Facing the Complexity Gap: Developing Leaders’ Reasoning Skills to Meet the Complex Task Demands of their Roles

Aiden M. A. Thornton, Bachelor of Arts (Social Science), Bachelor of Social Science (First Class Honours)

This thesis is presented for the degree of Doctor of Philosophy at The University of Western Australia

Business School, Department of Management and Organisations

2023
THESIS DECLARATION

I, Aiden M. A. Thornton, certify that:

This thesis has been substantially accomplished during enrolment in this degree. This thesis does not contain material which has been submitted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

In the future, no part of this thesis will be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree. This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text.

This thesis does not violate or infringe any copyright, trademark, patent, or other rights whatsoever of any person. The research involving human data reported in this thesis was assessed and approved by The University of Western Australia Human Research Ethics Committee. Approval #RA/4/1/6612 24 February 2014, 13 May 2015, 2 May 2016, 15 November 2016.

As per the research proposal, which was approved by The University of Western Australia, the author of this thesis undertook an internship at Lectica with the goal of understanding Lectica’s particular approach to cognitive development and measurement. Through this process, the author of this thesis eventually became a Certified Senior Lectical Analyst™ with over eight years of experience working with the Lectical Assessment System™ (LAS) and the Computerised Lectical Assessment System™ (CLAS). At no point was the author of this thesis an employee of Lectica and there was no commercial exchange for undertaking any of the analyses reported in this thesis. The author of this thesis has had no relationship with Lectica since January 2021.

This thesis does not contain work that I have published, nor work under review for publication.
Signed,

Aiden M. A. Thornton.
The 21st century is characterised by high levels of complexity. In order to influence organisational outcomes, leaders must enact leadership processes while contending with these complex conditions. This thesis aims to contribute to theory, empirical research, and practice pertaining to leaders’ ability to navigate organisational complexity. This aim was underpinned by three research questions which were addressed by four empirical studies.

Studies 1 and 2 address the first research question, which focuses on whether cognitive-developmental scores awarded by the Lectical Assessment System™ (LAS) and ego development scores awarded by the ego development scoring system, satisfy Kohlberg’s hard stage requirements (which define the primary form of complexity considered in this thesis). Hard stage requirements specify that developmental stages must be unidimensional, invariantly sequenced, qualitatively distinct, structured wholes, and hierarchically integrated. Hard stage requirements were operationalised for psychometric evaluation via the unidimensional Rasch model. Cognitive-developmental scores were provisionally demonstrated to satisfy all hard stage requirements. Ego development scores were demonstrated to satisfy the requirement of invariant sequence. However, they violated most other requirements even though they were tested on different samples and through various analytical procedures. Ego development scores seem to reflect a cumulative form of development that is more strongly related to the number of perspectives taken or unique words employed by test-takers. There are three potential implications. First, findings may suggest that one of the main theoretical postulates of cognitive-developmental theory has been successfully operationalised for the purpose of measurement. However, this does not appear to be the case for ego development. Second, findings may suggest that cognitive-developmental scores, but not ego development scores, may legitimately be used to make inferences about leaders’ ability to navigate complexity. Third, there are implications which are peculiar to cognitive and ego development, respectively.
Study 3 addresses the second research question, which focuses on whether the hierarchical complexity of leaders’ reasoning skills satisfies the task demands of their roles. Findings suggest a statistically significant increase in the hierarchical complexity of reasoning skills from mid-leaders to upper leaders to senior leaders but not from senior leaders to executive leaders. Findings also suggest a significant complexity gap between leaders’ reasoning skills and the task demands of their roles, particularly for senior and executive-level leaders.

Study 4 addresses the third research question, which focuses on whether the hierarchical complexity of leaders’ reasoning skills develops during participation in various leader development programs. Findings suggest leaders develop their reasoning skills, even though direct causal attributions could not be made. The rate of development appears to increase with more frequent and formative use of cognitive-developmental assessments but not with increased contact time. Given the size of the complexity gap and the rate at which hierarchical complexity appears to develop, leader development may need to be augmented by collective leadership development processes to reduce the impact of the complexity gap in individuals.
TABLE OF CONTENTS

Chapter 1: An Introduction to 21st Century Complexity, Leadership, and Hard Stage Developmental Models.................................................................1
   The Role of Leadership in the Context of 21st Century Complexity .................................................4
   The Intersection Between the Fields of Leadership and Complexity Science ..................................5
   Adult Development Theory as an Approach to Complexity ............................................................6
   Three Types of Adult Development Theories ....................................................................................9
   Formal Definitions of Adult Developmental Theories and Underlying Scales ...............................11
   The Importance of Hard Stages to the Field of Leadership .............................................................14
   Research Questions .........................................................................................................................19
   Contributions to Theory, Empirical Research, and Practice ............................................................23

Chapter 2: Two Psychometric Studies on Adult Development Scores and Hard Stage Requirements ...29
   Hard Stage Claims Made About Cognitive Development and Ego Development ......................30
   Contended Relation Between Cognitive Development and Ego Development ..........................30
   Distinguishing Between Hard Stages, Soft Stages, and Cultural Ages ........................................32

