of Clinical Epidemiology*, *62*, e1–e34.

Lincoln, Y., & Guba, E. (1985). *Naturalistic Inquiry*. London: Sage.

Mabey, C., & Finch-Lees, T. (2008). *Management and leadership development*. London: Sage.

Mabey, Christopher, & Ramirez, M. (2005). Does management development improve organizational productivity? A six country analysis of European firms. *International Journal of Human Resource Management*, *16*(7), 1067–1082.

Mabey, Christopher, & Thompson, A. (2001). Achieving Management Excellence: a survey of UK management development at the millennium. *European Business Review*, *13*(1). https://doi.org/10.1108/ebr.2001.05413aab.001

Maccoby, M., Norman, C. L., Norman, C. J., & Margolies, R. (2013). *Transforming health care leadership*. San Francisco: Jossey-Bass.

MacLeod, L. (2012). Making SMART goals smarter. *Physician Executive Journal*, 68–72.

MacPhail, A., Young, C., & Ibrahim, J. E. (2015). Workplace-based clinical leadership training increases willingness to lead. *Leadership in Health Services*, *28*(2), 100–118.

Malling, B., Mortensen, L., Bonderup, T., Scherpbier, A., & Ringstead, C. (2009). Combining a Leadership Course and Multi-Source Feedback Has No Effect on

Leadership Skills of Leaders in Postgraduate Medical Education. An Intervention Study with a Control Group. *Medical Education*, *9*(72), 1–7.

Mannion, R., Davies, H. T. O., & Marshall, M. N. (2005). Cultural characteristics of ‘high’ and ‘low’ performing hospitals. *Journal of Health Organisation and Management*, *19*(6), 431–439.

March, J. G., Sproull, L. S., & Tamuz, M. (2003). Learning from samples of one or fewer. *Quality and Safety in Health Care*, *12*(6), 465–472.

Marchal, B., Dedzo, M., & Kegels, G. (2010). A realist evaluation of the management of a well-performing regional hospital in Ghana. *BMC Health Service Research, 10*(1), 24.

Marcus, M. (2004). Preparing high-potential staff for the step up to leadership. *Canadian HR Reporter*, *17*(18), 11–12.

Mark, M. M. (2008). Emergence in and from quasi-experimental design and analysis. In S. N. Hesse-Biber & P. Leavy (Eds.), *Handbook of Emergent Methods* (pp. 87–110). London: Guilford Press.

Marshall, C., & Rossman, G. B. (2011). *Designing qualitative research* (5th ed.). Thousand Oaks, CA: Sage.

Martins, H. M. G. (2010). Why management and leadership education for internists? *European Journal of Internal Medicine*, *21*, 374–376.

Mason, J. (1997). *Qualitative researching*. Thousand Oaks, CA: Sage.

Maxim, P. S. (1999). *Quantitative Research Methods in the Social Sciences*. Oxford: Oxford University Press.

McAlearney, A. S. (2010). Executive Leadership Development in U.S. Health Systems/PRACTITIONER APPLICATION. *Journal of Healthcare Management*, *55*(3), 206–224.

McAlearney, A. S., & Butler, P. W. (2008). Using Leadership Development Programs to Improve Quality and Efficiency in Healthcare. *Journal of Healthcare Management, 53*(5), 319–332.

McAlearney, A. S., Fisher, D., Heiser, K., Robbins, D., & Kelleher, K. (2005). Developing Effective Physician Leaders: Changing Cultures and Transforming Organizations. *Hospital Topics*, *83*(2), 11–18.

McCall, M., & Morrison, L. (1988). *The lessons of experience: how successful executives develop on the job*. New York: Free Press.

McCauley, C. D. (2008). *Leader Development: A Review of Research*. Center for Creative Leadership.

McCormick, M. J. (2001). Self-Efficacy and Leadership Effectiveness: Applying Social Cognitive Theory to Leadership. *Journal of Leadership & Organizational Studies*, *8*(1), 22–33. https://doi.org/DOI: https://doi.org/10.1177/107179190100800102

McCormick, M. J., Tanguma, J., & López-Forment, A. S. (2002). Extending Self-Efficacy Theory to Leadership: A Review and Empirical Test. *Journal of Leadership Education*, *1*(2), 34–49.

McGurk, P. (2010). Outcomes of management and leadership development. *Journal of Management Development*, *29*(5), 457–470.

McKenna, M. K., & Pugno, P. A. (2006). *Physicians as Leaders Who, How and Why Now?* United Kingdom: Radcliffe-Oxford.

McKimm, J., & Phillips, K. (2009). Introduction. In J. McKimm & K. Phillips (Eds.), *Leadership and Management in Integrated Services*. Exeter: Learning Matters Ltd.

McKimm, J., & Swanwick, T. (2011). Leadership Development for Clinicians: What Are We Trying to Achieve? *The Clinical Teacher*, *8*, 181–185.

Miller, D. L., Umble, K. E., Frederick, S. L., & Dinkin, D. R. (2007). Linking learning methods to outcomes in public health leadership development. *Leadership in Health Services*, *20*(2), 97–123.

Millward, L., & Bryan, K. (2005). Clinical leadership: a position statement. *Leadership in Health Services*, *18*(2), xiii–xxxv.

Mintzberg, H. (1998). Covert leadership: Notes on managing professionals. *Harvard Business Review*, *76*(6), 140–147.

Montesino, M. U. (2002). Strategic alignment of training, transfer - enhancing behaviors, and training usage: a post-training study. *Human Resource Development Quarterly*, *13*(1), 89–108.

Morahan, P. S., Gleason, K. A., Richman, R. C., Dannels, S. A., & McDade, S. A. (2010). Advancing women faculty to senior leadership in US academic health centers: Fifteen years of history in the making. *Journal About Women Higher Education*, *3*(1), 137–162.

Mountford, J., & Webb, C. (2009). When clinicians lead. *Health International*, *9*, 18–25.

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2007). The leadership skills strataplex: Leadership skill requirements across organizational levels. *Leadership Quarterly*, *18*(2), 154–166. https://doi.org/10.1016/j.leaqua.2007.01.005

Murdock, J., & Brammer, C. (2011). A Successful Model of Leadership Development for Community Practice Physicians. *Physician Executive Journal*, *March/April*, 52–56.

Nabi, G., Liñán, F., Fayolle, A., Krueger, N., & Walmsley, A. (2017). The Impact of Entrepreneurship Education in Higher Education: A Systematic Review and Research Agenda. *Academy of Management Learning & Education*, *16*(2), 277–299. https://doi.org/https://doi.org/10.5465/amle.2015.0026

Nakanjako, D., Namagala, E., Semeere, A., Kigozi, J., Sempa, J., Ddamulira, J. B., … Sewankambo, N. (2015). Global health leadership training in resource-limited settings: a collaborative approach by academic institutions and local health care programs in Uganda. *Human Resources for Health*, *13*, 87. https://doi.org/10.1186/s12960-015-0087-2

NHS Institute for Innovation and Improvement and Academy of Medical Royal Colleges. (2010). Medical Leadership Competency Framework: Enhancing Engagement in Medical Leadership. Third Edition. NHS Institute for Innovation and Improvement.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organizational Science*, *5*(1), 14–37.

Northouse, P. G. (2006). *Leadership: Theory and Practice*. London: Sage.

NTL Institute. (n.d.). National Training Laboratories. Retrieved from http://www.ntl.org/

O’Connor, D., Green, S., & Higgins, J. (2008). Chapter 5: Defining the review question and developing criteria for including studies. In J. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (Version 5.0.0). The Cochrane Collaboration.

Oswald, A. J., Proto, E., & Sgroi, D. (2015). Happiness and productivity. *Journal of Labor Economics*, *33*(4), 789–822.

Palmer, P. (1998). *The Courage to Teach*. San Francisco: Jossey-Bass.

Patel, N., Brennan, P. J., Metlay, J., Bellini, L., Shannon, R. P., & Myers, J. S. (2015). Building the pipeline: the creation of a residency training pathway for future physician leaders in health care quality. *Journal of the Association of American Medical Colleges*, *90*(2), 185–190. https://doi.org/10.1097/ACM.0000000000000546

Patterson, M., Warr, P., & West, M. (77). Organizational climate and company productivity: the role of of employee affect and employee level. *Journal of Occupational and Organizational Psychology*, *77*, 193–216.

Peccei, R., Van de Voorde, K., & Van Veldhoven, M. (2013). HRM, well-being, and performance: a theoretical adn empirical review. In J. Paauwe, D. Guest, & P. Wright

(Eds.), *HRM and performance: achievements and challenges* (pp. 15–46). Chichester, UK: Wiley.

Peters, L., Baum, J., & Stephens, G. (2011). Creating ROI in leadership development. *Organizational Dynamics*, *40*, 104–109.

Peterson, T. (2012). Breaking the Mold: New Paradigms for Physician Leadership in Patient Safety. *Physician Executive*, *38*(3), 54–58.

Petriglieri, G., Wood, J. D., & Petriglieri, J. L. (2011). Up close and personal: Building foundations for leaders' development through the personalization of management learning. *Academy of Management Learning & Education*, *10*(3), 430–450.

Pfeffer, J. (2015). *Leadership BS*. New York: HarperCollins.

Pfeffer, J. (2016). Getting beyond the BS of leadership literature. *McKinsey Quarterly*, 1–7.

Pinnington, A. H. (2011). Leadership development: Applying the same leadership theories and development practices to different contexts? *Leadership*, *7*, 335–365.

Pless, N. M., Maak, T., & Stahl, G. K. (2011). Developing Responsible Global Leaders Through International Service-Learning Programs: The Ulysses Experience. *Academy of Management Learning & Education*, *10*(2), 237–260.

Powell, S. K., & Yalcin, S. (2010). Managerial training effectiveness. *Personnel Review*, *39*(2), 227–241.

Pradarelli, J. C., Jaffe, G. A., Lemak, C. H., Mulholland, M. W., & Dimick, J. B. (2016). Designing a leadership development program for surgeons. *Journal of Surgical Research*, *200*(1), 53–58. https://doi.org/10.1016/j.jss.2015.08.002

PRISMA. (2015). PRISMA: Transparent Reporting of Systematic Reviews and Meta-Analyses. Retrieved from http://www.prisma-statement.org/

Quaglieri, P. L., Penney, S. H., & Waldner, J. (2007). Developing future business and civic leaders: the Emerging Leaders Program. *Management Decision*, *45*(10), 1685–1694.

QuintilesIMS Institute. (2017). *Global Oncology Trends 2017* (pp. 1–47). Parsippany, NJ: QuintilesIMS Institute.

Raelin, J. A. (2011). From Leadership-as-Practice to Leaderful Practice. *Leadership*, *7*(2), 195–211.

Ready, D. A., & Conger, J. A. (2003). Why leadership-development efforts fail. *MIT Sloan Management Review*, *44*(3), 83–88.

Redecker, J. (2004). The legal overlay to succession planning. *Employee Relations Law Journal*, *30*(2), 23–30.

Reed, D., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association Between Funding and Quality of Published Medical Education Research. *Journal of the American Medical Association*, *298*(9), 1002–1009.

Richman-Hirsh, W. L. (2001). Posttraining interventions to enhance transfer: the moderating effects of work environments. *Human Resource Development Quarterly*, *12*(2), 105–120.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Science*, *4*, 155–169.

Robinson, V. M. J. (2010). From instructional leadership to leadership capabilities: Empirical findings and methodological challenges. *Leadership and Policy in Schools*, *9*(1), 1–26.

Rose, S. (2015). *Better leadership for tomorrow: NHS leadership review* (pp. 1–68). Department of Health.

Rosenman, E. D., Shandro, J. R., Ilgen, J. S., Harper, A. L., & Fernandez, R. (2014). Leadership Training in Health Care Action Teams: A Systematic Review. *Academic Medicine*, *89*(9), 1295–1306.

Rost, J. C. (1993). *Leadership for the Twenty-First Century*. Westport, CT: Praeger.

Roulston, K. (2010). *Reflective interviewing: A guide to theory & practice*. Thousand Oaks, CA: Sage.

Rousseau, D. M. (2006). Is there such a thing as ‘evidence-based management’? *Academy of Management Review, 31*(2), 256–269.

Rowan, M., & Huston, P. (1997). Qualitative research articles: information for authors and peer reviewers. *Canadian Medical Association Journal*, *157*(10), 1442–1446.

Rowland, D. (2016). Why leadership development isn’t developing leaders. *Harvard Business Review*, 1–5.

Russon, C., & Reinelt, C. (2004). The Results of an Evaluation Scan of 55 Leadership Development Programs. *Journal of Leadership and Organizational Studies, 10*(3), 104–107.

Sahlin-Andersson, K., & Engwell, L. (2002). *The Expansion of Management Knowledge*. Redwood City, CA: Stanford University Press.

Samani, M., & Thomas, R. J. (2016). Your leadership development programme needs an overhaul. *Harvard Business Review*, 1–5.

Sanfey, H., Harris, I., Pollart, S., & Schwartz, A. (2011). Evaluation of the University of Virginia Leadership in Academic Medicine Program. *Teaching and Learning in Medicine: An International Journal, 23*(4), 347–358.

Santos, A., & Stuart, M. (2003). Employee perceptions and their influence on training effectiveness. *Human Resource Management Journal, 13*(1), 27–45.

Sarto, F., & Veronesi, G. (2016). Clinical leadership and hospital performance: assessing the evidence base. *BMC Health Services Research*, *16*(Suppl 2), 169.

Satiani, B., Sena, J., Ruberg, R., & Ellison, E. C. (2014). Talent management and physician leadership training is essential for preparing tomorrow’s physician leaders. *Journal of Vascular Surgery*, *59*(2), 542–546. https://doi.org/10.1016/j.jvs.2013.10.074

Schwandt, T. A. (1998). Constructivist, interpretivist approaches to human inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *The landscape of qualitative research: theories and issues* (pp. 221–259). London: Sage.

Schwandt, T. A. (2000). Three Epistemological Stances for Qualitative Inquiry. Interpretivism, Hermeneutics and Social Constructionism. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (Second). Thousand Oaks, CA: Sage.

Schwartz, R. W., Pogge, C. R., Gillis, S. A., & Holsinger, J. W. (2000). Programs for the Development of Physician Leaders: A Curricular Process in Its Infancy. *Academic Medicine*, *75*(2), 133–140.

Schyns, B., Tymon, A., Kiefer, T., & Kerschreiter, R. (2013). New Ways to Leadership Development: A Picture Paints a Thousand Words. *Management Learning*, *44*(11), 11–24.

Seibert, S. E., Sargent, L. D., Kraimer, M. L., & Kiazad, K. (2017). Linking Developmental Experiences to Leader Effectiveness and Promotability: The Mediating Role of Leadership Self-Efficacy and Mentor Network. *Personnel Psychology*, *70*(2), 357–397. https://doi.org/10.1111/peps.12145

Shah, P., Cross, V., & Sii, F. (2013). Sailing a Safe Ship: Improving Patient Safety by Enhancing the Leadership Skills of New Consultant Specialist Surgeons. *Journal of Continuing Education in the Health Professions*, *33*(3), 190–200.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … the PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*, *349*(g7647), 1–25. https://doi.org/10.1136/bmj.g7647

Shanafelt, T. D., Gorringe, G., Menaker, R., Storz, K. A., Reeves, D., Buskirk, S. J., … Swensen, S. J. (2015). Impact of Organizational Leadership on Physician Burnout and

Satisfaction. *Mayo Clinic Proceedings*, *90*(4), 432–440. https://doi.org/10.1016/j.mayocp.2015.01.012

Sider, L. (2017, March 9). Where can I find information on Yale's 1953 goal study? Retrieved 25 August 2017, from http://ask.library.yale.edu/faq/175224

Simmonds, D., & Tsui, O. (2010). Effective design of a global leadership programme. *Human Resource Development International, 13*(5), 519–540.

Solansky, S. T. (2010). The evaluation of two key leadership development program components: Leadership skills assessment and leadership mentoring. *The Leadership Quarterly, 21*, 675–681.

Sonnino, R. E. (2016). Health care leadership development and training: progress and pitfalls. *Journal of Healthcare Leadership, 8,* 19–29.

Spurgeon, P., Mazelan, P., & Barwell, F. (2011). Medical Engagement: a crucial underpinning to organizational performance. *Health Services Management Research*, *24*(3), 114–120.

Squazzo, J. D. (2009). Cultivating Tomorrow's Leaders: Comprehensive Development Strategies Ensure Continued Success. *Healthcare Executive*, *24*(6), 8–20.

Stake, R. E. (1995). *The art of case study research*. London: Sage.

Steinert, Y., Naismith L, & Mann K. (2012). Faculty development initiatives designed to promote leadership in medical ducation. A BEME systematic review: BEME Guide No. 19. *The International Journal of Medical Technology*, *34*(6), 483–503. https://doi.org/doi: 10.3109/0142159X.2012.680937

Stergiopoulos, V., Maggi, J., & Sockalingam, S. (2009). Teaching the Physician-Manager Role to Psychiatric Residents: Development and Implementation of a Pilot Curriculum. *Academic Psychiatry*, *33*(2), 125–130. https://doi.org/10.1176/appi.ap.33.2.125

Sterne, J. A., Higgins, J. P., & Reeves, B. C. (2014). *A Cochrane Risk of Bias Assessment Tool: for Non-Randomised Studies of Interventions (ACROBAT NRSI)* (Version 1.0.0). Cochrane Colloquium. Retrieved from http://www.riskofbias.info

Stevens, A., Shamseer, L., Weinstein, E., Yazdi, F., Turner, L., Thielman, J., … Moher, D. (2014). Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *BMJ Open*, *348*(g3804), 1–29. https://doi.org/https://doi.org/10.1136/bmj.g3804

Stevens, J. P. (2012). *Applied Multivariate Statistics for the Social Sciences* (Fifth). New York: Routledge.

Stewart, J.-A. (2009). Evaluation of an action learning programme for leadership development of SME leaders in the UK. *Action Learning: Research and Practice*, *6*(2), 131–148.

Stoller, J. K. (2009). Developing Physician-Leaders: A Call to Action. *Journal of General Internal Medicine*, *24*(7), 876–878.

Stoller, J. K., Goodall, A. H., & Baker, A. (2016). Why the best hospitals are managed by doctors. *Harvard Business Review*, 1–6.

Strasser, D. C., Falconer, J. A., Stevens, A. B., Uomoto, J. M., Herrin, J., Bowen, S. E., & Burridge, A. B. (2008). Team training and stroke rehabilitation outcomes: a cluster randomized trial. *Archives of Physical Medicine & Rehabilitation*, *84*, 218–222.

Straus, S. E., Soobiah, C., & Levinson, W. (2013). The Impact of Leadership Training Programs on Physicians in Academic Medical Centers: A Systematic Review. *Academic Medicine*, *88*(5), 1–15.

Sullivan, G. M. (2011). Getting Off the 'Gold Standard': Randomized Controlled Trials and Education Research. *Journal of Graduate Medical Education*, *3*(3), 285–289. https://doi.org/10.4300/JGME-D-11-00147.1

Suutari, V., & Viitala, R. (2008). Management Development of Senior Executives Methods and Their Effectiveness. *Personnel Review*, *37*(4), 375–392.

Swanwick, T., & McKimm, J. (2012). Clinical leadership development requires system-wide interventions, not just courses. *The Clinical Teacher*, *9,* 89–93.