Study 1A: Cognitive-Developmental Scores Awarded by the Lectical Assessment System (LAS) ....34
   Approach to Cognitive Development .............................................................................................34
   Cognitive-Developmental Theory ..................................................................................................34
   Cognitive-Developmental Scale .....................................................................................................35
   Measuring Cognitive Development ...............................................................................................41
   Cognitive-Developmental Scoring Systems ....................................................................................42
   Cognitive-Developmental Scores ....................................................................................................44
   Evidence that Cognitive-Developmental Scores Satisfy Hard Stage Requirements ....................45
   Consistency Between LAS/HCSS Scores and Hard Stage Requirements .....................................46
   Review of Empirical Studies on LAS/HCSS Scores .....................................................................52
   Limitations of Previous Studies .......................................................................................................54

Introduction to the Unidimensional Rasch Model .......................................................................55

Rationale for Study 1A ......................................................................................................................59

Criteria for Study 1A .........................................................................................................................60

Method .............................................................................................................................................62
   Participants .....................................................................................................................................62
   Materials ........................................................................................................................................63
   Procedure .......................................................................................................................................65

Use of The Rasch Model to Analyse Lectical Phase Scores ...........................................................65

Analytical Approach ........................................................................................................................66

Results ...............................................................................................................................................68
   Step 1: Confirm if the Whole Dataset Could be Analysed Simultaneously ................................68
   Step 2: Summary Statistics ............................................................................................................72
   Step 3: Direct Tests of Hard Stage Requirements .........................................................................75

Evaluation of Criteria for Study 1A ..................................................................................................84
Study 1B: Ego Development Scores Awarded by the Ego Development Scoring System

Chapter 3: Additional Analyses Undertaken on Ego Development

Study 2A: Reanalysis of Earlier Ego Development Data
Study 2B: Scoring Ego Development Scoring Exemplars with a Cognitive-Developmental Scoring System

Correspondence Between Ego Development Stages and Cognitive-Developmental Stages ........ 168
The Ego Development Scoring System ............................................................................. 170
Logic Underpinning Study 2B ....................................................................................... 171
Using Lectical Phases as a Criterion Variable ............................................................... 172
Hypotheses for Study 2B .............................................................................................. 173
Method ......................................................................................................................... 174
Sample ......................................................................................................................... 174
Other Materials ............................................................................................................ 175
Procedure ...................................................................................................................... 176
Results ........................................................................................................................... 176
Hypothesis Tests for Study 2B ...................................................................................... 181
Discussion ...................................................................................................................... 181

Study 2C: Regression Analyses on the Cognitive-Developmental Scores and Ego Development Scores

Relation Between Ego Development and Cognitive-Developmental Scores ................. 185
Proxy Variables to Operationalise the Logic Underpinning Cultural Ages, Soft Stages, and Hard Stages ................................................................. 186
Operationalising the Logic Underpinning the Lectical Scale ......................................... 187
Operationalising the Logic Underpinning the Ego Development Scale ......................... 188
Hypotheses for Study 2C .............................................................................................. 188
Method ......................................................................................................................... 191
Sample ......................................................................................................................... 191
Materials ....................................................................................................................... 192
Procedure ...................................................................................................................... 193
Results ........................................................................................................................... 196
Hierarchical Clustering in the Data Structure ................................................................ 196
Intercorrelations Between Main Variables ..................................................................... 197
Sequential Regression on CLAS Scores ........................................................................ 198
Sequential Regression on SCT Rasch Scores ................................................................ 199
Comparison Between Sequential Multiple Regressions on CLAS Scores and SCT Rasch Scores ................................................................. 199
Hypotheses Tests for Study 2C ...................................................................................... 200
Discussion ...................................................................................................................... 201

Synthesised Discussion on Ego Development ................................................................ 204
Summary of Findings ..................................................................................................... 204
Potential Explanations ................................................................................................. 204
Addressing Potential Rebuttals from Ego Developmentalists ....................................... 207

Chapter 4: The Complexity Gap Between Leader Reasoning Skills and the Task Demands of Their Roles

Introduction to the Complexity Gap .............................................................................. 212
Philosophical Arguments.......................................................................................................................213
Arguments Informed by the Management and Leadership Literature ..................................................214
  Perceived Capability Gaps ......................................................................................................................215
  Inferred Capability Gaps .........................................................................................................................216
  Conceptual Capability Gaps ....................................................................................................................217
Arguments Informed by Developmental Psychology ...............................................................................219
Hypotheses for Study 3 ............................................................................................................................222
Method ....................................................................................................................................................223
  Participants ............................................................................................................................................223
  Materials ................................................................................................................................................224
  Procedure ..............................................................................................................................................225
Results ...................................................................................................................................................231
  Hierarchical Clustering in the Data Structure .......................................................................................231
  Relation between CLAS Scores and Management Layer .....................................................................232
  Model Fit ...............................................................................................................................................233
  Comparison Between Skill Complexity and Role Complexity Ranges ..................................................237
Hypothesis Tests for Study 3 .....................................................................................................................241
Discussion .............................................................................................................................................242