Tao, K., Li, X., Zhou, Q., Moher, D., Ling, C., & Yu, W. (2011). From QUOROM to PRISMA: A Survey of High-Impact Medical Journals' Instructions to Authors and a Review of Systematic Reviews in Anesthesia Literature. *PLoS One*, *6*(e27611). https://doi.org/https://doi.org/10.1371/journal.pone.0027611

Taylor, B. (2010). *Effective Medical Leadership*. Toronto: University of Toronto Press.

Taylor, N., Clay-Williams, R., Hogden, E., Braithwaite, J., & Groene, O. (2015). High performing hospitals: a qualitative systematic review of associated factors and practical strategies for improvement. *BMC Health Services Research*, *15*(244), 1–22. https://doi.org/10.1186/s12913-015-0879-z

Ten Have, E. C. M., Nap, R. E., & Tulleken, J. E. (2013). Quality improvement of interdisciplinary rounds by leadership training based on essential quality indicators of the Interdisciplinary Rounds Assessment Scale. *Intensive Care Medicine*, *39*(10), 1800–1807. https://doi.org/10.1007/s00134-013-3002-0

The Mid Staffordshire NHS Foundation Trust. (2013). *Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry* (No. HC 898-I). London.

Thomas, R. J., Jules, C., & Light, D. A. (2012). Making Leadership Development Stick. *Organizational Dynamics*, *41*, 72–77.

UK EQUATOR Centre. (n.d.). Reporting guidelines. Retrieved from https://www.equator-network.org/?post_type=eq_guidelines&eq_guidelines_study_design=systematic-reviews-and-meta-analyses&eq_guidelines_clinical_specialty=0&eq_guidelines_report_section=0&s=

Van Aerde, J. (2013). *Physician Leadership Development* (pp. 1–101). Alberta Health Services.

Van de Voorde, K., Paauwe, J., & Van Veldhoven, M. (2012). Employee well-being and the HRM-organizational performance relationship: a review of quantitative studies. *International Journal of Management Reviews*, *14*(4), 391–207.

Vardiman, P. D., Houghton, J. D., & Jinkerson, D. L. (2006). Environmental leadership development: Toward a contextual model of leader selection and effectiveness. *Leadership & Organization Development Journal*, *27*(1/2), 93–105.

Vimr, M., & Dickens, P. (2013). Building physician capacity for transformational leadership—Revisited. *Healthcare Management Forum*, *26*(1), 16–19. https://doi.org/10.1016/j.hcmf.2013.01.003

Vygotsky, L. S. (1978). *Mind in Society: The development of higher mental process*. Cambridge, MA: Harvard University Press.

Watkins, K. E., Lysø, I. H., & deMarrais, K. (2011). Evaluating executive leadership programmes: A theory of change approach. *Advances in Developing Human Resources*, *13*(2), 208– 239.

Watkins, M. (2003). *The first 90 days*. Boston, MA: Harvard Business School Press.

Weaver, S. J., Dy, S. M., & Rosen, M. A. (2014). Team-training in healthcare: a narrative synthesis of the literature. *British Medical Journal Quality and Safety*, *23*, 359–372.

Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2016, February 11). The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Retrieved from http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

Woodward, S. C. (2016, October). *The DNA of a Champion*. Churchill College, Cambridge.

World Medical Association. (2013). Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*, *310*, 2191–2194.

Xirasagar, S., Samuels, M., & Stoskopf, C. (2005). Physician leadership styles and effectiveness. *Medical Care Research and Reviews*, *62*(6), 720–740.

Yeo, R. K. (2007). Problem-based learning: a viable approach in leadership development? *Journal of Management Development, 26*(9), 874–894.

Yin, R. (2003). *Case study research: design and methods* (3rd ed.). London: Sage.

Yukl, G. A. (2010). *Leadership in Organizations*. New Jersey: Prentice Hall.

Zenger, J. H., & Folkman, J. (2003). Developing leaders. *Executive Excellence*, *20*(9), 5.

Zhang, J. (1999). *Effects of management training on trainees learning, job performance, and organization results: a meta-analysis of evaluation studies from 1983 1997* (Unpublished doctoral dissertation). Oklahoma State University.

Zook, C., & Allen, J. (2001). *Profit from the Core*. Cambridge, MA: Harvard Business Publishing.

9 APPENDIX

Appendix A: Background to Medical Leadership

Introduction: A Turning Point

The Canadian Royal College of Physicians claims that the medical profession is at a turning point in its history due principally to two interrelated concepts: quality control and leadership (B. Taylor, 2010). Health systems around the world are facing unprecedented, rising demands and pressure to offer better quality care with escalating costs and tightening budgets (CMO Clinical Advisor Alumni, 2012; Mountford & Webb, 2009; Rose, 2015; B. Taylor, 2010). Healthcare is becoming more and more expensive because of the increasing costs of medication and the fact that on average people are living longer and requiring extended, more complex care. In response to changing demands in the field, medical leadership is undergoing a paradigm shift from the traditional role that physicians have played to more collaborative leadership and administration (Sonnino, 2016). Until the past two or three decades, physicians did not receive any dedicated leadership or management development (Clark & Armit, 2010; Edmonstone, 2009; Ireri et al., 2011); they were considered leaders automatically by virtue of their profession. For reasons to be explained below, formal leadership development interventions for doctors are now becoming more widespread and considered by many to be increasingly important (Sonnino, 2016).

Regarding how the tension between care and cost has affected the healthcare situation in the UK, for example, the Rose Report (2015) claims that it has reached a "critical leadership tipping point" (p. 45). The answer, Rose (2015) says, is not more management but better leadership. This is not wholly unexpected; the National Health Service (NHS) has published a number of reports in the last several decades calling for improvement and change, which relates to leadership directly. These reports include the Darzi Report (2008), which declares a need for improved patient safety and quality of care, financial performance, and performance efficiency. Rose (2015) says that none of these recommended changes have been supported by the deliberate development of the skills needed to implement them. What is most needed, he says, is to focus on developing leaders across the NHS (Rose, 2015). Martins (2010) states that currently very few medical schools in the world actually include any kind of management and leadership in their curricula for students or for their faculty, thus, although more leadership development programmes are emerging, this remains an enterprise in its adolescence. Challenges associated with the tension described above and the urgency attached to overcoming them are not unique to the UK, nor is the rising number of leadership development programmes for physicians and healthcare professionals. The big issue is that the evidence base to inform and support these initiatives to make them optimal and guidance on how to measure leadership afterward are thin and, on the whole, very poorly reported.

Traditional Physician Leadership

As intimated previously, the evolving nature of medical leadership and the ensuing development initiatives spring from a long history of traditional roots. Historically, physicians have operated from a very hierarchical and autocratic position from which a doctor "dictates orders from the top of a management pyramid, with other caregivers carrying out those orders" (B. Taylor, 2010, p. 55). Stoller et al. (2016) describe traditional physician training as producing "heroic lone healers" who are "collaboratively challenged" and operate in command and control environments (p. 4). Generally, no one made clinical decisions without a physician's consent (Lee, 2010) because they were ultimately responsible for the patient; thus, there was a deeply embedded notion that physician autonomy was crucial to quality healthcare. Physicians also traditionally viewed empowering non-physician healthcare workers as a loss of control and functioned as what Maccoby et al. (2013) call "productive narcissists". Stoller (2009) says that this can be traced back to the scientific nature of medical school training, which typically values autonomous decision-making and personal achievement, resulting in doctors being often disinclined to collaborate or follow. Van Aerde (2013) and Bohmer (2012) suggest that this attitude continues to be true; however, they assert that individual physicians' excellence is no longer adequate enough to guarantee good patient outcomes. Taylor (2010) adds that because doctors are well educated, independent practitioners in a respected profession, they invariably feel a sense of entitlement, which creates further potential for conflict with colleagues. Thus, historically, doctors' authority was considered inherent and their approach was autocratic, stemming from the notion that they were ultimately accountable for the patients' care, which often resulted in a divide between them and other non-physician colleagues.

Medicine Versus Administration: the "We-They" Divide

The aforementioned dynamics of doctors' traditional self-conception and modus operandi often forged a "we-they" divide against non-physician administrators. In this case, physicians considered their relationship with patients vital to healthcare; whereas they viewed managers as prioritising budgetary concerns (B. Taylor, 2010). Rose (2015) describes a widespread, deep-rooted perception among doctors and nurses that management is "the dark side" (Goodall & Stoller, 2017), positioning themselves in opposition to this nefarious entity. An example of this is Davies and Harrison's (2003) survey of administrative and clinical healthcare leaders, which concluded that only 24 per cent of clinical directors believed that management was driven more by clinical rather than financial priorities. This negative perception intensified the closer respondents were to the front lines of patient care. Jones, McCay, and Keogh (2011) assert that typically, medical accountability was quite siloed: doctors made medical decisions, administrators made administrative decisions, nurses made nursing decisions, and central government made the funding decisions. These divisions were not only implicit; until recently, most hospitals had two organisational charts: one for hospital operations and one for medical staff (B. Taylor, 2010). Although theoretically, all healthcare professionals, whether physicians, nurses, or non-clinical administrators, have the same goal of providing effective patient-centred care (Taylor, 2010), the common physician approach to achieving this goal often generated the potential for professional clashes with colleagues.

As alluded to earlier, these differences in approaches are partly attributable to substantive features and differences in the training between doctors and non-physician administrators. Physician education has a clinical focus on treating individual patients one at a time, which McAlearney et al. (2005) suggest is in sharp contrast to the broad, institution-level focus on the organisation or the overall healthcare system that administrators require. Likewise, in terms of decision-making concerning patients, doctors are used to having the final say, whereas administrators often must build consensus among multiple divergent stakeholders. Furthermore, selection and compensation based on management and leadership tend to differ significantly between the two groups. In addition to relevant graduate degrees, such as an MBA, and a wealth of practical experience, administrators require proven exemplary managerial and leadership capabilities to be hired and are rewarded on these bases (McAlearney et al., 2005). Except perhaps at the most senior levels, doctors are often promoted based on clinical, teaching, or research accomplishments (M. Day, 2007; McAlearney et al., 2005) and receive no compensation for leadership. Other reasons for physician promotion to leadership roles include seniority, or simply by being the only one willing to take on the job.

Disincentives for Physician Leaders

In addition to the imbedded dichotomy between physicians and administrators and the suspicion of the motivation that the former have of the latter (Davies & Harrison, 2003), many factors can operate as deterrents for clinicians who might otherwise be willing to assume leadership roles. The first is a loss of or decrease in clinical time, which most physicians are reluctant to give up, since treating patients is why most became doctors in the first place. Other disincentives include a resulting sense of disenfranchisement (Edwards, Komacki, & Silversin, 2002) or a decrease in professional recognition and status. This is because those who give up clinical practice in favour of administrative positions can quickly be seen by their physician colleagues as losing or having lost touch with the realities of practical medicine. Leadership roles can also bring financial loss for doctors (Jones et al., 2011; Mountford & Webb, 2009; B. Taylor, 2010). In medical systems like Canada's, for example, where the bulk of physicians' salaries comes from patient consultations, taking on administrative roles (which take time away from clinical practice) actually *decrease* a doctor's salary because of the opportunity cost of missed clinical time (Taylor, 2010). Finally, junior doctors consistently report a lack of skills and training in management and leadership skills and cite a lack of exposure to administrative career pathways, opportunities to develop leadership skills, or formal support (Bohmer, 2012). This lack of preparation and support could explain why many medical leaders describe themselves as "accidental leaders" who stumbled into their roles without any formal training (Blumenthal, Bernard, Bohnen, & Bohmer, 2012). Taylor says that this results in clinicians being promoted to administrative leadership positions "for which they have never been trained, for which they may have little inclination, and for which they may be entirely unsuited" (B. Taylor, 2010, p. 39). For reasons about to be explained, the traditional medical leadership model of the independent, autocratic physician practitioners with little or no formal leadership training is no longer feasible given the changing nature of the field.

Reasons for Change: UK Reports

In the UK, there has been a gradual trend towards strategic medical leadership, highlighted by many national reports. (For a detailed elucidation of the history of medical leadership in the UK, see Spurgeon, Mazelan, and Barwell (2011)). Following the creation of the NHS in 1948, these include the 1967 Cogwheel Report, the Griffiths Report in 1983, the Darzi Report of 2008, and the Rose Report of 2015. One of the overarching themes among them is a reiteration of the need for doctors to be involved in leading and managing their medical practices. The emergence of this emphasis was not linear in its progression; however, and the pathway to the present situation was marked by periods of diminished clinical leadership. For example, Bohmer (2012) notes that many doctor practitioners interpreted the Griffiths Report's call for greater management as promoting non-clinical management. The rise in interest in medical leadership in the UK mirrors that in many other Western countries that are all facing similar challenges. The current situation in medical care demands a new kind of physician leader for three main reasons: medical technological developments and heightened budget concerns, patient safety, and doctors assuming senior leadership roles.

Reason #1: Increase in Technology, Healthcare Teams, and Budgetary Restrictions

The first major set of changes related to medical leadership includes advancing technology, an increase in the prevalence of healthcare teams, and augmenting budgetary constraints. Lee (2010) suggests that the adoption of new medical technologies often requires specially trained personnel, which precludes doctors from operating as independently as they had previously. In many cases, healthcare now is delivered on a multi-disciplinary continuum involving clinical teams that require complex coordination with doctors cooperating with a host of other healthcare practitioners, including other doctors, surgeons, pharmacists, nurses, technicians, and therapists (B. Taylor, 2010). McKimm and Phllips (2009) suggest that the current UK government policy is to promote integrated services and align those that were previously siloed. Clinical teams regularly form and dissolve rapidly and thus physician leaders must be able to quickly adapt to new situations and new team members who may not have worked together previously (Dine et al., 2011). Physicians also balance clinical responsibilities with those in the often-overlapping spheres of academic faculty at teaching hospitals and hospital administration. Hartley and Benington (2010) suggest that these new advancements rely as much on motivating staff and working within and across teams as on the techniques themselves. Thus, the developments in terms of delivering complex care with increasingly advanced technology requires leadership and teamwork that renders the aforementioned traditional hierarchical approach insufficient (Van Aerde, 2013).

In addition to inter-site complexity, healthcare centres are increasingly working with partner organisations, which requires further coordination (Hartley & Benington, 2010). Similarly, healthcare is expanding further to include non-traditional "complementary medical" providers such as chiropractors, massage therapists, and naturopaths who are lobbying to be involved in a more official capacity. In many places, these practitioners are asking to be able to order lab tests, prescribe drugs, and benefit from government funding in a similar way to medical doctors. This development broadens the pool of people involved in healthcare and makes effective coordination even more necessary. Some commentators suggest that this manifold expansion has led to confusion about the structure and role of individuals and teams in many organisations (Jong, Könings, & Czabanowska, 2014). Taylor (2010) warns that this ambiguity can lead to fragmentation and disorganisation which result in "redundant care and errors that raise costs and threaten quality" (p. 52). Another relevant factor is that patients are often better-informed because of the internet, which adds an additional level of intricacy to the conversation. Providing excellent medical services given these dynamics requires proficiency with management and leadership skills and approaches with which most doctors are unfamiliar (Bohmer, 2012). Lee (2010) argues that physicians who attempt to conduct themselves in the traditional style of independent, unilateral decision-makers cannot possibly provide the same quality care as healthcare teams.

Another related impetus for the changing role of physician leaders is that medical progress such as new drugs, tests, devices, and procedures, has caused hospital costs to rise inexorably (Catlin, Cowan, Heffler, & Washington, 2007; Cutler, 2004; Lee, 2010). Taylor (2010) says these developments create an “absolute need for improved efficiency” (p. 22) and greater fiscal responsibility, since there is massive pressure to improve the quality of care while at the same time managing budgets strictly. Doctors are central players and Xirasagar et al. (2005) suggest that they are the key to cost control and quality improvement. Also, health goals are increasingly targeted at predicting and preventing diseases within the whole population, rather than responding and treating individuals, whilst facing increasingly finite resources (Hartley & Benington, 2010; McKimm & Swanwick, 2011). Good leadership can improve quality and decrease costs by avoiding unnecessary tests and care (Maccoby et al., 2013). Bohmer (2012) adds that effective physician leaders understand the issues on both sides: the medical science and the organisational imperatives in terms of what is possible, feasible, and affordable and they are at the forefront of purchasing care. Therefore, as technology and the number of people involved in healthcare teams increase and the pressure to provide excellent care with greater financial efficiency heightens, the need for physicians to be competent in leadership is intensified.

Reason #2: Patient Safety and Creating Change: Two Examples

A second major reason for doctors to be involved in leadership was sparked by a 1999 report by the U.S. Institute of Medicine (IOM) entitled “To Err Is Human”. The report suggests that as many as 98,000 people die annually in the United States alone as a direct result of *preventable* medical errors in hospitals. Furthermore, data suggest that at least half of all surgical complications are avoidable (Haynes et al., 2009). These errors include improper transfusions, instruments or foreign bodies left in patients after surgery, wrong-site surgery, and mistaken patient identities. The report adds that this number exceeds the annual number of deaths caused by car accidents, breast cancer, and AIDS combined. It also reveals that these errors carry with them costs of between $17 billion and $29 billion in hospitals per year because of these so-called X factors. More than a decade later, the U.S. Department of Health and Human Services found still alarming numbers of medical errors (Peterson, 2012). IOM and Taylor (2010) do not blame individual doctors for the aforementioned preventable errors, but attribute many to what the IOM says is the decentralised and fragmented nature of the healthcare system. They argue that these errors can be prevented by designing processes within the health system that make it harder to do something wrong and easier to do something right (Maccoby et al., 2013). Taylor (2010) explains that alongside this development, the public's relationship with the medical profession has changed. Formerly, people accepted that physicians did their best and that sometimes unfortunate things simply happened; whereas now, there is less automatic deference towards medical authority and increased public awareness. Many believe that doctors and hospitals should be held accountable and liable for mistakes. These heightened expectations coincide with the amount of health information available online and a stronger desire for personalised and flexible care (Hartley & Benington, 2010). The data on preventable medical errors is one major factor that has sparked the global recognition of the need for greater leadership in healthcare.

The strong emphasis on quality control in Canada, the United States, and the UK is intimately linked to leadership. Patient care is becoming more standardised and leadership is viewed as playing a key part in generating consensus and compliance in terms of protocol. For example, Haynes et al. (2009) found that introducing a surgical checklist based on the WHO model in eight countries resulted in a startling 36 per cent drop in postoperative complications on average. Mortality rates also declined by a similar percentage, which equated to significantly fewer people dying *because* they were in hospital (Haynes et al., 2009). Even with such substantial and robust data, many hospitals, doctors, and surgeons are categorically averse to streamlining. Many physicians resist efforts to standardise processes and reduce variation because they term it "cookbook medicine" (Maccoby et al., 2013) and are reluctant to change their current practice. Haynes et al. (2009) also explain that the mechanism of improvement is not simple or unifactorial; implementing the checklist required systems and the behaviour of individual surgical teams to change. This was a challenge despite the fact that the items on the aforementioned 19-item checklist were basic, such as, "Confirms the patient's identity, surgical site, and procedure" and study participants reported that it only took roughly 90 seconds to complete each time.