Chapter 5: Longitudinal Growth in the Hierarchical Complexity of Leaders’ Reasoning Skills ........247
Approach to Researching the Hierarchical Complexity of Leaders’ Reasoning Skills ..............................247
The Leader Development Literature .........................................................................................................249
The General Cognitive Literature .............................................................................................................250
Cognitive-Developmental Literature (with Longitudinal Designs) ............................................................251
Cognitive-Developmental Literature (with Interventions) ........................................................................252
The Process of Cognitive Development .....................................................................................................257
Rationale for Study 4 ..................................................................................................................................258
Hypotheses for Study 4 .............................................................................................................................259
Method ....................................................................................................................................................259
  Participants ............................................................................................................................................259
  Materials ................................................................................................................................................261
  Procedure ..............................................................................................................................................263
Results ...................................................................................................................................................265
  Hierarchical Clustering in the Data Structure .......................................................................................265
  Descriptive Statistics ...............................................................................................................................266
  Model Fit ...............................................................................................................................................268
  Moderation of the Intercept and by Demographic Variables (i.e., Sex, Sector, Language) ..................272
  Predictors of Growth in the Combined Sample ......................................................................................273
  Relative Rates of Growth Between Developmental Providers ..............................................................274
  Hypothesis Tests for Study 4 ...................................................................................................................278
Discussion .............................................................................................................................................278

Chapter 6: General Discussion ..................................................................................................................282
Implications for Theory and Empirical Research .....................................................................................288
Implications for ego development ..........................................................................................................288
Implications for cognitive development ............................................................... 292
Implications for Practice ..................................................................................... 295
Broader Implications for the Fields of Adult Development and Leadership .......... 301
Limitations and Future Research Directions .................................................... 305
  General considerations ...................................................................................... 305
  Considerations about Studies 1 and 2 ............................................................... 309
  Considerations about Studies 3 and 4 ............................................................... 311
Concluding Reflections ....................................................................................... 313
References ........................................................................................................... 314
Appendix A: Hard Stage Claims Made by Ego Developmentalists ....................... 343
Appendix B: The Lectical Scale ......................................................................... 354
Appendix C: Lectica's Privacy and Legal Statements .......................................... 357
Appendix D: Additional Rasch Analyses Conducted on the LDMA ..................... 362
Appendix E: Studies on the Reliability of the SCT ............................................ 367
Appendix F: Relation Between SCT Scores and Scores Based on Various Cognitive-Developmental Models .............................................................................. 371
Appendix G: Illustrative Example of an SCT in the Public Domain ...................... 373
Appendix H: Participant Information and Consent Forms .................................... 376
Appendix I: SCT Stems Analyzed in This Thesis .............................................. 389
Appendix J: Additional Rasch Analyses Conducted on the SCT ......................... 393
Appendix K: Person-Item Threshold Distribution for a Particular Version of the SCT ........................................................................................................ 398
Appendix L: Scoring the Ego Development Scoring System with the Lectical Assessment System (LAS) (Supplementary Analysis) ...................................................... 399
Appendix M: Syntax for all Analyses Presented in Chapter 4 ............................ 402
Appendix N: Abbreviated Complexity Gap Analyses Undertaken on Various Lectical Assessments .... 411
Appendix O: Syntax for all Analyses Presented in Chapter 5 ................................ 428
TABLE OF TABLES

Table 1 Theoretical Basis of Each Study, Summary of Current Knowledge, Identification of Key Knowledge Gaps, and Contributions to Theory, Empirical Research, and Practice ........................................24
Table 2 Truncated Version of the Lectical Scale Relevant to Leader Development ...................................37
Table 3 Examples of Conceptions of a Good Leader at Successive Complexity Levels ................................39
Table 4 Demographic Details of Leaders for Study 1A ........................................................................63
Table 5 Steps Applied to the Rasch Analysis of the LDMA ....................................................................67
Table 6 Facet Analysis on Four LDMA Scorers .....................................................................................68
Table 7 Response Distribution Across Lectical Phases 10c to 12b ..........................................................69
Table 8 Uncentralised Threshold Locations for LDMA Items ..................................................................70
Table 9 Item Fit Statistics for LDMA Items ............................................................................................74
Table 10 PCA on Residuals for LDMA Items ............................................................................................76
Table 11 LDMA Item Loadings on Principal Component 1 (Organised by Strength of Loading) ..............77
Table 12 Subtest Analysis on the Magnitude of Multidimensionality in LDMA Items ..............................78
Table 13 Descriptive Statistics for Uncentralised Thresholds (LDMA Items) ...........................................80
Table 14 Definitional Attributes of Ego Development Stages ................................................................96
Table 15 Demographic Details of Leaders for Study 1B .....................................................................120
Table 16 Data Structure of the SCT Sample ............................................................................................122
Table 17 Summary of Leader Development Programs Pertaining to Study 1B .......................................123
Table 18 Steps Applied to the Rasch Analysis of the SCT .....................................................................125
Table 19 Facet Analysis on Three SCT Scorers .....................................................................................126
Table 20 Intercorrelations Between Five Versions of the SCT Included in Study 1B .............................128
Table 21 Response Distribution Across Nine Ego Development Stages .................................................129
Table 22 Uncentralised Threshold Locations for Stems with Disordered Thresholds ............................130
Table 23 Uncentralised Threshold Locations for Stems with Disordered Thresholds in the Seven-Stage Model ..................................................................................................................132
Table 24 Stem Fit Statistics for the Stems which Misfit on at Least One Statistical Test ............................137
Table 25 Analyses Conducted on the Unidimensionality of the SCT ......................................................140
Table 26 Descriptive Statistics for Uncentralised Thresholds Between Ego Development Stages ............145
Table 27 Frequency Distribution by TPR: Percentage of Stems Scored at Successive Ego Development Stages ........................................................................................................................................158
Table 28 Descriptive statistics for TPR Distributions (Total N of Stems = 2160) ....................................161
Table 29 Means, Standard Deviations, and One-Way Analysis of Variance (ANOVA) on Mean Scores Between Successive TPRs ........................................................................................................162
Table 30 OVL Between Successive TPR Distributions ..........................................................................164
Table 31 Cook-Greuter’s and Torbert’s Mapping Between Ego Development Stages and Cognitive-Developmental Stages ........................................................................................................................................169
Table 32 Illustrative Example of Scoring the Stem “When a child will not join in group activities ...” ....171
Table 33 Correspondence Between Ego Development Stages and Lectical Levels for the Stem “When a child will not join in group activities ...” ..................................................................................172
Table 34 Descriptive statistics for Lectical Phase Scores awarded to Exemplars for Each Ego Development Stage ........................................................................................................................................178
Table 35 Means, Standard Deviations, Mean Differences, and One-Way Analysis of Variance (ANOVA) on Lectical Phase Scores Between Successive Ego Development Stages .....................................180
Table 36 Demographic Details of Leaders for Study 2C ......................................................................192
Table 37 Grammatic Persons Used to Develop the ‘Unique Number of Perspectives’ Variable ..............195
Table 38 Variability Attributable to Hierarchical Clustering in the Data Structure for Study 2C ............196
Table 39 Descriptive Statistics and Correlations Between Variables for Study 2C .................................197
Table 40 Summary of Sequential Regression on CLAS Scores and SCT Rasch Scores ..........................200
Table 41 Demographic Details of Leaders for Study 3 .......................................................................224
Table 42 Number of Lectical Assessments Administered by Management Layer ..................................225
Table 43 Thresholds Applied to Unique Word Count by Years of Education and Approximate Grade ....226
Table 44  General Lectical Management Layers 2 Through 5 Including Descriptors and Role Complexity Ranges..................................................................................................................228
Table 45 Descriptive Statistics and Correlations Between Lectical Assessments Administered at Test-Time 1 for Provider 8 ........................................................................................................................................230
Table 46 Hierarchical Clustering in the Data Structure for Study 3 ..........................................................................................................................232
Table 47 Fit Statistics for the Four Models Applied to Study 3 ........................................................................................................................................234
Table 48 Frequencies, Percentages, and Chi-Square Goodness-of-Fit Tests on CLAS Scores ........................................................................240
Table 49 Demographic Details of Leaders for Study 4 ........................................................................................................................................261
Table 50 Number of Lectical Assessments Administered by Test-Time for Study 4 .............................................................................................262
Table 51 Thresholds Applied to Unique Word Count by Years of Education and Approximate Grade ....264
Table 52 Variability Attributable to Hierarchical Clustering in the Data Structure for Study 4 ..............265
Table 53 Descriptive Statistics for CLAS Scores at the Four Test-Times Included in Study 4 ..........266
Table 54 Fit Indices for the Four Models Applied to Study 4 ........................................................................................................................................269
Table 55 Descriptive Statistics for CLAS Scores at Test-Times 1 and 2 for Provider 5 and 12.........275
Table 56 Descriptive statistics for CLAS Scores at Test-Times 1 and 2 by Provider ............................277
Table 57 Relation Between 95% Confidence Intervals Constructed Around Slopes Between Providers .277
Table 58 Comparison of Findings Between Cognitive Development and Ego Development ..........284
TABLE OF FIGURES