A second related example comes from a 2005 report by the House of Commons Health Committee, which stated that between 25,000 and 32,000 people were dying in England annually from *preventable* blood clots contracted while in hospital, which also exceeds the number of deaths per year from car accidents, AIDS, and breast cancer *combined* (House of Commons Health Committee, 2005). Furthermore, the report estimates that the total cost to the UK of managing these clots was £640 million, compared to 60 – 80 per cent less projected to be incurred through effective prevention (thromboprophylaxis) (House of Commons Health Committee, 2005). This means that from a strictly financial perspective, prevention versus treatment of clots would result in a net savings of up to £512 million a year. This area was targeted because blood clots were considered the most common preventable cause of hospital death and the number-one strategy to improve patient safety in hospitals (W. H. Geerts et al., 2008). Despite these striking figures, a complex leadership programme was required to implement prevention guidelines nationally and, as importantly, to convince trusts and healthcare staff to follow them consistently and accurately (Lester et al., 2013). Maccoby et al. (2013) explain that "no government policy could by itself, cause healthcare organisations to improve quality and at the same time cut costs. To do so would require good leadership" (p. xxiv). Taylor (2010) argues that every medical leader is responsible for ensuring that all practitioners are on side and championing the processes of better care. Likewise, Peterson (2012) asserts that physicians are often the key to supporting the organisational culture that facilitates these ongoing quality and patient safety improvements. Therefore, even in the face of scientific evidence, credible reports, or authoritative guidelines, leadership is required to implement successful patient safety and quality improvements and inspire compliance among staff.

Reason #3: Leadership Roles

The final reason for the change in healthcare leadership dynamics is that beyond leading teams and departments, physicians are increasingly taking on senior administrative leadership roles (Taylor, 2010), such as the head of Human Resources or CEO. The NHS and other healthcare organisations are becoming progressively interested in having doctors in these roles, which Ireri et al. (2011) suggest has strong support in literature. The NHS's chief executive, David Nicholson, announced in 2007: "Within two years, we want a doctor applying for every chief executive post advertised … where clinicians and managers work together" (M. Day, 2007, p. 230). Although there are at present examples of doctors engaged in senior healthcare management, in the UK and the United States, there are very few doctors in chief executive positions (Falcone & Santiani, 2008; Horton, 2008; Ireri et al., 2011). Some question whether doctors *should* be in these positions (Dwyer, 2010); however, there is growing evidence that it leads to better outcomes on many levels.

Connecting Effective Leadership, Hospital Performance, and Patient Outcomes

There is a popular perception that effective leadership in healthcare is key to high-performing systems and improved outcomes (Baker, n.d.; CMO Clinical Advisor Alumni, 2012; Davies & Harrison, 2003; Ham & Dickenson, 2008). This belief is not restricted to top executives. Clinical leadership is described as "the core business of everyday medical care and public health" (CMO Clinical Advisor Alumni, 2012, p. 1847), critical to staff engagement, improved clinical, financial, and operational performance, and the delivery of high-quality, cost-effective care (CMO Clinical Advisor Alumni, 2012; Dine et al., 2011; Edmonstone, 2011; Jeon et al., 2013; Kim & Thompson, 2012; Millward & Bryan, 2005; Squazzo, 2009). Bruce Barraclough, Clinical Lead and Chair of the WHO Patient Safety Curriculum Guide, agrees, and writes that effective leadership is the essential ingredient necessary to secure the resources, improve quality, address risks, and provide the safest and best possible care in the complex modern healthcare environment (in B. Taylor, 2010). Physician leadership is also said to be pivotal in providing seamless care across professional, organisational, and geographical boundaries (Edmonstone, 2011). As mentioned previously, government policies alone are not enough to improve healthcare; leadership is crucial (Maccoby et al., 2013). Rather than lumping the onus exclusively on executives, Taylor (2010) argues that every medical leader is responsible for ensuring that all practitioners are on side, championing the processes of providing better care. Peterson (2012) reinforces this point, asserting that physicians are often the key to supporting the organisational culture that facilitates these ongoing quality and patient safety improvements. Therefore, at the individual, organisational, systemic, and national levels, many consider leadership to be essential to improving patient safety and care, while operating with increasing budgetary restrictions.

Beyond testimonials, there are several studies and systematic reviews that connect effective leadership with patient outcomes and high performing hospitals (Baker, n.d.; CMO Clinical Advisor Alumni, 2012; Davies & Harrison, 2003; Ham & Dickenson, 2008). For example, the Hockney and Bates (2010) review compared hospitals in the U.S. with consistent high performance over two years to those at the bottom deciles (low performing) based on internal medicine outcome measures. The findings of their semi-structured interviews with internists from seven different sites revealed that "leadership characteristics" was identified as one of the key differentiating features in high-performing hospitals (Hockey & Bates, 2010). Equally, low performing sites reported having transient leadership and disconnected access between frontline physicians and senior leaders (Hockey & Bates, 2010). Bloom et al. (2014) also found a positive link between effective leadership in UK hospitals and improved outcomes such as survival rates from general surgery, lower staff turnover, and shorter patient length of stay in hospital. Spurgeon et al. (2011) compared UK hospitals' Quality Commission ratings of patient mortality, error rates and patient care, and 100 per cent compliance with levels of service provision. They analysed differences among the top ten and bottom ten hospitals on the Medical Engagement Scale (MES) (a reliable and valid psychometric instrument). The authors claim that there was a statistically significant correlation between sites where doctors are more engaged with maintaining and enhancing the performance and outstanding clinical and financial outcomes (Spurgeon et al., 2011). They identified factors including continuity of leadership at the executive level and an explicit strategy at improving engagement with clinical staff as features of the top ten hospitals (Spurgeon et al., 2011). Interviews with key managers and senior clinicians, including the Chief Executive and the Medical Director, in the Mannion et al. (2005) study revealed that a key point of divergence between high and low-performing hospitals in England according to the NHS star system was leadership and management orientation. The authors conclude that there was a strong relationship between hospital leadership and hospital performance. From a systemic point of view, Clay-Williams et al. (2014) assert that effective leadership and clinical champions were among the common features of successful large-scale hospital and system-wide interventions to improve patient outcomes. Bloom et al. (2014) also suggest that financial performance is significantly better in U.S. hospitals with a higher management score. Finally, the systematic review by Taylor et al. (2015) identified effective leadership across the organisation as one of the key themes characterising high-performing hospitals in three different continents.

These make it unsurprising that effective leadership is thought to be central to implementing the NHS reforms (Jones et al., 2011; Rose, 2015), which explains why physician leadership was prioritised in the 2008 NHS review (Darzi, 2008; Horton, 2008) and made the focus of the Rose Report in 2015. With these points in mind, Straus, Soobiah, and Levinson (2013) conclude that the role of the physician leader simply cannot be overemphasised and Taylor (2010) adds that effective clinical leadership is necessary to guarantee the "well-being of the profession and certainly that of the patient" (p. 3). Therefore, there are several examples of high performing hospitals and improved patient and financial outcomes being correlated with reports of effective leadership, which reinforce the connection between the two made in the previous paragraphs.

Medical Leadership at All Levels

As mentioned throughout, leadership is becoming widely recognised increasingly as an essential skill for all physicians (Mountford & Webb, 2009; Pradarelli et al., 2016; Rose, 2015; Satiani et al., 2014; Straus et al., 2013). Bohmer (2012) suggests that working doctors exercise the most influence over the key processes and microsystems necessary to improve the overall health system performance significantly, as well as medical outcomes, such as error rates, and terminal outcomes, such as readmission and mortality rates. Likewise, Mountford and Webb (2009) suggest that frontline clinicians are vital to realising organisations' visions by improving services on a daily basis. Van Aerde (2013) and others advocate physician leadership programmes at all levels (Rose, 2015), yet leadership training is decidedly lacking in today's medical schools and residency programs (Martins, 2010; Pradarelli et al., 2016). MacPahail et al. (2015) extend the need for effective leadership to other healthcare professionals, stating that transforming healthcare systems to improve patient safety and quality of care requires engagement and leadership on the part of all clinical staff (Rose, 2015). A McKinsey report asserts that the most successful healthcare organisations treat all employees as potential leaders, especially clinicians (Mountford & Webb, 2009). More on interdisciplinary leadership development is discussed in chapters five, six, and seven.

Evidence of the Effectiveness of Physician Leaders at the Highest Levels

There are further claims that it can be beneficial to have physicians in the most senior leadership roles. For example, the paper by Bloom et al. (2014) claims that hospital level management scores are strongly correlated with standard hospital outcomes such as heart attack survival rates in all nine countries in the four continents that they studied. The first characteristic that they discovered was associated with higher management quality is the higher percentage of managers with a medical degree. Similarly, a study by McKinsey and the London School of Economics (LSE) analysed 126 hospitals across the UK by interviewing general managers and heads of clinical departments regarding whether and how they had implemented proven management practices in hospitals. The researchers then grouped responses into good, average, and poor practices for 27 behaviour dimensions and assigned an overall management score out of five to each hospital. They report two interesting findings: the first is that better target setting, talent management, and business leadership by doctors was correlated with lower rates of infection in hospitals and of readmission, more satisfied patients and more productive staffs, and higher financial margins (Castro, Dorgan, & Richardson, 2008). Second, they found that hospitals with general managers with a clinical degree performed 50 per cent higher on drivers of performance compared with hospitals with non-clinical leaders (Castro et al., 2008). They conclude that clearly defined roles for doctors in the running of hospitals, as well as appropriate skills, might drive best-practice management generally.

As a final example, Goodall (2011) did an extensive study involving the top-100 U.S. hospitals in 2009, as identified by the US News and World Report's "Best Hospitals", which used the Index of Hospital Quality (IHQ) to measure the performance of nearly 5,000 hospitals across 16 specialties. Goodall analysed the hospital ratings involving three specialties: cancer, digestive disorders, and heart surgery, along with whether the CEOs were physicians or not. Goodall (2011) reports a statistically significant correlation between outstanding hospital performance and physician CEOs, though the overall percentage of physician CEOs in hospitals is quite low (Darzi, 2008; Gunderman & Kanter, 2009; Halligan, 2008). Of particular note, of the 21 highest-ranked hospitals, 16 of the CEOs were physicians (Goodall, 2011). In a related article, Stoller, Goodall, and Baker (2016) conclude that the inverse scenario is also true: the separation of clinical and managerial knowledge inside hospitals has been associated with worse management. While Goodall cautions that this does not prove that doctors make better leaders than non-physicians, others are confident that placing physicians in leadership positions can result in improved hospital performance and patient care (Candace & Giordano, 2009; Darzi, 2008; Dwyer, 2010; Falcone & Santiani, 2008; Halligan, 2008; Horton, 2008; Stoller, 2009). Thus, although the evidence does not yet confirm that all CEOs of hospitals should be doctors, it does indicate that healthcare organisations can flourish with physicians at the helm and the implicit implication is that physicians in senior leadership roles should have the training and support they need to be successful.

The Rationale for Senior Physician Leaders

Many reasons have been offered for why having physicians in senior leadership roles in healthcare institutions can be beneficial. Falcone and Satiani (2008) suggest that physician leaders with years as clinicians are seen as having acquired peer-to-peer credibility (Stoller et al., 2016), which Bloom et al. (2014) add can lead to colleagues having more trust in them and eliminate the “we-they” mentality (Goodall, 2011). As a nuance to this point, when physicians stop seeing patients altogether in favour of full-time administrative duties, they can be seen by other doctors as fading out of touch with the realities of day-to-day physician life. More work needs to be done comparing the outcomes of physician leaders who are still practicing clinicians versus full-time administrators. Regardless, the duty of care that physicians have for patients as clinicians extends even when they take on administrative roles and give up their clinical practice. Goodall (2011) and Bloom et al. (2014) argue that in addition to their credibility, doctor leaders have acquired a deep intuitive knowledge of the medical side of their organisations, which helps with communicating with and managing clinical staff, decision-making, and developing and implementing institutional strategy (Castro et al., 2008; Mountford & Webb, 2009). Jones, McCay, and Keogh (2011) state that clinicians are the “principal drivers of healthcare, with a unique insight into and expertise in healthcare need, challenges, and delivery” (p. 1). Another advantage of having doctors at the highest levels of trusts and hospitals is that they best understand the aforementioned inherent tension between cost and patient welfare and can anticipate the potential impact of policy changes (Bohmer, 2012). Bloom et al. (2014) add that physician leaders may be able to better able to overrule powerful incumbents who object to needed re-organisations. Finally, on a political level, doctors are also in a position to guide politicians to keep health delivery and funding structures focused on patient well-being (Darzi, 2008). Therefore, advocates of physicians in senior leadership roles cite the credibility that goes along with their clinical experience and knowledge and can serve to guide their decision making, while keep the patient as the foremost priority.

“Expert” Physician Leaders

There is good evidence that the best results come not only with physician leaders at the helm of healthcare organisations, but when these physicians also have outstanding clinical achievements (Stoller et al., 2016). Stoller et al. assert that this is an example of what they call “expert [in the core business of the organisation] leaders” (Stoller et al., 2016) and have reported similar findings in other professional domains. These include exemplary academics leading universities and NBA basketball coaches who were former all-stars significantly outperforming coaches who had been players but not all-stars and coaches who had never played professionally. One reason that they offer for this is that when managers have direct experience of what is required to perform a job at the highest standard, they appear more credible, may be more likely to offer the right resources and foster a positive work environment for success, establish appropriate goals, accurately evaluate others’ performance, and make excellent hiring decisions, since they know firsthand what “great” looks like (Stoller et al., 2016). Goodall and Stoller (Goodall & Stoller, 2017) also suggest that expert leaders provide better feedback, allow more autonomy, and attract the best applicants. Stoller et al. (2016) also add that employees who have a boss who is an expert in the core business tend to have high levels of job satisfaction and low intentions of quitting and can thereby contribute to enhancing individual and organisational performance (Edmans, 2011; Keller, Julian, & Kedia, n.d.; A. Oswald et al., 2015; Patterson, Warr, & West, 77). As well, organisations that select expert leaders, such as an outstanding physician executive, Stoller et al. (2016) say, may send a positive message about organisational priorities, which in healthcare is about putting patients first. These leaders may also be best suited to be able to identify other top candidates when hiring and may assist in attracting talented medical personnel to their hospitals because of their professional status (Stoller et al., 2016). Therefore, there are many reasons why having an outstanding physician leader is said to contribute positively to the workplace environment, recruiting, and perceived organisational priorities.

As mentioned earlier, the number of physicians serving in a formal leadership capacity has grown in response to an imminent need in the field (Kuo et al., 2010; Straus et al., 2013). The UK's General Medical Council (2009) states that leadership is an expectation of all doctors and advocates that leadership competence must be an integral part of every doctor's training and learning. Castro et al. (2008) suggest that developing clinicians' leadership should improve the overall management of hospitals and ultimately, the quality and productivity of the healthcare they provide. Many assert that given the current state of healthcare, there is an urgent need for "clinically trained administrators who govern the human and financial resources within healthcare organisations" (Murdock & Brammer, 2011, p. 52). Despite many doctors' aversion to lessening or abandoning their clinical practices, Robert Naylor, chief executive of University College London Hospitals NHS Foundation Trust, asserts that many physicians can have a greater influence on the health of individuals by adopting leadership roles than by treating individual patients (M. Day, 2007). The belief in the importance of physicians joining the administrative ranks has given birth to the profession of Medical Management and the establishment of formal medical specialties such as Medical Managers in the United Kingdom (Dwyer, 2010). One of the main drivers for this was the Francis Report (2013) following the inquiry into the hundreds of patients who died as a result of poor care in Stafford hospital (UK). Francis's (2013) recommendations focused heavily on leadership and management and concluded that "*healthcare management and leadership is, or should be treated as a profession* (p. 241). Given the current nature of Western healthcare systems, physicians are now seen as requiring leadership and management capabilities, especially those who assume formal administrative roles in organisations. It seems, however, that the need for physician-executives far overshadows the scarcity of individuals who are skilled in both clinical medicine and management (Ackerly et al., 2011). If more physicians are indeed to rise to senior administrative positions, it is imperative that they have the appropriate training and support to be successful (McKimm & Swanwick, 2011; Stoller et al., 2016; Straus et al., 2013).

The Lack of Skills and Training

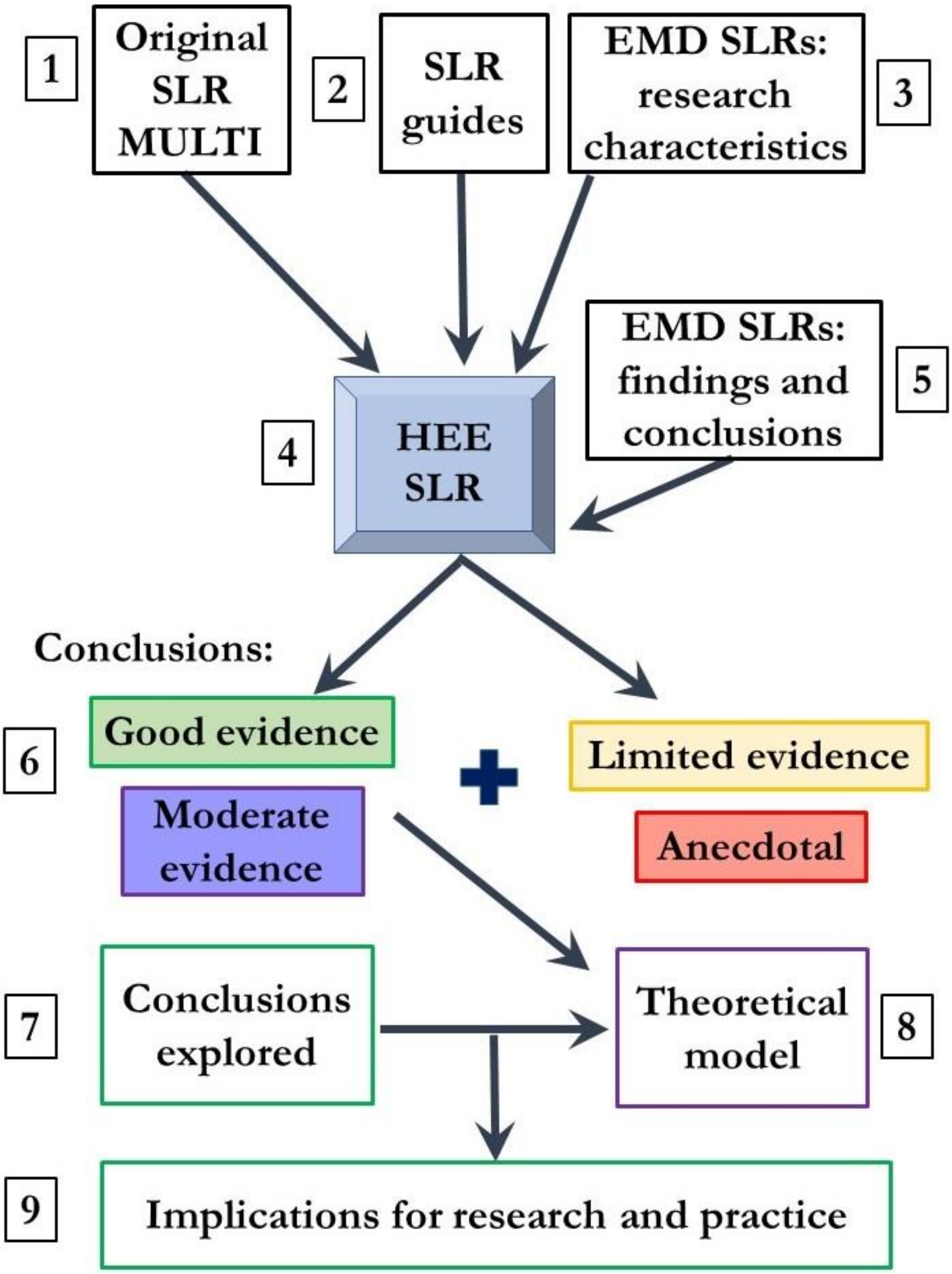
Despite the recurring reiterations of the need for developing physician leaders, calling it "essential" (Castro et al., 2008; McAlearney et al., 2005) and "critical to the sustainability of the healthcare industry" (Ireri et al., 2011, p. 18), many physician executives are said to lack key leadership skills (Straus et al., 2013). Doctors are often promoted based on outstanding clinical care, research, or teaching expertise, not usually based on exemplary leadership and often having had little or no leadership training (Ackerly et al., 2011; M. Day, 2007; Ireri et al., 2011; H. Korschun et al., 2010; McAlearney et al., 2005; Mountford & Webb, 2009;

Sonnino, 2016). The result is that doctors tend to develop leadership capabilities though ad hoc, on-the-job learning, which is not a sufficient strategy overall given the increasing demands in the field (Blumenthal et al., 2014; Rose, 2015). The unfortunate result is that physicians' performance in their new roles is reported as often being mediocre or worse (Lazarus, 2009; Stoller, 2009). Similarly, McKinsey & Company reported that there are significant skills and knowledge deficits among middle and senior management NHS staff in the UK compared to their counterparts in industry and private health care (Ireri et al., 2011). A different McKinsey report declares that opportunities to nurture clinical leadership capabilities are scarce (Mountford & Webb, 2009). Finally, as mentioned previously, many junior doctors feel that they lack the leadership capabilities and opportunities to develop (Bohmer, 2012). Without structured training in leadership, Martins (2010) suggests that the risk is that aspects of doctors' practice will be left to trial and error or will remain under- or altogether un-developed, resulting in underperformance (Rose, 2015). In medicine, this kind of underperformance can lead to a loss in confidence in physician–managers, burnout, or, worse, can result in mismanagement of systems and ultimately jeopardise patient safety (Ackerly et al., 2011). The Chief Medical Officer (CMO) Clinical Advisor Alumni (2012) go so far as to say that if physicians do not get adequate leadership training, a crisis is imminent. Watkins (2003) states that 30 – 50 per cent of senior leaders fail or quit within 18 months of a new appointment, which is a situation no organisation can afford. For these reasons, it is clear to medical leadership proponents that programmatic approaches to leadership development are required at various stages of physicians' careers (Rose, 2015; Sonnino, 2016; Swanwick & McKimm, 2012; Van Aerde, 2013). The demands, complexity, and responsibilities of a modern healthcare organisation are too important to be left to "accidental leaders" (Rose, 2015; Satiani et al., 2014; Shah et al., 2013).

Background to Medical Leadership Summary

Therefore, given the changing nature of the healthcare field, including calls to improve the quality of care and at the same time tighten budgets, there is no question that imminent change is needed (Lee, 2010; Mountford & Webb, 2009; Rose, 2015). Medical leadership affects all branches of healthcare: public health, national health policies, trust and hospital administration and improvement, clinical teams, and individual patients. Everyone agrees that physicians are integral to the inevitable change in some way (Mountford & Webb, 2009) and many feel that this includes doctors taking on senior leadership roles. There is reported to be a dearth of clinicians who have the necessary leadership training and support (Rose, 2015) and the programmes that do exist are allegedly seldom based on empirical evidence or well evaluated. This has led to claims of leader underperformance and programmes that are suboptimal, which is a tremendous waste of valuable financial, human, and professional resources. This is a situation we can no longer afford (Rowland, 2016). All stakeholders are interested to know what works optimally and what the evidence is to substantiate those claims.

Appendix B: PhD Methodology Full



HEE Research Protocol Team Members

Role	Organisation
Director of Faculty	The Staff College: Leaders in Healthcare
Curriculum Director	The Staff College: Leaders in Healthcare
PhD researcher (author of this thesis)	The University of Cambridge
Lead instructor	The Staff College: Leaders in Healthcare
Fellow in Medical Education	Health Education England (HEE)
Education Development Manager	HEE
Research Associate	University of Liverpool School of Medicine
Junior doctor	HEE North West
Education Project Support officer	HEE North West
Associate Dean, Leadership	HEE North West
Junior doctor	HEE North West

HEE Literature Search Strategy

#	Database	Terms	Hits	Total
1	ABI	lead* (title), all others (subject)	237	237
	ABI	All terms (title)	104	341
2	Business Source Complete	lead* (title), all others (subject)	974	1315
3	Embase	lead* (title), all others (subject)	3891	5206
	Embase	All terms (title)	190	5396
4	ERIC	lead* (title), all others (subject)	809	6205
	ERIC	All terms (title)	414	6619
5	Medline/PubMed (NCBI)	lead* (title), all others (subject)	5055	11674
6	Scopus	All terms (title)	3390	15064
7	Web of Science	All terms (title)	3926	18990

Appendix C: Full Data Analysis: MULTI

Multi SLR Codes – Design, Data, and Sample 1/6

			Data Collected		Sample						
#	First Author and Publication Year	Methodology	Qual[a]/ Quant[b]/ Both[c]	Methods	# Part's[e]	# Control	F	M	Mean Age	Level of Seniority	Domain
1	Leslie (2005)	Case study	Both	Questionnaire, statistical analysis	56	-	31	25	-	Pediatricians	Healthcare
2	Mabey (2005)	Survey	Both	Statistical analysis, interviews	179	-	-	-	-	HR managers and line managers	179 firms
3	McAlearney (2005)	Case study	Both	Questionnaire	-	-	-	-	-	Physicians unspecified	Healthcare
4	Parry (2005)	Quasi-experiment	Both	Questionnaire	50	-	-	-	-	Mid-level managers	Private and public sector
5	Umble (2005)	Case study	Qual	Interview, document analysis	25	-	-	-	-	Senior leaders	Healthcare
6	Boaden (2006)	Case study	Qual	Questionnaire, document analysis, conversation analysis, programme observation	250	-	-	-	-	Prospective and current HR directors	Healthcare
7	Gilipin-Jackson (2006)	Case study	Qual	Questionnaire, interviews, programme observation	18	-	-	-	-	Senior and middle managers	Healthcare
8	Iles (2006)	Case study	Qual	Interviews, document analysis, programme observation	20	-	-	-	-	Chief or senior executives	Various
9	Terrion (2006)	Case study	Qual	Interviews	9	-	4	5	-	Senior directors	Higher education
10	Bowles (2007)	Case study	Qual	Questionnaire, statistical analysis, document analysis	59	-	5	54	-	Middle and executive managers	Military
					Mean: 103.7		**Total:** 1135	**Total:** 1951			

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[e] Participants

Multi SLR Codes – Design, Data, and Sample 2/6

			Data Collected		Sample						
#	First Author and Publication Year	Methodology	Qual[a]/ Quant[b]/ Both[c]	Methods	# Part's[e]	# Control	F	M	Mean Age	Level of Seniority	Domain
11	Dexter (2007)	Case study	Qual	Document analysis, interviews	32	-	-	-	-	Middle managers	Politics (municipal)
12	Hayes (2007)	Case study	Both	Questionnaire, document analysis	258	-	-	-	-	Supervisors and managers	Gaming industry
13	Ladyshewsky (2007)	Case study	Both	Questionnaire, document analysis	15	-	4	11	-	Middle level managers	Public sector
14	Miller (2007)	Case study	Both	Questionnaire	210	-	-	-	51	Senior public health leaders	Public health
15	Quaglieri (2007)	Case study	Both	Questionnaire, interviews	40-45	-	-	-	35	Mid-career executives	Business
16	Yeo (2007)	Case study	Qual	Interview	20	-	-	-	-	Middle, senior, and top managers	Manufactoring engineering
17	Butler (2008)	Action research	Both	Questionnaire	35	-	-	-	27-33	MBA and non MBA students	Business
18	D'Netto (2008)	Survey	Both	Questionnaire	206	-	69	137	37.26	Mixed: front line managers to CEO's	18 different industries
19	Magner (2008)	Case Study	Both	Questionnaire, interviews, document analysis	-	-	-	-	-	Young leaders (28-38)	Business
20	Raudenbush (2008)	Case study	Both	Questionnaire, interview	14	-	-	-	-	NASS managers	Government (federal)
					Mean: 103.7		**Total:** 1135	**Total:** 1951			

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[e] Participants

Multi SLR Codes – Design, Data, and Sample 3/6

			Data Collected		Sample						
#	First Author and Publication Year	Methodology	Qual[a]/ Quant[b]/ Both[c]	Methods	# Part's[e]	# Control	F	M	Mean Age	Level of Seniority	Domain
21	Suutari (2008)	Survey	Both	Questionnaire	878	-	386	492	-	Senior managers	Business
22	Dragoni (2009)	Case Study	Quant	Questionnaire	433	-	131	302	40	Junior managers, part-time MBA students, supervisors	Various (business, health care, military)
23	DeRue (2009)	Case study	Both	Questionnaire, interviews	60	-	15	45	33.4	Middle and senior managers	For and non-profit organisations
24	Drew (2009)	Case study	Qual	Questionnaire, interviews	8	-	-	-	-	Academic and professional (administrative) senior supervisory staff	Higher Education
25	Edmonstone (2009)	Case study	Both	Questionnaire, interviews	218	-	-	-	-	Senior medical leaders (primary and secondary care and public health)	Healthcare
26	Malling (2009)	Case Study	Quant	Questionnaire, statistical analysis	20	28	-	-	-	Consultant Responsible for Education	Healthcare
27	Murdock (2009)	Case study	Quant	Questionnaire	100	-	-	-	-	Community practice physicians	Healthcare
28	Stewart (2009)	Action research	Qual	Interviews, questionnaire, programme observations, focus groups, peer reviews	19	-	2	17	-	Small and medium sized business enterprise leaders	Business
					Mean: 103.7		**Total:** 1135	**Total:** 1951			

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[e] Participants

Multi SLR Codes – Design, Data, and Sample 4/6

			Data Collected		Sample						
#	First Author and Publication Year	Methodology	Qual[a]/ Quant[b]/ Both[c]	Methods	# Part's[e]	# Control	F	M	Mean Age	Level of Seniority	Domain
29	Ardts (2010)	Survey	Quant	Questionnaire	453	-	186	267	40	17% lower management, 56% upper, and 27% higher	Business
30	Cherry (2010)	Case study	Qual	Unclear	141	-	-	-	-	Junior physicians	Healthcare
31	Dalakoura (2010)	Survey	Quant	Questionnaire	112	-	-	-	-	Human resources managers of multinational firms	Business
32	Hassan (2010)	Experiment	Quant	Questionnaire	12	12	3	21	-	Area managers	Healthcare
33	Johnson (2010)	Case study	Quant	Questionnaire, statistical analysis	294	-	84	207	-	Mid and senior level managers	Business, nonprofit, and public sector
34	McGurk (2010)	Case study	Qual	Interviews, document analysis	26	-	-	-	-	Middle managers	Public social services
35	Simmonds (2010)	Case study	Both	Questionnaire, programme observation, interviews	282	-	-	-	-	Managers	Business
36	Abrell (2011)	Case study	Quant	Questionnaire	25	9	-	-	-	Managers	Pharmaceuticals
37	Chochard (2011)	Case study (multiple)	Both	Questionnaire, interviews	158	-	-	-	-	Managers	Unclear
38	Edmonstone (2011)	Case study	Both	Questionnaire, document analysis, interivews	125	-	-	-	-	Potential senior clinical leaders	Healthcare
					Mean: 103.7		**Total:** 1135	**Total:** 1951			

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[e] Participants

Multi SLR Codes – Design, Data, and Sample 5/6

			Data Collected		Sample						
#	**First Author and Publication Year**	**Methodology**	**Qual[a]/ Quant[b]/ Both[c]**	**Methods**	**# Part's[e]**	**# Control**	**F**	**M**	**Mean Age**	**Level of Seniority**	**Domain**
39	Petriglieri (2011)	Experiment	Qual	Document analysis, interviews	55	-	18	72	31	MBA students	Business
40	Pinnington (2011)	Survey	Both	Questionnaire	192	-	-	-	-	Chief executive or HR manager	Public, private, and nonprofit
41	Pless (2011)	Case study	Both	Interviews, document analysis, questionnaire	70	-	-	-	-	Executives and partners	Business
42	Sanfey (2011)	Grounded theory	Both	Questionnaire	142	-	22	32	-	Doctors, academics, and medical staff	Healthcare
43	Watkins (2011)	Case study (two)	Qual	Interviews, document analysis	56	-	-	-	-	Senior executives	Global healthcare
44	Berg (2012)	Case study	Qual	Interviews, surveys, programme observation	14	-	2	12	-	Middle managers and project managers	Business
45	Coloma (2012)	Case study	Both	Questionnaire	166	-	69	97	48	Middle managers	Human service
46	DeRue (2012)	Quasi experiment	Both	Questionnaires, document analysis, statistical analysis	93	80	54	119	28	First-year MBA students	Business
47	Galli (2012)	Case study	Qual	Observations, document analysis, questionnaire, interviews	30	-	-	-	-	Chief and senior executives and senior directors with potential	Business
48	Thomas (2012)	Case study	Qual	Questionnaires, programme observations, statistical analysis	87	-	-	-	-	Managers	Business
					Mean: 103.7		**Total:** 1135	**Total:** 1951			

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[e] Participants

Multi SLR Codes – Design, Data, and Sample 6/6

			Data Collected		Sample						
#	First Author and Publication Year	Methodology	Qual[a]/ Quant[b]/ Both[c]	Methods	# Part's[e]	# Control	F	M	Mean Age	Level of Seniority	Domain
49	Getha-Taylor (2013)	Case study	Both	Questionnaire, interviews	45	-	-	-	-	Senior and mid-level managers	Government
50	Jeon (2013)	Experiment	Both	Questionnaire, statistical analysis	24 sites	-	-	-	-	Middle managers	Healthcare
51	Vimr (2013)	Case study	Both	Questionnaire	-	-	-	-	-	Physician leaders	Healthcare
52	Blumenthal (2014)	Action research	Both	Questionnaire, programme observation	16	-	6	10	-	Residents (internal medicine)	Healthcare
53	de Jong (2014)	Case study	Qual	Questionnaire, Focus group interview	12	-	-	-	-	Healthcare professionals	Healthcare
54	Kwamie (2014)	Case Study	Both	Document analysis, participant observation, interviews	23	-	17	6	-	District managers	Healthcare
55	Aarons (2015)	Action research	Both	Questionnaire, focus group interview	6	5	7	4	39.58	Team leaders	Healthcare
56	Grotrian-Ryan (2015)	Case study	Both	Questionnaire, interviews	36	-	20	16	54	University academics	Higher education
					Mean: 103.7		**Total:** 1135	**Total:** 1951			

[a] Qualitative data only
[b] Quantitative data only
[c] Both qualitative and quantitative

[e] Participants

Multi SLR Codes – Programme 1/6

		Programme(s)				
#	**First Author and Publication Year**	**Selection criteria**	**Location**	**In-House vs External**	**Length**	**Developmental Activities**
1	Leslie (2005)	Nominated	US	External	3 days	Lectures, role-plays, team projects, presentations, small group discussions, self-assessment surveys, case studies, review of video taped scenarios, and PDP's
2	Mabey (2005)	Unclear	Europe (six countries)	Both	Unclear	Workshops, lectures, job rotation, internships, mentoring, coaching, e-learning, formal qualifications
3	McAlearney (2005)	Applied and selected	US	Unclear	Unclear	Lectures, group discussions, case studies, guest speakers
4	Parry (2005)	Unclear	Unclear	Both	3 months	360's, PDP's, coaching, action learning, reflection
5	Umble (2005)	Unclear	US	Unclear	1 year	MSF, feedback, coaching, required reading, lectures, group discussions, case studies, simulations, workshops, guest speakers, and a team project
6	Boaden (2006)	Unclear	UK	In-house	Unclear	Lectures, action learning projects
7	Gilipin-Jackson (2006)	Unclear	Canada	In-house	9 months	Lectures, coaching, assignments, journal
8	Iles (2006)	Nominated by chief or senior executives	UK	External	1 year	Guest speakers, group discussions, coaching, networking
9	Terrion (2006)	Volunteered	Canada	Unclear	Unclear	Workshops, psychometric/personality tests, 360's, and guest speakers
10	Bowles (2007)	Volunteered	US	In-house	1 year	Workshops

Multi SLR Codes – Programme 2/6

		Programme(s)				
#	**First Author and Publication Year**	**Selection criteria**	**Location**	**In-House vs External**	**Length**	**Developmental Activities**
11	Dexter (2007)	Unclear	UK	In-house	10 months	360's, discussion with a psychologist, lectures, assignments, examinations, journal, team project
12	Hayes (2007)	Unclear	Canada	In-house	Unclear	PDP's, coaching
13	Ladyshewsky (2007)	Applied and selected	Australia	In-house	2 years	360's, peer coaching
14	Miller (2007)	Unclear	US	External	1 year	Team-based action learning project, coaching, leadership assessment, psychometrics, 360's, PDP's, workshops, simulations, films, small-group work, group discussions, required readings, and distance-learning conference calls
15	Quaglieri (2007)	Nominated	US	External	10 months	Guest speakers, workshops, team project, required readings, lectures, presentation
16	Yeo (2007)	Unclear	Singarpore	In-house	3 days	Workshops
17	Butler (2008)	Unclear	United States	External	4 months	Coaching, lectures, required readings, group discussions, case studies, and simulations
18	D'Netto (2008)	Top 200 Australian companies	Australia	In-house	Unclear	Not listed
19	Magner (2008)	Unclear	South Africa	Unclear	2 years	Lectures, reflection, networking groups, storytelling, group discussions, site visits, guest speakers
20	Raudenbush (2008)	Applied and selected	US	Unclear	1 year	Workshops, PDP's, action learning project, required reading, coaching

Multi SLR Codes – Programme 3/6

		Programme(s)				
#	First Author and Publication Year	Selection criteria	Location	In-House vs External	Length	Developmental Activities
21	Suutari (2008)	Invitation from a database	Finland	Unclear	Unclear	Action learning, mentoring, job rotation, required readings, lectures, workshops, outdoor development, customized approaches, films, and e-learning
22	Dragoni (2009)	Unclear	United States	Unclear	Unclear	Job assignment
23	DeRue (2009)	Volunteered	United States	External	Unclear	Critical incidents, on-the-job learning
24	Drew (2009)	Nominated by their supervisors	Australia	In-house	1 year	360's
25	Edmonstone (2009)	Unclear	England	External	1 year	Workshops and PDP's
26	Malling (2009)	Unclear	Denmark	External	6 months	Workshops, 360's, assignments, and PDP's
27	Murdock (2009)	Volunteered	US	External	20 weeks	Workshops, guest speakers, required reading
28	Stewart (2009)	Unclear	UK	External	6 - 9 months	Action learning, workshops, coaching