Figure 1  Three Types of Adult Developmental Theories and Underlying Scales: Cultural Ages, Soft Stages, and Hard Stages .......................................................................................................................... 13
Figure 2  Relation Between the Three Research Questions and Studies Undertaken ................................................. 22
Figure 3  Item-Person Map of Scores Awarded to Moral Reasoning Interviews (Dawson et al., 2005, p. 181) .................................................................................................................................................. 47
Figure 4  Category Probability Curves of Lecical Levels 7 Through 12 Identified in Scores Awarded to Moral Reasoning Interviews (Dawson et al., 2005, p. 184) .......................................................... 49
Figure 5  Item Equating Map of Scores Awarded to Moral Reasoning Interviews (Dawson et al., 2005, p. 185) ................................................................................................................................................ 51
Figure 6  Illustrative Example of Scores that Do Satisfy Hard Stage Requirements in a Rasch Analysis ...... 57
Figure 7  Illustrative Example of Scores that Do Not Satisfy Hard Stage Requirements in a Rasch Analysis 58
Figure 8  Person-Item Threshold Distribution for the LDMA ....................................................................................... 70
Figure 9  Threshold Probability Curve for LDMA Item 1 .......................................................................................... 71
Figure 10  Threshold Map of 5 LDMA Items from 10c to 12b .................................................................................. 72
Figure 11  Person-Item Location Distribution for the LDMA .................................................................................... 73
Figure 12  Item Characteristic Curve for LDMA Item 1 ......................................................................................... 75
Figure 13  Person-Item Threshold Distribution for the LDMA .................................................................................. 80
Figure 14  Category Probability Curve for LDMA Item 5 ....................................................................................... 82
Figure 15  Item Equating Graph for 5 LDMA Items ................................................................................................. 83
Figure 16  Item Equating Graph for Item 5 Only ....................................................................................................... 83
Figure 17  Common Item Equating Conducted on Five Versions of the SCT Showing Person Ability
(Vertical) and Mean SCT Stem (Horizontal) ........................................................................................................ 128
Figure 18  Person-Item Threshold Distribution for the SCT .................................................................................. 130
Figure 19  Illustrative Example of Threshold Probability Curve for SCT Stem 1 ................................................. 131
Figure 20  Threshold Map for the 55 Stems and the Location of Stages in the Seven-Stage Model 133
Figure 21  Illustrative Example of Item Characteristic Curve for Stem 44 from the SCT .............................. 134
Figure 22  Person-Item Location Distribution for the SCT Showing Person Ability (top) and Average Stem
Location (Bottom) Along the Logit Scale ........................................................................................................ 136
Figure 23  Illustrative Example of Item Characteristic Curve for Stem 14 from the SCT ................................. 138
Figure 24  Person-Item Threshold Distribution for the SCT .................................................................................. 144
Figure 25  Category Probability Curve for Stems 43 and 8 .............................................................................. 147
Figure 26  Category probably curve for SCT Stem 8 ......................................................................................... 147
Figure 27  Item Equating Graph for all 42 Stems ................................................................................................. 148
Figure 28  Distribution by TPR: Percentage of Stems Scored at Successive Ego Development Stages .... 156
Figure 29  Boxplot of TPR Distributions .............................................................................................................. 160
Figure 30  Boxplot of Lecical Phase Scores Awarded to Exemplars for Each Ego Development Stage 177
Figure 31  Three Scenarios: Increase, Statis, and Regression in Lecical Phase Scores .................................. 178
Figure 32  Exploratory Analysis: Boxplot of Lecical Phase Scores Awarded to Exemplars for Each Ego
Development Stage ........................................................................................................................................... 181
Figure 33  Mitigation of Conflating Factors Which May Have Impacted Earlier Research .......................... 186
Figure 34  Conceptual Relation between Proxies and CLAS Scores ................................................................. 189
Figure 35  Conceptual Relation between Proxies and SCT Scores .................................................................. 190
Figure 36  Average Lecical Score and Role Complexity by Management Layer ............................................. 222
Figure 37  Boxplot of CLAS Scores by Management Layer ............................................................................ 233
Figure 38  Piecewise Model Applied to CLAS Scores for Study 3 ................................................................. 237
Figure 39  Boxplot of CLAS Scores and Role Complexity Ranges by Management Layer ............................ 238
Figure 40  CLAS Score by Test-time 1 Through 4 for a Random Sample of 250 Persons ............................ 267
Figure 41  CLAS Score by Month of Administration Post Test-Time 1 for a Random Sample of 250 Persons ........................................................................................................................................ 267
Figure 42  Time-Variant Regression Model Applied to CLAS Scores .......................................................... 272
ACKNOWLEDGEMENTS

This research was supported by an Australian Government Research Training Program (RTP) Scholarship. I would like to extend my gratitude to Professor David V. Day from the Kravis Leadership Institute at Claremont McKenna College and Dr Theo Dawson from Lectica™ for their supervisory support. I acknowledge Dr Theo Dawson from Lectica, Professor Bill Torbert from Action Inquiry Associates and Boston College, Professor Sean Esbjörn-Hargens from MetaIntegral™ and John F. Kennedy University, and Dr Niki Vincent, formerly from The Leaders Institute of South Australia, for their generous provision of data. I would also like to thank Dr Theo Dawson for providing me with a thorough understanding of Lectica’s particular approach to cognitive development, the Lectical Assessment System™, and the Computerized Lectical Assessment System™. I would like to thank Professor David Andrich, Dr Ida Marais, and Dr Irene Styles from the University of Western Australia for your most excellent teaching of the Rasch model. I would also like to thank the Psychometrics Laboratory at the University of Western Australia for providing the RUMM2030 software, which enabled me to perform the psychometric analyses reported in this thesis (Andrich et al., 2015). I would like to thank Professor Gillian Yeo, Dr Sandra Kiffin-Petersen, and Ms Robyn Oliver from the University of Western Australia, who provided an incredible amount of support in times of great need. I would like to thank Professor Michael Commons from Harvard University and Professor Patrice Miller from Salem State University for supporting my interest in novel applications of The Model of Hierarchical Complexity to the science and practice of leadership over the recent past. I am truly indebted to Dr Roy Kong for his tireless support through the trials and tribulations of undertaking such an ambitious research project. Without your unconditional support, I would not have made it to the finish line. Last but certainly not least, I extend my deepest gratitude to my professional mentor, who will remain anonymous (for now). Thank you for reminding me of the many ways in which this thesis was more than merely an intellectual pursuit. Thank you for helping me to skilfully navigate the difficult decisions about the various paths I could have taken after completing this project. Veritas omnia vincit.
Chapter 1: An Introduction to 21st Century Complexity, Leadership, and Hard Stage Developmental Models

Leaders in the 21st century are confronted with an unprecedented level of complexity (Crutzen & Stoermer, 2000; Helbing, 2013; Rasmus, 2016; Taleb, 2012). A considerable body of literature describes the complex environmental, social, political, ethical, and organisational issues with which many leaders are currently contending (Gahan et al., 2016; Glenn & Florescu, 2018; IBM, 2010; Meadows et al., 2004; National Intelligence Council, 2012; Naughtin et al., 2022; Randers, 2012; World Economic Forum, 2022).

The complexity of social systems, including organisations, manifests in various ways. The U.S. Military initially proposed the term VUCA in the 1980s to characterise the volatile, uncertain, complex and ambiguous conditions in which strategic leadership needed to be developed (Barber, 1992; Bennis & Nanus, 1986). The term volatility was used to specify conditions which change rapidly; uncertainty to specify conditions which have weak predictability; complexity to specify conditions in which causal relations are difficult to establish; and ambiguity to specify conditions in which truth and meaning are unclear (Stiehm, 2002). The term VUCA is now used in a broad range of contexts to describe the complex conditions in which many 21st century leaders operate. The Global Financial Crisis in 2007 and the COVID-19 pandemic provide two salient examples among many.
In the context of systems science, complexity is typically conceived to be an emergent property of social systems (Holland, 1995; Von Bertalanffy, 1968; Weinberg, 2001). Complexity emerges as a function of the number of diverse elements of which social systems are comprised and the interdependencies between them (Gharajedaghi, 1999). As the interdependencies between elements increase in density, social systems become increasingly unstable and exhibit volatile, uncertain, complex, and ambiguous behaviour (Helbing, 2013). Because complexity emerges from properties inherent to social systems, there is a general trend for systems to increase in complexity. This is evident in how society has become increasingly complex as it has increased in size, used more advanced technology, and cooperated with other systems (Nolan & Lenski, 2008; Stewart, 2014). This general trend is apparent from the increased complexity ensuing from the Industrial Revolution (circa 1760) and the Great Acceleration beginning in the 1950s, through which an exponential increase in socio-economic activity across a broad range of indicators has been observed (Steffan et al., 2015).

In this context, it is no surprise that complexity has been amongst the most significant challenges experienced by organisational leaders for over a decade. In a survey of 1541 senior leaders across the globe, 60% reported that they were experiencing high or very high levels of complexity, 79% expected to experience a high or very high level of complexity over the coming years, and yet only 49% felt adequately prepared to navigate these expected levels of complexity (IBM, 2010). The major drivers of complexity included market factors, technology, macroeconomics, and people factors. Solutions suggested by the senior leaders included more creative forms of leadership, reinventing customer relationships, and greater operational agility. The report concluded by arguing that “We occupy a world that is connected on multiple dimensions, and at a deep level — a global system of systems. That means, among other things, that it is subject to systems-level failures, which require systems-level thinking about the effectiveness of its physical and digital infrastructures” (IBM, 2010, p. 3). In a more recent study, over 8000 Australian leaders reported facing a similar set of complex challenges, including market pressures, government
regulation, technological disruption, and the volatility and uncertainty of the external operating environment (Gahan et al., 2016).

More contemporary studies suggest that the digital era and the COVID-19 pandemic instantiate complex conditions that impact leadership. For example, a recent meta-analysis of 54 studies examined the impact of digitization on leadership at macro and micro scales (Cortellazzo et al., 2019). At the macroscale, the digital era was shown to create conditions in which leaders must foster new relationships with their organisations, adopt technology as a partial solution to navigating organisational complexity, grapple with inherently ethical dilemmas, and use digital technologies to influence society at large. At the microscale, the digital era was shown to create conditions in which leaders must embrace the complexity of C-level roles, cultivate new leadership and digital skills, and adopt new team leadership practices. Analogously to the impact of the digital era, the COVID-19 pandemic has also created social complexity that has had significant consequences for leadership. Based on analyses of various case studies, some researchers argue that COVID-19 surfaced two major ‘fault lines’ of leadership that would not have otherwise emerged: narcissism and ideological rigidity (Maak et al., 2021). Narcissism included leaders’ tendency to focus excessively on their needs for social approval and dominance. Ideological rigidity included the tendency to defend existing worldviews and beliefs at the expense of more adaptive behavioural responses.