Multi SLR Codes – Programme 4/6

		Programme(s)				
#	**First Author and Publication Year**	**Selection criteria**	**Location**	**In-House vs External**	**Length**	**Developmental Activities**
29	Ardts (2010)	Mixed	Netherlands	In-house	Unclear	Not listed
30	Cherry (2010)	Unclear	United States	Unclear	9 months	Workshops, mentoring, lectures, action learning, case study analysis, guest speakers, group discussion, peer feedback, team project, peer support, networking
31	Dalakoura (2010)	The firm had ≥ 50 employees, an HRM department, and the HRM manager had served for at least two years	Greece	Both	Unclear	Not listed
32	Hassan (2010)	Unclear	Unclear	In-house	1 year	Feedback, workshops, role playing
33	Johnson (2010)	Unclear	United States	External	5 days	Assignments, group activities, reflection, coaching, PDP's, presentation, 360's
34	McGurk (2010)	Unclear	UK	In-house	1 year	Lectures
			UK	In-house	3 days	Reflection, workshops
			UK	In-house	6 months	Workshops, action learning projects
35	Simmonds (2010)	Unclear	UK	In-house	Unclear	Workshops, 360's, action learning projects, PDP's, required reading
36	Abrell (2011)	Unclear	Germany	In-house	1 year	Lectures, role-play, group discussion, 360's, peer coaching, PDP's
37	Chochard (2011)	Unclear	Switzerland	In-house	1 - 2 days, 4 - 9 days, and 5 - 13 days	Not listed
38	Edmonstone (2011)	Applied and selected	Scotland	In-house	1 year	Workshops, 360's, psychometrics, action learning, PDP's, observed case study, guest speakers, coaching, and job shadowing

Multi SLR Codes – Programme 5/6

		Programme(s)				
#	**First Author and Publication Year**	**Selection criteria**	**Location**	**In-House vs External**	**Length**	**Developmental Activities**
39	Petriglieri (2011)	Volunteered	Unclear	External	6 months	Psychotherapy
40	Pinnington (2011)	Unclear	UK	In-house	Unclear	Action learning projects, job assignment, networking, 360, coaching, mentoring
41	Pless (2011)	Unclear	Several countries	In-house	Unclear	Action learning projects, team building, coaching, 360's, reflection, meditation and yoga, storytelling, service learning
42	Sanfey (2011)	Nominated by department chair	United States	In-house	10 weeks	360's, psychometric tests, workshops
43	Watkins (2011)	Unclear	United States	In-house	4 months	Action learning projects, 360's, group cohort discussions, coaching, mentoring
44	Berg (2012)	Unclear	Unclear	Unclear	Unclear	Workshops, case studies, group discussion, facilitator feedback, peer and team coaching, assignments
45	Coloma (2012)	Nominated by supervisors	United States	External	5 months	360's, PDP's, workshops, journaling, coaching, required reading, group discussion, and networking
46	DeRue (2012)	Unclear	United States	External	9 months	Team building, simulation, internships, case competition, facilitator feedback, coaching, reflection
47	Galli (2012)	Unclear	Europe	In-house	Unclear	360-degree feedback, coaching, mentoring, job assignment, workshops, action learning, networking, site visits
48	Thomas (2012)	Unclear	United States	In-house	Unclear	Psychometric tests, 360's, assessment, coaching, PDP's, mentors, team action learning project

Multi SLR Codes – Programme 6/6

		Programme(s)				
#	First Author and Publication Year	Selection criteria	Location	In-House vs External	Length	Developmental Activities
49	Getha-Taylor (2013)	Unclear	United States	External	3 days	Lectures, discussion, case study analysis, group activities
		Unclear	United States	External	1 month	Lectures, group discussions, case study analysis, group activities, simulation/role plays, 360's, action learning projects
50	Jeon (2013)	Volunteered	Australia	In-house	10 - 12 months	Action learning projects, 360's, case studies, coaching
51	Vimr (2013)	Required	Canada	In-house	8 months	360's, lectures, small group discussions, reflection, journaling, required readings, case studies, team action learning projects, and coaching
52	Blumenthal (2014)	Volunteered	United States	In-house	1 month	Required reading, assignments, role plays, small group discussions, lectures, case study analysis, group discussion
53	de Jong (2014)	Unclear	UK	External	8 weeks	Lectures, small group sessions, e-learning
54	Kwamie (2014)	Randomly selected	Ghana	External	6 months	Workshops, coaching, facilitation teams to help projects, and group action learning projects
55	Aarons (2015)	Volunteered	United States	In-house	6 months	360's, coaching, workshops
56	Grotrian-Ryan (2015)	Nominated	United States	External	1 year	Mentoring

Multi SLR Codes – Outcomes 1/6

#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Reported Benefits and Outcomes
1	Leslie (2005)	Self	Sub #[e]	Both	Pre[j], D[K], P[l], PP[m]	1, 2b, 3a	3	1) PPE; 2b) Increased skills and competencies; 3a) Met individual goals
2	Mabey (2005)	Self	Sub #, sub desc[f]	Prog[h]	P	4a	1	4a) Organisational productivity
3	McAlearney (2005)	Self	Sub desc	Both	D, PP	1, 2a, 2b, 3a	4	1) PPE, 2a) Increased confidence, aspirations to lead, 2b) Increased leadership skills, increased teamwork skills, 3a) Increased leadership behaviours, increased leadership effectiveness
4	Parry (2005)	Self, sub[a], peer, sup[b]	Obj[g]	Part[i]	P	3a, 3b, 4a	3	3a) Increased leadership behaviours, 3b) Satisfaction with leadership, 4a) Extra effort of followers
5	Umble (2005)	Self	Sub desc	Both	PP	2a, 3a, 4a	3	2a) Changed perspective, increased confidence; 3a) Increased leadership behaviours; 4a) Increased social capital
6	Boaden (2006)	Self	Sub #, sub desc	Both	D, PP	2a, 2b, 3a, 4a	4	2a) Personal development; 2b) Increased knowledge; 3a) Increased leadership behaviours, increased influence, positive impact on their careers; 4a) Raised profile of the department
7	Gilipin-Jackson (2006)	Self, sup	Sub #, sub desc, obj	Part	PP	2a, 2b, 3a, 3b	4	2a) Increased self-awareness, increased systems thinking; 2b) Increased skills, increased knowledge; 3a) Increased leadership behaviours; 3b) Increased leadership behaviours (verified by observers)
8	Iles (2006)	Self	Sub desc, obj	Both	D, PP	1, 3a	2	3a) Networking benefits
9	Terrion (2006)	Self	Sub desc	Both	P	1, 2a, 2b, 3a	4	2a) Increased self-awareness, greater appreciation of others' perspectives; 2b) Increased skills; 3a) Networking benefits
10	Bowles (2007)	Facil[c], stats[d]	Sub desc, obj	Part	D, PP	3b	1	3b) Meeting individual goals/higher levels of quota achievement

[a] Subordinate
[b] Superior
[c] Facilitator
[d] Statistics

[e] Subjective numbers
[f] Subjective descriptions
[g] Objective

[h] Programme
[i] Participants

[j] Pre-programme
[k] During
[l] Post
[m] Post-post
[n] Retro post

Multi SLR Codes – Outcomes 2/6

#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Reported Benefits and Outcomes
11	Dexter (2007)	Self, sup	Sub desc, obj	Part	P	2b, 3a, 4a	3	2b) Improved self-management, improved teamwork skills; increased leadership competence; 3a) Networking benefits; 4a) Better organisational processes
12	Hayes (2007)	Self, sup	Sub #	Both	Pre, D, P, PP	1, 2b, 3a, 4a	4	1) PPE's; 2b) Increased knowledge, increased leadership skills, increased communication skills, 3a) Meeting individual goals, improved leadership competence; 4a) Increased organisational performance
13	Ladyshewsky (2007)	Self, sub, peer	Sub #, sub desc	Both	Pre, D, P	2b, 3b	2	2b) Increased skills; 3b) Improved leadership behaviours
14	Miller (2007)	Self	Sub #, sub desc	Both	PP, Retro[n]	1, 2a, 2b, 3a, 4a	5	1) PPE's; 2a) Increased confidence and self-awareness; 2b) Increased leadership knowledge and skills; 3a) Increased leadership behaviours; 4a) General organisational benefits, developing and strengthening their organizations' collaborative relationships, and developing or implementing a new programme.
15	Quaglieri (2007)	Self	Sub #, obj	Both	Pre, D, P, PP	1, 2b, 4a	3	1) PPE's; 2b) Increased skills; 4a) Increased recruitment, improved retention
16	Yeo (2007)	Self	Sub desc	Both	PP	2a, 2b	2	2a) Increased capacity to learn; 2b) Increased leadership competence
17	Butler (2008)	Self	Sub #, sub desc	Both	Pre, P	1, 2b	2	2a) Increased self-awareness; 2b) Increased leadership skills, increased communication skills
18	D'Netto (2008)	Self	Sub #, sub desc	Prog	PP	1, 2a, 2b, 3a	4	1) PPE's; 2b) Increased leadership skills
19	Magner (2008)	Self	Sub #, sub desc	Both	P	1, 2b	2	1) PPE's; 2a) Increased awareness of the role they could play, increased systems thinking, greater appreciation of others' perspectives, increased self-awareness; 2b) Increased leadership competence; 3a) Networking benefits
20	Raudenbush (2008)	Self, sub, peer, sup, facil, stats	Sub #, sub desc	Part	Pre, D, P, PP	1, 2a, 2b	3	1) PPE's; 2b) Increased leadership competence

[a] Subordinate
[b] Superior
[c] Facilitator
[d] Statistics

[e] Subjective numbers
[f] Subjective descriptions
[g] Objective

[h] Programme
[i] Participants

[j] Pre-programme
[k] During
[l] Post
[m] Post-post
[n] Retro post

Multi SLR Codes – Outcomes 3/6

#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Reported Benefits and Outcomes
21	Suutari (2008)	Self	Sub desc	Part	PP	1, 2a	2	1) PPE's; 2a) Broadening of understanding
22	Dragoni (2009)	Self, sup	Sub #	Both	Unclear	2a, 2b	2	1) PPE's; 2b) Increased leadership competency
23	DeRue (2009)	Self, sup	Sub #, sub desc	Part	P	2a, 2b	2	2a) Increased self-awareness; 2b) Increased leadership skills
24	Drew (2009)	Self, sub, peer, sup	Sub #, sub desc, obj	Part	Pre, P	2b, 3b	2	2b) Increased leadership skills; 3b) Improved leadership behaviours
25	Edmonstone (2009)	Self, fac	Sub desc	Prog	PP	1, 2a, 2b, 3a, 4a	5	1) PPE; 2a) Greater appreciation of others' perspectives, increased engagement, enhanced common identity, increased confidence; 2b) Increased leadership skills; 3a) Increased leadership behaviours, networking benefits, developed PDP's; 4a) Having launched a new initiative
26	Malling (2009)	Self, sub, peer	Sub #, sub desc	Both	Pre, PP	1, 2b, 3a	3	1) PPE; 2b) Increased knowledge; 3a) Increased leadership behaviours
27	Murdock (2009)	Self	Sub #, sub desc	Both	Pre, P	1, 2a, 2b, 3a, 3b	5	1) PPE; 2a) Increased aspirations to lead; 2b) Increased leadership skills; 3a) Increased leadership behaviours; 3b) Having taken on a leadership role
28	Stewart (2009)	Self	Sub #, sub desc, obj	Both	P	1	1	1) PPE's

[a] Subordinate
[b] Superior
[c] Facilitator
[d] Statistics

[e] Subjective numbers
[f] Subjective descriptions
[g] Objective

[h] Programme
[i] Participants

[j] Pre-programme
[k] During
[l] Post
[m] Post-post
[n] Retro post

Multi SLR Codes – Outcomes 4/6

#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Reported Benefits and Outcomes
29	Ardts (2010)	Self	Sub #, obj	Both	N/A	1	1	1) PPE
30	Cherry (2010)	Peer, sup, fac	Sub desc	Part	D, PP	1, 2b, 3b	3	1) PPE's; 2a) Increased aspirations to lead, increased self-awareness, increased leadership self-identity; 2b) Increased knowledge and skills; 3a) Increased leadership behaviours, networking benefits; 3b) Promotions
31	Dalakoura (2010)	Stats	Sub #	Prog	P	4a	1	4a) Increased financial and market performance of the organisation
32	Hassan (2010)	Self, sub	Sub #	Part	Pre, PP	2b, 3b	2	2b) Increased developing others, developed building relationships skills, 3b) Increased leadership behaviours
33	Johnson (2010)	Self, sub, peer, sup	Sub #	Part	PP, Retro	2a, 2b, 3b, 4a	4	2b) Increased developing others, developed building relationships skills, 3b) Increased leadership behaviours
34	McGurk (2010)	Self	Sub desc	Part	Unclear	2a, 2b, 3a, 4a	4	2a) Developed a deeper understanding of organisational strategy, 2b) Increased leadership effectiveness, developed interpersonal skills, increased teamwork skills; 3a) Increased leadership behaviours; 4a) Service improvements
35	Simmonds (2010)	Self, peer	Sub desc	Both	Pre, P	1, 2a, 2b, 3a, 4a	5	1) PPE; 2a) Increased commitment, increased personal development; 2b) Developed networking skills; 3a) Increased leadership behaviours; 4a) Increased organisational productivity, greater staff engagement, higher sales, and reduced staff turnover
36	Abrell (2011)	Self, sub, peer, sup	Sub #	Part	Pre, P	2b, 3b	2	2b) Increased leadership skills; 3b) Improved supervisor's rating of increased leadership behaviour
37	Chochard (2011)	Self, sup	Sub #, sub desc, obj	Both	Pre, P	1, 2b, 3b	3	1) PPE's; 2b) Increased leadership skills; 3b) Increased leadership behaviours
38	Edmonstone (2011)	Self, sub, peer, sup	Sub desc, obj	Prog	Pre, P	1, 2a, 2b, 3b, 4a, 4b	6	1) PPE's; 2a) Increased self-awareness, increased resilience, increased engagement; 2b) Developing interpersonal and networking skills; 3b) Colleagues' feedback on behaviour changes, promotions, 4a) Policy changes, developed organisational capacity; 4b) Implementing action learning projects; Other) Having joined a mentoring network

[a] Subordinate
[b] Superior
[c] Facilitator
[d] Statistics

[e] Subjective numbers
[f] Subjective descriptions
[g] Objective

[h] Programme
[i] Participants

[j] Pre-programme
[k] During
[l] Post
[m] Post-post
[n] Retro post

Multi SLR Codes – Outcomes 5/6

#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Reported Benefits and Outcomes
39	Petriglieri (2011)	Self	Sub desc	Both	Pre, D, P	1	1	1) PPE's
40	Pinnington (2011)	Facil	Sub desc, obj	Prog	N/A	1	1	1) PPE's
41	Pless (2011)	Self	Sub desc	Both	PP	2a, 2b, 3a	3	2a) Increased personal development, increased knowledge, increased servant leadership attitude, increased awareness of responsible global leadership; 2b) Developed interpersonal skills, developed their sense of responsibility and ethics; 3a) Increased leadership behaviours
42	Sanfey (2011)	Self	Sub #, sub desc, obj	Part	Pre, P, PP	1, 2a, 2b, 3b	4	1) PPE's; 2a) Increased aspirations to lead, increased self-awareness, increased leadership self-identity; 2b) Increased knowledge and skills; 3a) Increased leadership behaviours, networking benefits; 3b) Promotions
43	Watkins (2011)	Self, sub, peer, sup	Sub desc	Both	P, PP	1, 2a, 2b, 3b, 4a, 4b	6	1) PPE's; 2a) Increased commitment, increased confidence; 2b) Increased leadership skills, developed communication skills, developed teamwork skills, developed mentoring skills; 3b) Promotions, implemented tools; 4a) Development of a common language; 4b) Having implemented action learning projects
44	Berg (2012)	Self, sub, sup	Sub desc	Both	Pre, D, P	1, 2a, 2b, 3a	4	1) PPE's; 2a) Increased confidence, increased self-awareness, increased self-efficacy, having developed one's own leadership style; 2b) Increased knowledge, developed a series of tools; 3a) Increased leadership behaviours
45	Coloma (2012)	Self, sup	Sub #, obj	Both	Pre, P, PP	1, 2a, 2b, 3a, 3b	5	1) PPE's; 2a) Broadened perspective, increased confidence; 2b) Increased knowledge, increased leadership skills, increased political savvy, increased communication skills, increased leadership competence; 3a) Networking benefits, increased work performance; 3b) Promotions
46	DeRue (2012)	Peer, sup, fac, stats	Sub #	Part	Pre, D, P	3a	1	3a) Increased leadership behaviours
47	Galli (2012)	Self	Sub desc	Both	Unclear	1, 3a, 3b	3	1) PPE's; 3a) Increased leadership behaviour; 3b) Awards
48	Thomas (2012)	Self, sub, peer, sup, facil, stats	Sub #, sub desc	Part	Pre, P	2a, 3a, 4a	3	2a) Enhanced common identity; 3a) Increased leadership behaviours; 4a) Reduced staff absenteeism

[a] Subordinate
[b] Superior
[c] Facilitator
[d] Statistics

[e] Subjective numbers
[f] Subjective descriptions
[g] Objective

[h] Programme
[i] Participants

[j] Pre-programme
[k] During
[l] Post
[m] Post-post
[n] Retro post

Multi SLR Codes – Outcomes 6/6

#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Reported Benefits and Outcomes
49	Getha-Taylor (2013)	Self, peer	Sub desc	Part	PP	1, 2a, 2b, 3a	4	1) PPE's; 2a) Increased self-awareness, increased appreciation for the value of collaboration; 2b) Increased leadership skills; 3a) Increased leadership behaviours
		Self, sub, peer, sup	Sub desc	Part	Unclear			
50	Jeon (2013)	Self	Sub desc, obj	Both	Pre, P, PP	N/A	N/A	N/A
51	Vimr (2013)	Self	Sub #, sub desc	Both	P	1, 2a, 3a, 4b	4	1) PPE's; 2a) Improved self-awareness, developed a systems view; 3a) Increased leadership behaviours; 4b) Having implemented action learning projects
52	Blumenthal (2014)	Self	Sub #, sub desc	Both	P	1, 2a, 2b	3	1) PPE's; 2a) Increased confidence, increased self-awareness, increased awareness of different leadership styles, increased interest in further training; 2b) Increased knowledge and skills
53	de Jong (2014)	Self	Sub desc	Prog	P	1, 2a, 3a	3	1) PPE's, 2a) Greater awareness of leadership challenges in public health, 3a) Gained experience as leaders
54	Kwamie (2014)	Self	Sub #, sub desc	Prog	P	3b, 4b	2	3b) Promotions; 4b) Achieved clinical targets
55	Aarons (2015)	Self, sub	Sub #, sub desc	Both	Pre, P, PP	1, 2a, 2b, 3a	4	1) PPE's; 2a) Increased appreciation for the utility of training; 2b) Increased knowledge, increased leadership competence, greater ability to manage change; 3a) Increased leadership behaviours
56	Grotrian-Ryan (2015)	Self	Sub desc	Both	P	1	1	1) PPE's

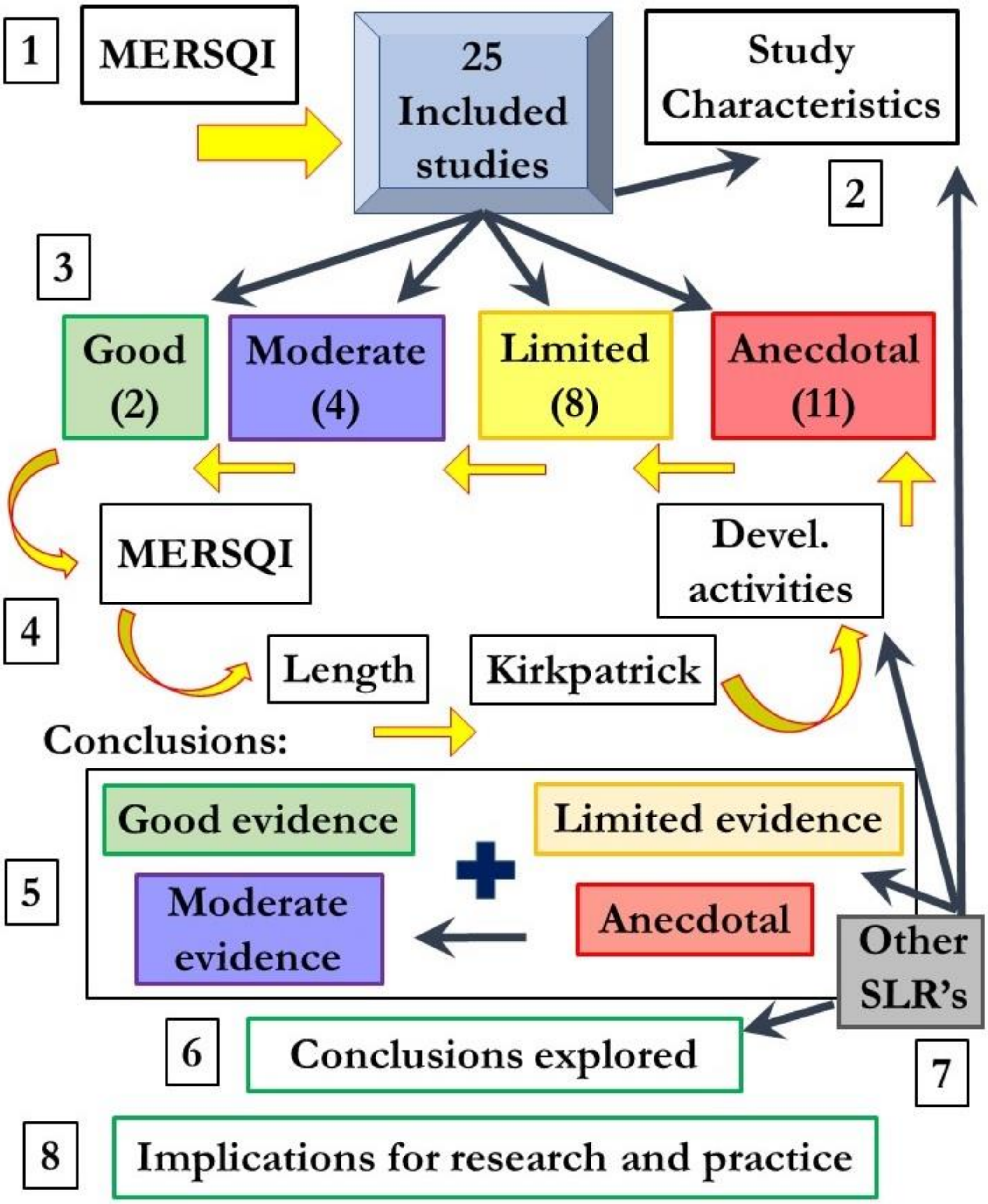
[a] Subordinate
[b] Superior
[c] Facilitator
[d] Statistics

[e] Subjective numbers
[f] Subjective descriptions
[g] Objective

[h] Programme
[i] Participants

[j] Pre-programme
[k] During
[l] Post
[m] Post-post
[n] Retro post

Appendix D: Full Data Analysis: HEE



HEE SLR Codes – Design, Data, and Sample 1/4

			Data Collected			Samples						
#	First Author and Publication Year	Methodology	Qual[a]/ Quant[b]/ Both[c]	Methods	Response Rate	# Part's[e]	# Control	Female	Male	Mean Age	MD's[f] Only?	Level of Seniority
1	Hemmer (2007)	Action research/ Case study	Qual	Questionnaire, document analysis	N/A	16	-	-	-	-	N	Residents (internal)
2	Korschun (2007)	Action research/ Case study	Both	Questionnaire, interviews	79%	70	-	-	-	-	N	Mixed
3	Miller (2007)	Case study	Both	Questionnaire	66%	210	-	-	-	51	N	Senior leaders
4	Dannels (2008)	Experiment	Quant	Questionnaire	38 - 71%	78	468	78	0	-	N	Senior faculty
5	Bergman (2009)	Multiple case study	Both	Questionnaire, focus group interviews	74%	109	-	95	14	-	N	First-line managers
6	Edmonstone (2009)	Case study	Qual	Questionnaire, interviews, document analysis	56.50%	218	-	-	-	-	N	Senior leaders
						Mean: 72.5		**Total:** 304	**Total:** 157		**Y Total:** 15	

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[d] Unclear: questionnaire or interview

[e] Participants

[f] Medical doctors

HEE SLR Codes – Design, Data, and Sample 2/4

			Data Collected			Samples						
#	First Author and Publication Year	Methodology	Qual[a]/ Quant[b]/ Both[c]	Methods	Response Rate	# Part's[e]	# Control	Female	Male	Mean Age	MD's[f] Only?	Level of Seniority
7	Malling (2009)	Quasi-experiment	Quant	Questionnaire, statistical analysis	77.25%	20	28	-	-	-	Y	Consultant (education)
8	Murdock (2009)	Case study	Quant	Questionnaires	N/A	100	-	-	-	-	Y	Physicians unspecified
9	Cherry (2010)	Case study	Qual	Unclear	N/A	141	-	-	-	-	Y	Junior physicians
10	Day (2010)	Case study	Quant	Questionnaire, document analysis	53%	100	73	-	-	-	Y	Surgeons unspecified (Orthopaedics)
11	Kuo (2010)	Case study	Both	Questionnaire	93.50%	15	-	-	-	-	Y	Residents (pediatric)
12	Edmonstone (2011)	Case study	Both	Questionnaire, document analysis, interviews	N/A	125	-	-	-	-	N	Potential senior clinical leaders
						Mean: 72.5		**Total:** 304	**Total:** 157		**Y Total:** 15	

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[d] Unclear: questionnaire or interview

[e] Participants

[f] Medical doctors

HEE SLR Codes – Design, Data, and Sample 3/4

#	First Author and Publication Year	Methodology	Data Collected			Samples						
			Qual[a]/ Quant[b]/ Both[c]	Methods	Response Rate	# Part's[e]	# Control	Female	Male	Mean Age	MD's[f] Only?	Level of Seniority
13	Sanfey (2011)	Grounded theory	Both	Questionnaire	50%	142	-	50	92	-	N	Mixed
14	Bearman (2012)	Case study	Qual	Questionnaire, MSF	92%	12	-	-	-	-	Y	Residents (surgical trainees)
15	Shah (2013)	Case study	Qual	Video analysis, q/i[d]	N/A	40	-	-	-	-	Y	Consultant (ophthalmic surgeons)
16	Ten Have (2013)	Quasi-experiment	Quant	Experiment, questionnaire	100%	9	10	7	12	-	Y	Midlevel surgeons
17	Vimr (2013)	Action research/ Case study	Both	Questionnaire	N/A	-	-	-	-	-	Y	Physicians unspecified
18	Blumenthal (2014)	Action research	Both	Questionnaire, observation	100%	16	-	10	6	30	Y	Residents (Internal medicine)
						Mean: 72.5		**Total:** 304	**Total:** 157		**Y Total:** 15	

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[d] Unclear: questionnaire or interview

[e] Participants

[f] Medical doctors

HEE SLR Codes – Design, Data, and Sample 4/4

			Data Collected			Samples						
#	First Author and Publication Year	Methodology	Qual[a]/ Quant[b]/ Both[c]	Methods	Response Rate	# Part's[e]	# Control	Female	Male	Mean Age	MD's[f] Only?	Level of Seniority
19	Dickey (2014)	Action research/ Case study	Qual	Unclear	N/A	-	-	-	-	-	Y	Residents (psychiatry)
20	MacPhail (2014)	Case study	Both	Questionnaire, interviews	70%	39	-	32	7	-	N	Middle and senior leaders
21	Satiani (2014)	Case study	Both	Questionnaire, q/i	N/A	-	-	-	-	-	Y	Physicians (early to mid career high potentials)
22	Nakanjako (2015)	Case study	Qual	Document analysis	N/A	15	-	-	-	-	N	Physicians unspecified
23	Patel (2015)	Action research	Both	Questionnaire, q/i, statistical analysis	77%	62	-	-	-	-	Y	Residents
24	Fernandez (2016)	Quasi-experiment	Both	Questionnaire	60%	37	-	26	11	46	Y	Junior fellows
25	Pradarelli (2016)	Case study	Qual	Interviews	100%	21	-	6	15	47	Y	Surgeon unspecified
						Mean: 72.5		**Total:** 304	**Total:** 157		**Y Total:** 15	

[a] Qualitative data only

[b] Quantitative data only

[c] Both qualitative and quantitative

[d] Unclear: questionnaire or interview

[e] Participants

[f] Medical doctors

HEE SLR Codes – Programme 1/4

		Programme(s)								
#	First Author and Publication Year	Selection Criteria	Location	Faculty	Sites	Needs Assessment	Goals?	In House vs External	Length	Pedagogical Methods
1	Hemmer (2007)	Unclear	US	Mixed	Single	Unclear	Unclear	In-house	1 year	Required reading, workshops, small group sessions, lectures, final exam, final presentations, group project, case studies, and team building exercises
2	Korschun (2007)	Nominated	US	Mixed	Single	Yes	Yes	In-house	5 months	Lectures, guest speakers, case study analysis, workshops, experiential exercises, group discussions, individual assessment and feedback, one-on-one executive coaching, psychometrics, 360's, individual counseling, PDP's, and mentoring
3	Miller (2007)	Unclear	US	Unclear	Single	Unclear	Unclear	External	1 year	Team-based action learning project, coaching, leadership assessment, psychometrics, 360's, PDP's, skill-building seminars, simulations, films, small-group work, discussions, readings, and distance-learning conference calls
4	Dannels (2008)	Applied and selected	US	Unclear	Single	Unclear	Yes	External	Unclear	Unclear
5	Bergman (2009)	Volunteered	Sweden	External	Single	Unclear	Unclear	In-house	1 week	Workshops, facilitator feedback, facilitated discussion, peer support, self-reflection
6	Edmonstone (2009)	Nominated	England	Mixed	Multiple	Yes	Yes	External	1 year	Workshops and PDP's

HEE SLR Codes – Programme 2/4

		Programme(s)								
#	First Author and Publication Year	Selection Criteria	Location	Faculty	Sites	Needs Assessment	Goals?	In House vs External	Length	Pedagogical Methods
7	Malling (2009)	Volunteered	Denmark	Unclear	Single	Yes	Unclear	External	6 months	Workshops, 360's, written assignments, and PDP's
8	Murdock (2009)	Unclear	US	Mixed	Single	Unclear	Yes	External	20 weeks	Workshops, guest speakers, reading assignments
9	Cherry (2010)	Unclear	US	Internal	Single	Yes	Yes	In-house	9 months	Workshops, mentoring, lectures, action learning, case study analysis, guest speakers, facilitated discussion, peer feedback, team challenge, peer support, networking
10	Day (2010)	Applied and selected	US	Unclear	Single	Unclear	Yes	External	1 year	Mentoring
11	Kuo (2010)	Applied and selected	US	Internal	Single	Yes	Yes	In-house	3 years	Workshops, required reading, e-portfolio's, mentoring, role plays, case study analysis, psychometric tests, facilitator feedback, advisory groups, small group discussion, job shadowing, guest speakers
12	Edmonstone (2011)	Applied and selected	Scotland	Unclear	Single	Yes	Yes	In-house	1 year	Workshops, 360's, psychometrics, action learning, PDP's, observed case study exercise, guest speakers, coaching, and job shadowing

HEE SLR Codes – Programme 3/4

	Programme(s)									
#	First Author and Publication Year	Selection Criteria	Location	Faculty	Sites	Needs Assessment	Goals?	In-House vs External	Length	Pedagogical Methods
13	Sanfey (2011)	Nominated	US	Mixed	Single	No	Yes	In-house	10 weeks	360's, psychometric tests, workshops
14	Bearman (2012)	Nominated	Australia and New Zealand	Unclear	Single	Unclear	Yes	In-house	2 days	Workshops, simulations, role-plays, peer feedback, facilitator feedback, small group discussions, 360's
15	Shah (2013)	Unclear	UK	External	Single	Yes	Yes	In-house	2 days	Team challenges, role plays, small group discussions, presentations, and self-reflection
16	Ten Have (2013)	Required	Netherlands	Internal	Single	Unclear	Unclear	In-house	1 day	Workshops, simulations, video taped, facilitator feedback,
17	Vimr (2013)	Required	Canada	Unclear	Single	Unclear	Unclear	In-house	8 months	360's, lectures, small group discussions, self-reflection, journaling, readings, case studies, team action learning projects, and coaching
18	Blumenthal (2014)	Volunteered	US	Internal	Single	Yes	Yes	In-house	1 month	Reading assignments, role plays, small group discussions, lectures, case study analysis, facilitated discussion

HEE SLR Codes – Programme 4/4

	Programme(s)									
#	First Author and Publication Year	Selection Criteria	Location	Faculty	Sites	Needs Assessment	Goals?	In-House vs External	Length	Pedagogical Methods
19	Dickey (2014)	Volunteered	US	Internal	Single	No	Yes	In-house	4 years	Workshops, reading assignments, simulations, mentoring, lectures, action learning, experiential activities
20	MacPhail (2014)	Nominated	Australia	Mixed	Single	No	Yes	In-house	9 - 10 months	Guest speaker, small group discussions, reading assignments, facilitated discussion, observed case study, self-reflection
21	Satiani (2014)	Nominated	US	Mixed	Single	Yes	Yes	In-house	18 months	Workshops, reading assignments, small group discussions, presentations, and journals
22	Nakanjako (2015)	Applied and selected	Uganda	Mixed	Single	Unclear	Yes	In-house	1 year	Lectures, small group discussions, case study analysis, mentoring, action learning, and online modules.
23	Patel (2015)	Volunteer	US	Internal	Single	Unclear	Yes	In-house	2 years	Lectures, required readings, videos, small group discussions, simulations, mentoring, action learning, selected online modules, and facilitated discussion
24	Fernandez (2016)	Nominated	US	Mixed	Single	No	Yes	External	3.5 days	Workshops, small group discussions, 360's, coaching, and psychometrics
25	Pradarelli (2016)	Applied and selected	US	Internal	Single	Yes	Unclear	In-house	8 months	Case study analysis, workshops, action learning projects, reading assignments, 360's, guest speakers, PDP's, coaching, and peer feedback

HEE SLR Codes – Outcomes 1/4

		Measurements									
#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Outcome Metrics	Kirkpatrick Levels	Total	Reported Outcomes and Benefits
1	Hemmer (2007)	Self, fac[a]	Sub #[d]	Both	Base[i], P[j], PP[k]	1, 2b	2	1) PPE's; 2b) Increased knowledge tests results	1, 2b	2	1) PPE's; 2b) Increased knowledge tests results
2	Korschun (2007)	Self	Sub desc[e], obj[f]	Both	PP	1, 2a, 3b	3	1) PPE's; 2a) Increased aspirations to lead, increased engagement, increased commitment; 3b) Retention, promotions, have taken on more leadership responsibilities	1, 2a, 2b, 3a, 3b, 4a, 4b	7	1) PPE's; 2a) Increased aspirations to lead, increased engagement, increased commitment; 2b) Increased knowledge and interpersonal and teamwork skills; 3a) Increased leadership effectiveness, networking benefits, have taken on more responsibility; 3b) Retention, promotions, have taken on a leadership role, increased committee involvement; 4a) Having launched a new initiative; 4b) Having implemented action learning projects
3	Miller (2007)	Self	Sub #, sub desc	Both	PP, Retro P[l]	2a, 2b, 3a, 4a	4	2a) Increased confidence and self-awareness; 2b) Increased leadership knowledge and skills; 3a) Increased leadership behaviours; 4a) General organizational benefits, developing and strengthening their organizations' collaborative relationships, and developing and implementing a new programme	1, 2a, 2b, 3a, 4a	5	1) PPE's; 2a) Increased confidence and self-awareness; 2b) Increased leadership knowledge and skills; 3a) Increased leadership behaviours; 4a) General organisational benefits, developing and strengthening their organizations' collaborative relationships, and developing or implementing a new programme.
4	Dannels (2008)	Self	Sub desc, obj	-	Base, PP	2a, 2b, 3b	3	2a) Increased aspirations to lead; 2b) Increased knowledge; 3b) Improved MSF pre and post, promotions	2a, 2b, 3b	3	2a) Increased aspirations to lead; 2b) Increased knowledge; 3b) Improved MSF pre and post, promotions, higher participation in further leadership developmemt following programmes
5	Bergman (2009)	Self	Sub desc	Both	Pre[m], PP	1, 2a, 2b, 3a	4	1) PPE's; 2a) Increased confidence; 2b) Increased knowledge; 3a) Increased leadership behaviours	1, 2a, 2b, 3a	4	1) PPE's; 2a) Increased confidence and self-awareness; 2b) Increased knowledge and communication skills; 3a) Increased leadership behaviours
6	Edmonstone (2009)	Self, fac	Sub desc	Prog[g]	PP	1, 3a	2	1) PPE's; 3a) Increased leadership behaviours	1, 2a, 2b, 3a, 4a	5	1) PPE; 2a) Greater appreciation of others' perspectives, increased engagement, enhanced common identity, increased confidence; 2b) Increased leadership skills; 3a) Increased leadership behaviours, networking benefits, developed PDP's; 4a) Having launched a new initiative

[a] Facilitator	[d] Subjective numbers	[g] Participants	[i] Basline
[b] Subordinate	[e] Subjective descriptions	[h] Programme	[j] Post
[c] Superior	[f] Objective		[k] Post-post
			[l] Retrospective post
			[m] Pre
			[n] During
			[o] Retrospective pre