At the scale of country-level leaders, a recent survey conducted by the World Economic Forum (2022) showed that 63% of respondents reported feeling worried or concerned about the outlook for the world, and 79% of respondents expected a consistently volatile landscape with multiple surprises or fractured trajectories which separated winners from losers. Over the coming two years, the most imminent threats to the world were considered to be extreme weather patterns, failure of climate action, livelihood crises, the erosion of social cohesion, the prevalence of infectious diseases such as COVID-19, and the deterioration of mental health. The report also explored the systemic interdependencies between economic, environmental, geopolitical, societal, and technological risks and surfaced the complex implications for leadership.
It seems reasonable to conclude that leaders must cultivate types of leadership that will enable them to keep step with current and future manifestations of complexity. The types of leadership that are well suited to 21st complexity may be informed by various practices, including compassion, evidence-based decision-making, adaptive leadership, sensemaking, emotional stability, and innovative types of communication (Dirani et al., 2020; Maak et al., 2021; Uhl-Bien, 2021).

**The Role of Leadership in the Context of 21st Century Complexity**

The field of leadership has enjoyed a history of more than a century and comprises different schools of thought about what leadership is, why it is important, and how it emerges. While a detailed review of the field of leadership is beyond the scope of this thesis, there has been a historical trend for conceptions of leadership to shift in emphasis from psychological traits (the 1920s-1950s), to behaviours (1950s-1960s), to environmental contingencies (1970s-1980s), to a more integrative combination of information-processing (1990s-2020s), charismatic (1980s-2020s), interpersonal (1980s-2020s), contextual (1990s-2020s), and evolutionary factors (2020s) over time (Antonakis & Day, 2018). This has led some scholars to assert that “there are almost as many definitions of leadership as there are persons who have attempted to define the concept” (Bass, 1990, p. 11). Despite the multitude of ways in which leadership is conceived, many schools of thought instantiate an enquiry into leadership as a process of social influence, which has a positive impact on the social systems in which it is nested (Antonakis & Day, 2018; Bass, 2008; Day, 2012). In this context, leadership is a process that can be enacted by individual leaders or collectives (Day et al., 2014). Individual leaders may be regarded as persons at any layer in an organisational hierarchy who practice the process of leadership (Day et al., 2014).

Because influence is a definitional attribute of leadership, leadership processes may be ideally positioned to positively influence organisational outcomes in the context of 21st century complexity. Leadership is correlated with, and sometimes causally related to, a variety of favourable outcomes for individuals, teams, and organisations (Ahearne et al., 2005; De Waal & Sivro, 2012;
Recent research has shown that leadership is related to individual outcomes, such as followers’ motives and task performance (Kehr et al., 2022; Zhang et al., 2022), as well as creativity and innovation (Knotts et al., 2022). Leadership has also been shown to relate to group outcomes such as fostering a positive ethical climate (Neubert et al., 2022), social competence and the empowerment of others (Gloria & Alison, 2022), and higher levels of team performance in certain contexts (Miao et al., 2019). Finally, leadership has been shown to relate to various organisational outcomes, such as organisational citizenship and culture (Karadağ, 2015; Newman et al., 2017), and may explain up to 45% of variability in organisational performance (Day & Lord, 1988).

The Intersection Between the Fields of Leadership and Complexity Science

While the role of leadership in relation to influencing general organisational outcomes has become increasingly well established, there have been relatively few attempts to integrate the field of leadership with the field of complexity science. In a review of the theoretical foci of articles published in The Leadership Quarterly between 2010 and 2019, only 1.9% of articles reflected an explicit focus on contextual, adaptive, and complexity-related leadership theory (Gardner et al., 2020). Arguably, there are three main points of intersection between the two fields: complexity leadership theory (Uhl-Bien et al., 2007), leader complexity theory (Hannah et al., 2009; Hooijberg et al., 1997; Lord et al., 2011), and adult development theory (McCaulley et al., 2006). According to complexity leadership theory, complexity is an emergent property of organisational systems that needs to be fostered so that organisations can match their internal complexity to that of their environmental context. This is accomplished by engaging networks of leaders in emergent self-organisation (Goldstein et al., 2010; Marion & Uhl-Bien, 2001; Uhl-Bien et al., 2007). Various examples of theorising about leader complexity have been presented in the literature, including the Leaderplex Model (Hooijberg et al., 1997), leader complexity (Hannah et al., 2009), requisite
complexity (Lord et al., 2011) and the leadership complexity model (Day & Lance, 2004). While a more detailed discussion of these models is provided in Chapter 4, the general postulate is that leaders must draw upon various psychological and social resources to adapt the complexity of their behavioural responses, so they are commensurate with the complexity of their environmental conditions. Behavioural complexity is positively related to leader effectiveness, and leader effectiveness emerges when there is a sound fit between leader and environmental complexity (Day & Lance, 2004; Denison et al., 1995). Leader effectiveness is typically defined as a process of mobilising and influencing followers to accomplish organisational goals by performing at a level that exceeds normal expectations (Yukl, 2006, 2008). Finally, adult development theory has been posited to characterise the qualitatively distinct stages of developmental complexity through which leaders and/or specific attributes of leaders might develop (McCauley et al., 2006). The general contention is that leaders who can perform at higher developmental stages are better equipped to navigate complexity.