HEE SLR Codes – Outcomes 2/4

#	First Author and Publication Year	Measurements Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Outcome Metrics	Kirkpatrick Levels	Total	Reported Outcomes and Benefits
7	Malling (2009)	Self, sub[b], peer	Sub #, sub desc	Both	Pre, PP	1, 2b, 3a, 3b	4	1) PPE's; 2b) Increased knowledge; 3a) Increased leadership behaviours; 3b) Improved MSF pre and post	1, 2b, 3a	3	1) PPE; 2b) Increased knowledge; 3a) Increased leadership behaviours
8	Murdock (2009)	Self	Sub #, sub desc	Both	Pre, P	1, 2a, 2b, 3a, 3b	5	1) PPE's; 2a) Increased aspirations to lead; 2b) Increased leadership skills; 3a) Increased leadership behaviours; 3b) Have taken on a leadership role	1, 2a, 2b, 3a, 3b	5	1) PPE; 2a) Increased aspirations to lead; 2b) Increased leadership skills; 3a) Increased leadership behaviours; 3b) Having taken on a leadership role
9	Cherry (2010)	Peer, sup[c], fac	Sub desc	Part[h]	N/A	1, 2b	2	1) PPE's; 2b) Increased leadership skills	1, 2b, 3b	3	1) PPE's; 2b) Increased leadership skills; 3b) Research publications
10	Day (2010)	Self	Sub #	Part	Pre, base	2a, 2b, 3b	3	2a) Increased confidence; 2b) Increased knowledge and skills; 3b) Have taken on a leadership role, increased committee involvement, research publications	2a, 2b, 3a, 3b	4	2a) Increased confidence, 2b) Increased knowledge and skills, 3a) Positive impact on their careers; 3b) Having taken on a leadership role, increased committee involvement, research publications, increased academic rank, hospital administrative rank (chair or chief)
11	Kuo (2010)	Self, statistics	Sub #, sub desc, obj	Both	P, PP	1, 3a, 3b	3	1) PPE's; 3a) Positive impact on one's career; 3b) Awards won, grants earned, and research publications, have taken on a leadership role, promotions	1, 2a, 2b, 3a, 3b	5	1) PPE's; 2a) Increased aspirations to lead; 2b) Increased leadership competence; 3a) Positive impact on their careers; 3b) Awards won, grants earned, and research publications, having taken on leadership roles, promotions
12	Edmonstone (2011)	Self, sub, peer, sup	Sub desc, obj	Both	Pre, P	1, 3b, 4a, 4b	4	1) PPE's; 3b) Colleagues' feedback on behaviour changes, promotions; 4a) Policy changes; 4b) Having implemented action learning projects; Other) Having joined a mentoring network	1, 2a, 2b, 3b, 4a, 4b	6	1) PPE's; 2a) Increased self-awareness, increased resilience, increased engagement; 2b) Developing interpersonal and networking skills; 3b) Colleagues' feedback on behaviour changes, promotions, 4a) Policy changes, developed organisational capacity; 4b) Implementing action learning projects; Other) Having joined a mentoring network

[a] Facilitator	[d] Subjective numbers	[g] Participants	[i] Basline
[b] Subordinate	[e] Subjective descriptions	[h] Programme	[j] Post
[c] Superior	[f] Objective		[k] Post-post
			[l] Retrospective post
			[m] Pre
			[n] During
			[o] Retrospective pre

HEE SLR Codes – Outcomes 3/4

	Measurements										
#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Outcome Metrics	Kirkpatrick Levels	Total	Reported Outcomes and Benefits
13	Sanfey (2011)	Self	Sub #, sub desc, obj	Part	Pre, P, PP	1, 2a, 2b, 3a, 3b	5	1) PPE's; 2a) Increased aspirations to lead; 2b) Increased knowledge and skills; 3a) Increased leadership behaviours; 3b) Promotions	1, 2a, 2b, 3a, 3b	5	1) PPE's; 2a) Increased aspirations to lead, increased self-awareness, increased leadership self-identity; 2b) Increased knowledge and skills; 3a) Increased leadership beahviours, networking benefits; 3b) Promotions
14	Bearman (2012)	Self	Sub #	Prog	P	1	1	1) PPE's	1	1	1) PPE's
15	Shah (2013)	Self	Sub desc	Prog	P	1	1	1) PPE's; (the others were unclear)	1, 2a, 2b, 3a	4	1) PPE's; 2a) Increased engagement, 2b) Increased knowledge and skills, 3a) Increased leadership behaviours
16	Ten Have (2013)	Peer, fac	Obj	Part	Pre, PP	3b	1	3b) Improved MSF pre and post	3b	1	3b) Improved MSF pre and post
17	Vimr (2013)	Self	Sub #, sub desc	Both	P	1, 2a, 3a, 4b	4	1) PPE's; 2a) Improved self-awareness; 3a) Increased leadership behaviours; 4b) Having implemented action learning projects	1, 2a, 3a, 4b	4	1) PPE's; 2a) Improved self-awareness, developed a systems view; 3a) Increased leadership behaviours; 4b) Having implemented action learning projects
18	Blumenthal (2014)	Self	Sub #, sub desc	Both	P	1, 2a, 2b	3	1) PPE's; 2a) Increased confidence; 2b) Increased knowledge and skills	1, 2a, 2b	3	1) PPE's; 2a) Increased confidence, increased self-awareness, increased awareness of different leadership styles, increased interest in further training; 2b) Increased knowledge and skills

[a] Facilitator	[d] Subjective numbers	[g] Participants	[i] Basline
[b] Subordinate	[e] Subjective descriptions	[h] Programme	[j] Post
[c] Superior	[f] Objective		[k] Post-post
			[l] Retrospective post
			[m] Pre
			[n] During
			[o] Retrospective pre

HEE SLR Codes – Outcomes 4/4

		Measurements									
#	First Author and Publication Year	Raters	Type of Data	What	When	Kirkpatrick Levels	Total	Outcome Metrics	Kirkpatrick Levels	Total	Reported Outcomes and Benefits
19	Dickey (2014)	N/A	Sub desc	Prog	N/A	1	1	1) Authors' perceptions of programme strengths	1, 2b	2	1) Authors' perceptions of programme strengths; 2b) Developed negotiation skills
20	MacPhail (2014)	Self, sup	Sub desc, obj	Both	PP	1, 2a, 2b, 3a, 3b	5	1) PPE's; 2a) Increased aspirations to lead, increased leadership capacity; 2b) Increased leadership knowledge and skills; 3a) Increased leadership behaviours; 3b) Have taken on a leadership role, retention, promotions.	1, 2a, 2b, 3a, 3b	5	1) PPE's; 2a) Increased aspirations to lead, increased leadership capacity, plan to change their approach to patient care; 2b) Increased leadership knowledge and skills, developed ideas for improving patient care; 3a) Increased leadership behaviours; 3b) Have taken on a leadership role, retention, promotions.
21	Satiani (2014)	Self, sup	Sub desc	Both	D[o], P	1, 2b, 3b	3	1) PPE's; 2b) Increased leadership skills; 3b) Supervisors' ratings of increased leadership behaviour	1, 2a, 2b, 3a, 3b	5	1) PPE's; 2a) Increased self-awareness and confidence; 2b) Increased leadership skills, increased negotiation skills, developed interpersonal skills; 3a) Networking benefits; 3b) Supervisors' ratings of increased leadership skill levels and changes in behaviour.
22	Nakanjako (2015)	Self	Sub desc, obj	Part	P	2b, 3a, 3b, 4b	4	2b) Increased leadership skills; 3a) Have taken on more responsibility; 3b) Retention, awards won; 4b) Having implemented action learning projects	2b, 3a, 3b, 4a, 4b	5	2b) Increased leadership skills, increased leadership capability; 3a) Have taken on more responsibility; 3b) Retention, awards won, research publications; 4a) General organisational benefits, increased organisational capacity; 4b) Having implemented action learning projects, used innovative approaches to improve healthcare delivery
23	Patel (2015)	Self, fac	Sub #, sub desc	Both	Pre, P	1, 2a, 2b, 3b, 4b	5	1) PPE's; 2a) Increased confidence, increased aspirations to lead; 2b) Increased leadership knowledge; 3b) Have taken on a leadership role; 4b) Having implemented an action learning project	1, 2a, 2b, 3b, 4b	5	1) PPE's; 2a) Increased confidence, increased aspirations to lead; 2b) Increased leadership knowledge; 3b) Have taken on a leadership role, 4b) Having implemented an action learning project
24	Fernandez (2016)	Self	Sub #, sub desc, obj	Both	P, PP, Retro Pre[o]	1, 2a, 2b, 3a, 3b, 4b	5	1) PPE's; 2a) Increased confidence; 2b) Increased leadership skills; 3a) Increased leadership behaviours; 3b) Promotions; 4b) Self-reports of providing better healthcare to patients	1, 2a, 2b, 3a, 3b, 4b	6	1) PPE's; 2a) Increased confidence; 2b) Increased leadership skills, increased communication and teamwork skills; 3a) Increased leadership behaviours, have taken on more responsibility, positive impact on their careers; 3b) Promotions; 4b) Self-reports of providing better healthcare to patients
25	Pradarelli (2016)	Self, sub, peer, sup	Sub desc	Both	Base, P	1, 2b, 3a	3	1) PPE's; 2b) Increased knowledge and skills; 3a) Positive impact on their careers	1, 2a, 2b, 3a	4	1) PPE's; 2a) Increased self-awareness and confidence; 2b) Increase knowledge and skills; 3a) Positive impact on their careers

[a] Facilitator	[d] Subjective numbers	[g] Participants	[i] Basline
[b] Subordinate	[e] Subjective descriptions	[h] Programme	[j] Post
[c] Superior	[f] Objective		[k] Post-post
			[l] Retrospective post
			[m] Pre
			[n] During
			[o] Retrospective pre

Appendix E: Results of the Bivariate Statistical Analysis

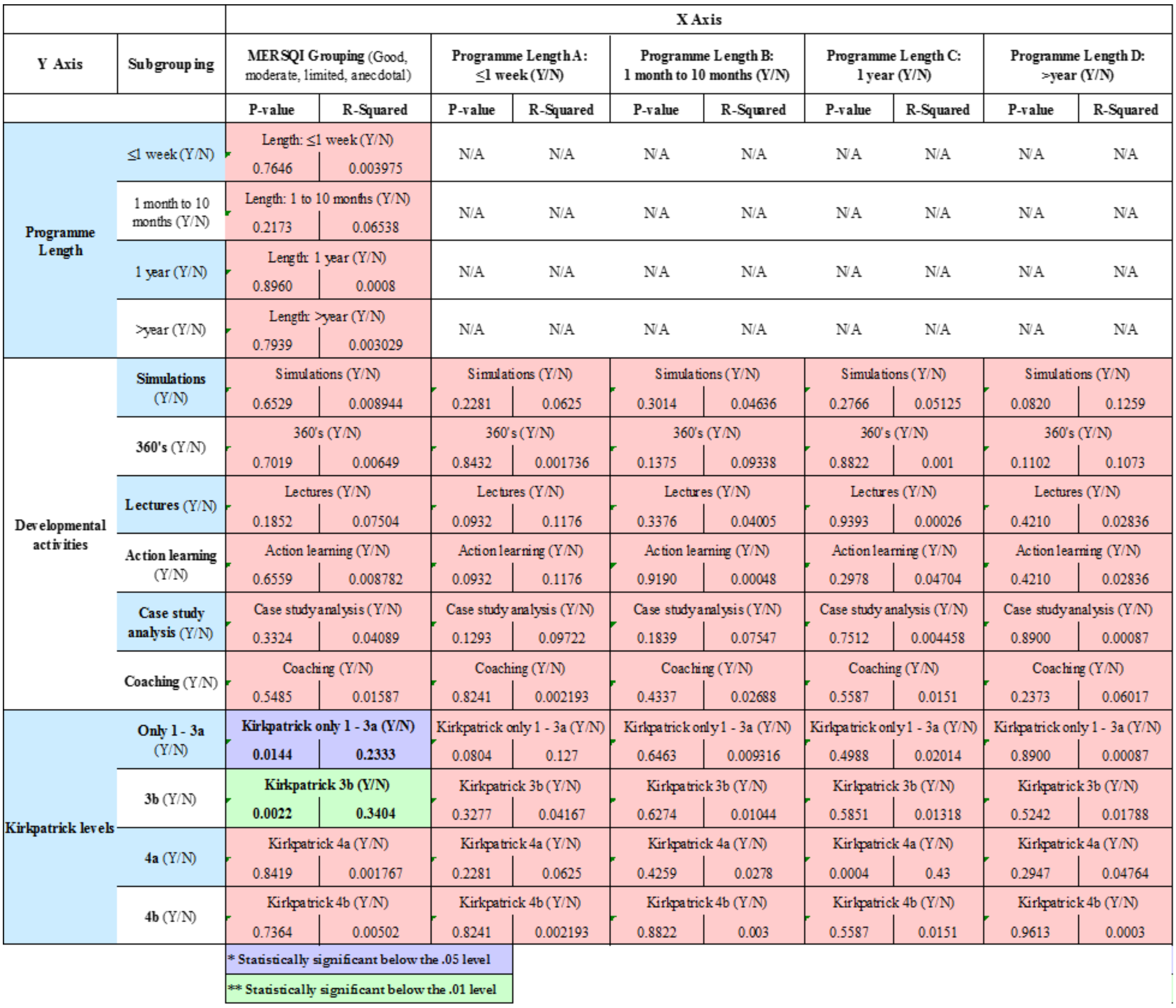
Bivariate Linear Regression Results: MERSQI and Length (1/2)

		X Axis									
Y Axis	**Subgrouping**	**MERSQI Grouping** (Good, moderate, limited, anecdotal)		**Programme Length A: ≤1 week (Y/N)**		**Programme Length B: 1 month to 10 months (Y/N)**		**Programme Length C: 1 year (Y/N)**		**Programme Length D: >year (Y/N)**	
		P-value	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**
Data type	Qualitative only (Y/N)	**Qualitative only (Y/N)***		Qualitative only (Y/N)		Qualitative only (Y/N)		Qualitative only (Y/N)		Qualitative only (Y/N)	
		0.0049	**0.2962**	0.6836	0.007353	0.4530	0.02473	0.2978	0.04704	0.7558	0.004291
	Quantitative only (Y/N)	**Quantitative only (Y/N) ****		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)	
		0.002	**0.3463**	0.>0.9999	7.70E-34	0.8432	0.001738	0.8241	0.002195	0.2947	0.04764
	Both (Y/N)	Both (Y/N)		Both (Y/N)		Both (Y/N)		Both (Y/N)		Both (Y/N)	
		0.8605	0.00137	0.7036	0.00641	0.5893	0.01289	0.4307	0.02724	0.2564	0.05565
MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		**MD's only (Y/N)**		MD's only (Y/N)	
		0.6243	0.0106	0.3277	0.04167	0.6274	0.01044	**0.0115**	**0.2473**	0.0804	0.129
Methodology	Case study (Y/N)	Case study (Y/N)		Case study (Y/N)		Case study (Y/N)		Case study (Y/N)		Case study (Y/N)	
		0.1734	0.07903	0.8484	0.001623	0.4039	0.0305	0.1323	0.09577	0.8022	0.002785
Methods	**Questionnaire** (Y/N)	Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)	
		0.0795	0.1277	>0.9999	0	0.8432	0.001738	0.8241	0.002195	0.7956	0.002978
	Interviews (Y/N)	Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)	
		0.5485	0.01587	0.8241	0.002193	0.4337	0.02688	0.3196	0.04308	0.2373	0.06017
Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)	
		0.7286	0.006153	0.1073	0.1246	0.8669	0.001441	0.5587	0.0151	0.4760	0.02573
Selection criteria	Applied and selected (Y/N)	Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)	
		0.0719	0.134	0.1737	0.07895	0.2766	0.05125	0.0942	0.119	0.9613	0.0003
	Nominated (Y/N)	Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)	
		0.3324	0.04089	0.5242	0.01786	0.6719	0.0079	0.4988	0.02014	0.8900	0.00087
	Volunteered (Y/N)	Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)	
		0.4174	0.02879	0.7956	0.002976	0.6343	0.03	0.2373	0.06017	0.1102	0.1073
Faculty	Internal (Y/N)	Internal (Y/N)		Internal (Y/N)		Internal (Y/N)		Internal (Y/N)		Internal (Y/N)	
		0.7087	0.006184	0.6719	0.007937	0.6719	0.007939	0.0859	0.123	0.0217	0.2089
	Mixed (Y/N)	Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)	
		0.2173	0.06538	0.4259	0.02778	0.5295	0.01743	0.4337	0.02688	0.6343	0.03
Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)	
		0.1990	0.07069	0.6719	0.007937	0.6463	0.009316	0.1828	0.07583	0.1881	0.07409

*** Statistically significant below the .05 level**

**** Statistically significant below the .01 level**

Bivariate Linear Regression Results: MERSQI and Length (2/2)



		X Axis									
Y Axis	Subgrouping	MERSQI Grouping (Good, moderate, limited, anecdotal)		Programme Length A: ≤1 week (Y/N)		Programme Length B: 1 month to 10 months (Y/N)		Programme Length C: 1 year (Y/N)		Programme Length D: >year (Y/N)	
		P-value	R-Squared	P-value	R-Squared	P-value	R-Squared	P-value	R-Squared	P-value	R-Squared
Programme Length	≤1 week (Y/N)	Length: ≤1 week (Y/N)		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		0.7646	0.003975								
	1 month to 10 months (Y/N)	Length: 1 to 10 months (Y/N)		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		0.2173	0.06538								
	1 year (Y/N)	Length: 1 year (Y/N)		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		0.8960	0.0008								
	>year (Y/N)	Length: >year (Y/N)		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		0.7939	0.003029								
Developmental activities	Simulations (Y/N)	Simulations (Y/N)		Simulations (Y/N)		Simulations (Y/N)		Simulations (Y/N)		Simulations (Y/N)	
		0.6529	0.008944	0.2281	0.0625	0.3014	0.04636	0.2766	0.05125	0.0820	0.1259
	360's (Y/N)	360's (Y/N)		360's (Y/N)		360's (Y/N)		360's (Y/N)		360's (Y/N)	
		0.7019	0.00649	0.8432	0.001736	0.1375	0.09338	0.8822	0.001	0.1102	0.1073
	Lectures (Y/N)	Lectures (Y/N)		Lectures (Y/N)		Lectures (Y/N)		Lectures (Y/N)		Lectures (Y/N)	
		0.1852	0.07504	0.0932	0.1176	0.3376	0.04005	0.9393	0.00026	0.4210	0.02836
	Action learning (Y/N)	Action learning (Y/N)		Action learning (Y/N)		Action learning (Y/N)		Action learning (Y/N)		Action learning (Y/N)	
		0.6559	0.008782	0.0932	0.1176	0.9190	0.00048	0.2978	0.04704	0.4210	0.02836
	Case study analysis (Y/N)	Case study analysis (Y/N)		Case study analysis (Y/N)		Case study analysis (Y/N)		Case study analysis (Y/N)		Case study analysis (Y/N)	
		0.3324	0.04089	0.1293	0.09722	0.1839	0.07547	0.7512	0.004458	0.8900	0.00087
	Coaching (Y/N)	Coaching (Y/N)		Coaching (Y/N)		Coaching (Y/N)		Coaching (Y/N)		Coaching (Y/N)	
		0.5485	0.01587	0.8241	0.002193	0.4337	0.02688	0.5587	0.0151	0.2373	0.06017
Kirkpatrick levels	Only 1 - 3a (Y/N)	**Kirkpatrick only 1 - 3a (Y/N)**		Kirkpatrick only 1 - 3a (Y/N)		Kirkpatrick only 1 - 3a (Y/N)		Kirkpatrick only 1 - 3a (Y/N)		Kirkpatrick only 1 - 3a (Y/N)	
		0.0144	**0.2333**	0.0804	0.127	0.6463	0.009316	0.4988	0.02014	0.8900	0.00087
	3b (Y/N)	**Kirkpatrick 3b (Y/N)**		Kirkpatrick 3b (Y/N)		Kirkpatrick 3b (Y/N)		Kirkpatrick 3b (Y/N)		Kirkpatrick 3b (Y/N)	
		0.0022	**0.3404**	0.3277	0.04167	0.6274	0.01044	0.5851	0.01318	0.5242	0.01788
	4a (Y/N)	Kirkpatrick 4a (Y/N)		Kirkpatrick 4a (Y/N)		Kirkpatrick 4a (Y/N)		Kirkpatrick 4a (Y/N)		Kirkpatrick 4a (Y/N)	
		0.8419	0.001767	0.2281	0.0625	0.4259	0.0278	0.0004	0.43	0.2947	0.04764
	4b (Y/N)	Kirkpatrick 4b (Y/N)		Kirkpatrick 4b (Y/N)		Kirkpatrick 4b (Y/N)		Kirkpatrick 4b (Y/N)		Kirkpatrick 4b (Y/N)	
		0.7364	0.00502	0.8241	0.002193	0.8822	0.003	0.5587	0.0151	0.9613	0.0003

* Statistically significant below the .05 level

** Statistically significant below the .01 level

Bivariate Linear Regression Results: Kirkpatrick Outcomes Levels (1/2)

		X Axis							
Y Axis	**Subgrouping**	**Kirkpatrick Levels: 1 - 3a only (Y/N)**		**Kirkpatrick 3b Behaviour (objective) (Y/N)**		**Kirkpatrick 4a** Organisational (Y/N)		**Kirkpatrick 4b** Benefit to Patients (Y/N)	
		P-value	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**
Data type	Qualitative only (Y/N)	**Qualitative only (Y/N) ****		**Qualitative only (Y/N)**		Qualitative only (Y/N)		Qualitative only (Y/N)	
		0.0068	**0.278**	**0.0129**	**0.2404**	0.6836	0.007355	0.3767	0.03414
	Quantitative only (Y/N)	Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)	
		0.2428	0.05883	0.5333	0.01711	0.2281	0.0627	0.1737	0.07897
	Both (Y/N)	Both (Y/N)		**Both (Y/N)**		Both (Y/N)		**Both (Y/N)**	
		0.1292	0.09724	**0.0428**	**0.1669**	0.5674	0.01444	**0.0491**	**0.16**
MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		**MD's only (Y/N)**		MD's only (Y/N)	
		0.4878	0.02118	>0.9999	1.37E-35	**0.0011**	**0.377**	0.5851	0.01318
Methodology	Case study (Y/N)	Case study (Y/N)		Case study (Y/N)		Case study (Y/N)		Case study (Y/N)	
		0.9457	0.00023	0.7547	0.004331	0.2443	0.05846	0.2156	0.06587
Methods	**Questionnaire** (Y/N)	Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)	
		0.0804	0.129	0.3277	0.04169	>0.9999	4.82E-37	0.8241	0.002195
	Interviews (Y/N)	Interviews (Y/N)		Interviews (Y/N)		**Interviews (Y/N)**		Interviews (Y/N)	
		0.7512	0.00448	0.5851	0.01318	**0.0359**	**0.1778**	0.5587	0.0151
Sample size (# of participants)		**Sample size (# of participants) ***		Sample size (# of participants)		**Sample size (# of participants)**		Sample size (# of participants)	
		0.0407	**0.1934**	0.8519	0.001787	**0.0115**	**0.2791**	0.8041	0.00315
Selection criteria	Applied and selected (Y/N)	Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)	
		0.4988	0.02014	0.1958	0.07166	0.3699	0.03511	0.5587	0.0151
	Nominated (Y/N)	Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)	
		0.3618	0.0363	0.4878	0.02118	0.5242	0.01788	0.7512	0.004458
	Volunteered (Y/N)	Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)	
		0.3047	0.04575	0.6719	0.007939	0.2947	0.04764	0.9613	0.0003
Faculty	Internal (Y/N)	Internal (Y/N)		Internal (Y/N)		Internal (Y/N)		Internal (Y/N)	
		0.3224	0.0426	0.8630	0.001325	0.1292	0.09724	0.4988	0.02014
	Mixed (Y/N)	Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)	
		0.1719	0.07961	0.1881	0.07409	0.2281	0.0627	0.9190	0.00048
Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)	
		0.0547	0.1514	0.4878	0.02118	0.5242	0.01788	0.4988	0.02014

* Statistically significant below the .05 level

** Statistically significant below the .01 level

Bivariate Linear Regression Results: Kirkpatrick Outcomes Levels (2/2)

		X Axis							
Y Axis	**Subgrouping**	**Kirkpatrick Levels: 1 - 3a only (Y/N)**		**Kirkpatrick 3b Behaviour (objective) (Y/N)**		**Kirkpatrick 4a** Organisational (Y/N)		**Kirkpatrick 4b** Benefit to Patients (Y/N)	
		P-value	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**
Programme Length	≤1 week (Y/N)	Length: ≤1 week (Y/N)		Length: ≤1 week (Y/N)		Length: ≤1 week (Y/N)		Length: ≤1 week (Y/N)	
		0.0804	0.129	0.3277	0.04169	0.2281	0.0627	0.8241	0.002195
	1 month to 10 months (Y/N)	Length: 1 to 10 months (Y/N)		Length: 1 to 10 months (Y/N)		Length: 1 to 10 months (Y/N)		Length: 1 to 10 months (Y/N)	
		0.6463	0.009316	0.6274	0.01044	0.4259	0.0278	0.8822	0.003
	1 year (Y/N)	Length: 1 year (Y/N)		Length: 1 year (Y/N)		**Length: 1 year (Y/N)**		Length: 1 year (Y/N)	
		0.4988	0.02014	0.5851	0.01318	**0.0004**	**0.43**	0.5587	0.0151
	>year (Y/N)	Length: >year (Y/N)		Length: >year (Y/N)		Length: >year (Y/N)		Length: >year (Y/N)	
		0.8900	0.00087	0.5242	0.01788	0.2947	0.04764	0.9613	0.0003
Developmental activities	**Simulations** (Y/N)	Simulations (Y/N)		Simulations (Y/N)		Simulations (Y/N)		Simulations (Y/N)	
		0.1839	0.07547	0.2516	0.05673	0.8432	0.001738	0.8822	0.001
	360's (Y/N)	360's (Y/N)		360's (Y/N)		360's (Y/N)		360's (Y/N)	
		0.6463	0.009316	0.7466	0.00465	0.2281	0.0627	0.0780	0.1291
	Lectures (Y/N)	Lectures (Y/N)		Lectures (Y/N)		Lectures (Y/N)		Lectures (Y/N)	
		0.4888	0.02109	0.5044	0.01963	0.6836	0.007355	0.0378	0.1746
	Action learning (Y/N)	Action learning (Y/N)		Action learning (Y/N)		**Action learning (Y/N)**		**Action learning (Y/N)**	
		0.2544	0.0561	0.8681	0.001227	**0.0085**	**0.2649**	**0.0010**	**0.3826**
	Case study analysis (Y/N)	Case study analysis (Y/N)		Case study analysis (Y/N)		Case study analysis (Y/N)		Case study analysis (Y/N)	
		0.3224	0.0426	0.8630	0.001325	0.5242	0.01788	0.7512	0.004458
	Coaching (Y/N)	Coaching (Y/N)		Coaching (Y/N)		**Coaching (Y/N)**		**Coaching (Y/N)**	
		0.4988	0.02014	0.5851	0.01318	**0.0359**	**0.1778**	**0.0035**	**0.3154**
Kirkpatrick levels	**4a** (Y/N)	N/A	N/A	Kirkpatrick 4a (Y/N)		N/A	N/A	N/A	N/A
				>0.9999	0.1.156e-035				
	4b (Y/N)	N/A	N/A	Kirkpatrick 4b (Y/N)		**Kirkpatrick 4b (Y/N)**		N/A	N/A
				0.1958	0.07166	**0.0359**	**0.1778**		

*** Statistically significant below the .05 level**

**** Statistically significant below the .01 level**

Bivariate Linear Regression Results: Developmental Activities (1/2)

		X Axis											
Y Axis	**Subgrouping**	**Simulations** (Y/N)		**360's** (Y/N)		**Lectures** (Y/N)		**Action Learning** (Y/N)		**Case Study Analysis** (Y/N)		**Coaching** (Y/N)	
		P-value	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**
Data type	Qualitative only (Y/N)	Qualitative only (Y/N)		Qualitative only (Y/N)		Qualitative only (Y/N)		Qualitative only (Y/N)		Qualitative only (Y/N)		Qualitative only (Y/N)	
		0.9190	0.00047	0.4530	0.02473	0.2010	0.07003	0.7007	0.006544	0.1004	0.115	0.3767	0.03414
	Quantitative only (Y/N)	Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)		Quantitative only (Y/N)	
		0.4259	0.0278	0.4259	0.0278	0.0932	0.1178	0.0932	0.1178	0.1292	0.09724	0.1737	0.07897
	Both (Y/N)	Both (Y/N)		Both (Y/N)		Both (Y/N)		Both (Y/N)		Both (Y/N)		**Both (Y/N)**	
		0.8012	0.002814	0.1749	0.07855	0.8963	0.00077	0.3400	0.03966	0.7605	0.00414	**0.0491**	**0.16**
MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)		MD's only (Y/N)	
		0.1881	0.07409	0.7466	0.00465	0.8681	0.001227	0.5044	0.01963	0.863	0.001325	0.5851	0.01318
Methodology	Case study (Y/N)	Case study (Y/N)		Case study (Y/N)		**Case study (Y/N)**		Case study (Y/N)		Case study (Y/N)		Case study (Y/N)	
		0.4039	0.0305	0.4029	0.0305	**0.0326**	**0.1837**	0.6935	0.0069	0.9457	0.0004	0.7470	0.00463
Methods	**Questionnaire** (Y/N)	Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)		Questionnaire (Y/N)	
		0.8432	0.001738	0.4259	0.0278	0.1449	0.09009	0.1449	0.09009	0.0804	0.129	0.8241	0.002195
	Interviews (Y/N)	Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)		Interviews (Y/N)	
		0.2562	0.05568	0.1802	0.0769	0.9601	0.00013	0.9601	0.0003	0.2681	0.05305	0.3196	0.04308
Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)		Sample size (# of participants)	
		0.4306	0.03137	0.6385	0.01127	0.4962	0.02348	0.079	0.1465	0.1676	0.09301	0.2963	0.05441
Selection criteria	Applied and selected (Y/N)	Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)		Applied and selected (Y/N)	
		0.2766	0.05125	0.8822	0.0012	0.3767	0.03414	0.9393	0.00026	0.1828	0.07583	0.5587	0.0151
	Nominated (Y/N)	Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)		Nominated (Y/N)	
		0.6463	0.009316	0.1839	0.07547	0.2544	0.0561	0.2544	0.0561	0.3618	0.0363	0.7512	0.004458
	Volunteered (Y/N)	Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)		Volunteered (Y/N)	
		0.5443	0.0164	0.6343	0.03	0.4210	0.02838	0.4210	0.02838	0.1811	0.07409	0.2373	0.06017
Faculty	Internal (Y/N)	**Internal (Y/N) ***		Internal (Y/N)		Internal (Y/N)		Internal (Y/N)		**Internal (Y/N)**		Internal (Y/N)	
		0.0206	**0.2121**	0.1719	0.07961	0.1004	0.115	0.4888	0.02109	**0.0447**	**0.164**	0.4988	0.02014
	Mixed (Y/N)	Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)		Mixed (Y/N)	
		0.0547	0.1514	0.8432	0.001738	0.9190	0.00048	0.4530	0.02473	0.6719	0.007939	0.8822	0.00099
Location (In-house v external)		Location (In-house v external)		Location (In-house v external)		**Location (In-house v external)**		Location (In-house v external)		Location (In-house v external)		Location (In-house v external)	
		0.1719	0.07961	0.6719	0.007939	**0.0329**	**0.185**	0.2544	0.0561	0.0547	0.1514	0.7512	0.004458

*** Statistically significant below the .05 level**

**** Statistically significant below the .01 level**

Bivariate Linear Regression Results: Developmental Activities (2/2)

		X Axis											
Y Axis	**Subgrouping**	**Simulations** (Y/N)		**360's** (Y/N)		**Lectures** (Y/N)		**Action Learning** (Y/N)		**Case Study Analysis** (Y/N)		**Coaching** (Y/N)	
		P-value	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**	**P-value**	**R-Squared**
Programme Length	≤1 week (Y/N)	Length: ≤1 week (Y/N)		Length: ≤1 week (Y/N)		Length: ≤1 week (Y/N)		Length: ≤1 week (Y/N)		Length: ≤1 week (Y/N)		Length: ≤1 week (Y/N)	
		0.2281	0.0627	0.8432	0.001738	0.0932	0.1178	0.0932	0.1178	0.1292	0.09724	0.3767	0.03412
	1 month to 10 months (Y/N)	Length: 1 to 10 months (Y/N)		Length: 1 to 10 months (Y/N)		Length: 1 to 10 months (Y/N)		Length: 1 to 10 months (Y/N)		Length: 1 to 10 months (Y/N)		Length: 1 to 10 months (Y/N)	
		0.3014	0.04636	0.1375	0.09338	0.3376	0.04005	0.9190	0.00048	0.9190	0.00048	0.3767	0.03412
	1 year (Y/N)	Length: 1 year (Y/N)		Length: 1 year (Y/N)		Length: 1 year (Y/N)		Length: 1 year (Y/N)		Length: 1 year (Y/N)		Length: 1 year (Y/N)	
		0.2766	0.05125	0.8822	0.001	0.9393	0.00026	0.2978	0.04704	0.1839	0.07547	0.3767	0.03412
	>year (Y/N)	Length: >year (Y/N)		Length: >year (Y/N)		Length: >year (Y/N)		Length: >year (Y/N)		Length: >year (Y/N)		Length: >year (Y/N)	
		0.4210	0.02838	0.1102	0.1073	0.4210	0.02838	0.4210	0.02838	0.8900	0.00087	0.3767	0.03412
Developmental activities	**360's** (Y/N)	360's (Y/N)			N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		0.8432	0.001738										
	Lectures (Y/N)	Lectures (Y/N)		Lectures (Y/N)		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		0.3376	0.04003	0.453	0.02473								
	Action learning (Y/N)	Action learning (Y/N)		Action learning (Y/N)		**Action learning (Y/N)**		N/A	N/A	N/A	N/A	N/A	N/A
		0.3376	0.04003	0.3376	0.04005	**0.0007**	**0.4001**						
	Case study analysis (Y/N)	Case study analysis (Y/N)		Case study analysis (Y/N)		**Case study analysis (Y/N)**		Case study analysis (Y/N)		N/A	N/A	N/A	N/A
		0.6719	0.007939	0.6463	0.009316	**0.0068**	**0.278**	0.4888	0.02109				
	Coaching (Y/N)	Coaching (Y/N)		Coaching (Y/N)		Coaching (Y/N)		**Coaching (Y/N)**		Coaching (Y/N)		N/A	N/A
		0.8822	0.001	<0.0001	0.5616	0.9393	0.00026	**0.0378**	**0.1746**	0.7512	0.004458		

*** Statistically significant below the .05 level**

**** Statistically significant below the .01 level**

Appendix F: Descriptions of the Evaluation of the Best Quality Studies

What follows is a description of how the best calibre studies measured their programmes (two good evidence studies and four moderate evidence, six total).

The following are fuller descriptions of each of the **good evidence** studies:

Dannels et al. (2008): their experiment used a questionnaire featuring closed-ended selections relating to leadership positions and participation in leadership development (outcome Levels 2a, 2b, and 3b) on the first day of a leadership intervention (baseline) to responses several years later (post-post). Several iterations of participants' responses were compared to those of a large control group (468 people) who had applied to the intervention but were rejected. Statistically significant increases were reported in the experiment group for 12 of the 15 leadership indicators.

Ten Have et al. (2013): their quasi-experiment featured a validated instrument (the IDR Assessment Scale) of Likert scale questions completed by trained raters on ten quality indicators related to leading medical interdisciplinary ICU rounds. The ratings concerned developing a plan of care for patients and the process of deciding and communicating that plan. Participants were rated before, after, and six weeks following a one-day intervention that included workshops, followed by videotaped simulations with peer and expert feedback. Participants' results were compared to identical ratings using the same instrument of physicians of the same specialty who had not received the training. Conclusions are that the experiment group's scores had improved compared to their original scores and to those of the control group in eight of the ten indicators.

The following are fuller descriptions of the **moderate evidence** studies:

Malling et al. (2009): their quasi-experiment featured multi-source feedback (MSF) concerning technical, human, administrative, and citizenship behaviour skills and was offered at baseline and a year following a six-month leadership intervention. Participants and equal-calibre colleagues who had not participated in the intervention were evaluated and both groups had similar baseline scores. In the post-post ratings, the control groups' scores had remained constant. The experiment group's self-ratings had increased; however, the other raters' assessments of them had not increased from the original scores.

Day et al. (2010): their case study involved questionnaires and document analysis. The questionnaires were completed by participants of a year-long mentorship intervention to assess their confidence in their leadership skills. These were collected at baseline and more than a year following the programme for several years' worth of iterations. Their CVs were also compared at the same times, along with a balanced control group of surgeons who applied to the programme but were not accepted. Both sets of CVs were analysed for leadership role attainment and research productivity. Findings were that intervention participants reported an increase in confidence in 7/8 categories, with this increase being statistically significant in three categories. Furthermore, intervention participants' leadership roles and research productivity increased by two to three times as much as did those in the control group.

Kuo et al. (2010): their case study involved a three year-long residency programme and distributed questionnaires to participants following the programme and years later. They reported increased competence as a leader, a positive impact on their clinical skills, increased motivation, and a positive impact on long-term career goals. The evidence offered for the latter was measured by grants won, academic publications, attainment of leadership roles and promotions, and awards won, including compared to residents at the same university who had not taken the programme.

Patel et al. (2015): their action learning methodology offered residents a two-year leadership training intervention through quality improvement (QI) and patient safety (PS). Participants completed pre and post questionnaires and reported that the intervention increased their ability to lead QI/PS activities in the future, as well as their motivation to pursue leadership positions. Many implemented their action learning projects, which directly benefited their patients. Lastly, expert faculty rated them using a validated instrument regarding clinical scenarios, which resulted in a three to four-point increase on a scale of 15 following the intervention. Thus, there were reported outcome benefits at Levels 1, 2a, 2b, 3b, and 4b.

Appendix G: Resource Guides for the Second Conclusion Explored

Leskiw and Singh (2007)'s Six Key Factors for Effective Leadership Development

Six key factors were found to be vital for effective leadership development:

- a thorough needs assessment,
- the selection of a suitable audience,
- the design of an appropriate infrastructure to support the initiative,
- the design and implementation of an entire learning system,
- an evaluation system, and
- corresponding actions to reward success and improve on deficiencies.

Van Aerde's (2013) Best Practice Themes

Six themes of leadership development best practice plans:

- developing pervasive mentoring relationships,
- identifying and codifying leadership talent,
- enhancing high potentials' visibility,
- assigning action-oriented developmental activities,
- leadership development through teaching and,
- reinforcing an organisational culture of leadership development.

Gilpin-Jackson and Bushe's (2007) Characteristics of a Positive Training Transfer Environment

Five characteristics of a positive training transfer environment:

- social support from supervisors and/or peers
- adoption environment that is conducive to innovation
- continuity and maintenance, which means support for the long-term maintenance of learning application
- situational context – that participants have the opportunity to apply their skills in the workplace
- systemic forces, referring to learning norms and culture, along with available resources