While all three points of intersection may be worthy paths of pursuit, complexity leadership theory and theorising about leader complexity are relatively nascent. For example, complexity leadership theory has informed a few empirical studies and conceptual commentaries (Tourish, 2019), and theorising about leader complexity has only begun to be operationalised in the leadership literature (Day et al., 2009; Day & Lance, 2004). Adult development theory, however, provides a relatively mature conception of the stages of complexity through which leadership attributes can develop and therefore forms a valuable link between the fields of leadership and complexity science.

**Adult Development Theory as an Approach to Complexity**

Adult development theory was introduced to the field of leadership relatively recently, even though theories of adult development have been used in applied epistemology, education, psychotherapy, and other disciplines for considerably longer (Baldwin, 1905; Powers et al., 1983). In
the leadership literature, the possibility of leaders developing a more ‘complicated understanding’ was first raised in the 1980s. Complicated understanding was defined as leaders’ ability to grapple with complex organisational challenges by taking and coordinating diverse perspectives (Bartunek et al., 1983; Weathersby et al., 1982). McCauley et al. (2006) described how adult development theory could account for leaders’ ability to navigate complexity, such as adaptive organisational change.

Day et al. (2009) included adult development stages in their integrative conception of leader development alongside the constructs of leader identity and expertise. More recently, Brownlow’s (2022) doctoral dissertation presents a systematic review of published and unpublished literature on the relation between specific types of adult development stage theories and leadership, as well as developing and validating a questionnaire which is contended to measure a particular type of adult development stages. While there is interest in applications of adult development theory to account for leaders’ ability to navigate 21st complexity, the program of work required to accomplish this has been unclear to date.

Although multiple scholars have presented taxonomies of adult development theories (e.g., Demick & Andreoletti, 2003; Robinson, 2013), there is no universal agreement on how they should be classified. Despite this, the primary adult development theories currently employed in the field of leadership would arguably include constructive-development theory (Kegan, 1982), cognitive-developmental theory (Commons & Pekker, 2008; Dawson-Tunik, 2006; Fischer & Bidell, 2006), ego development theory (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 2013), and Gravesian stages of life conditions (Beck & Cowan, 1996; Cowan & Todorovic, 2003, 2005; Laloux, 2014). Constructive-developmental theory describes how the self traverses a sequence of stages toward higher levels of differentiation and integration (Kegan, 1982). Cognitive-developmental theory describes the way reasoning skills develop through stages of increasing hierarchical complexity in a variety of knowledge domains (Armon & Dawson, 1997; Commons & Pekker, 2008; Fischer & Bidell, 2006; Kitchener et al., 2006; Kohlberg, 1981; Piaget, 1954). Ego development theory describes how the self—or the worldview constructed by the self—develops through a sequence of stages towards
increasing complexity and ultimately, ego transcendence (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 2013). Gravesian stages describe how individuals and cultures grow through increasingly complex life conditions (Beck & Cowan, 1996; Cowan & Todorovic, 2003, 2005; Laloux, 2014). Some scholars use the term ‘constructive-developmental’ as a broad way of referring to several stage theories (Brownlow, 2022; McCauley et al., 2006). Strictly speaking, the term only applies to Kegan’s (1982) stages of the evolving self and has been adopted by some ego developmentalists in specific contexts e.g., Cook-Greuter (1999). This term does not necessarily generalise to cognitive-developmental and other types of stage theories throughout the adult development literature.

While the aforementioned theories claim to consist of stages that increase in complexity, relative few scholars appear to have provided quantitative empirical evidence to demonstrate that their respective stages reflect an increase in a particular form of complexity that is typically considered in developmental psychology, i.e., hierarchical complexity (Commons, Li, et al., 2014; Dawson et al., 2005; Dawson-Tunik, 2006; Wilson, 1989). This presents a considerable challenge for the field of leadership because many of the theories cited previously—and the assessments based on these theories—are used by leadership scholars and practitioners to make inferences about leaders’ ability to navigate complexity. However, the evidence to support these contentions may be questionable. For example, The Leadership Quarterly published a series of articles in which Kegan’s constructive-developmental theory and assessment were used to make inferences about leaders’ ability to navigate complexity despite the limited empirical evidence in support of the construct measured by the assessment (Bartone et al., 2007; McCauley & Palus, 2021; Russell & Kuhnert, 1992; Strang & Kuhnert, 2009; Valcea et al., 2011). Torbert and his associates argue that only leaders who perform above a particular stage of ego development are equipped to lead complex organisational transformations (Rooke & Torbert, 1998). Laloux (2014) advocates for various leadership practices, which are claimed to foster the development of individuals and organisations in service of their ability to navigate VUCA conditions. McCauley’s (2006) review of adult development theories mentioned above asserts, but does not demonstrate, that there is a direct correspondence
between Kegan’s constructive developmental stages (1982), Kohlberg’s (1969, 1981) cognitive-developmental theory, and Torbert’s ego development theory (Fisher et al., 1987; Torbert, 1987c, 1991a). Brownlow’s (2022) doctoral dissertation implicitly assumes, but again does not demonstrate, a direct correspondence between Piaget’s (1952) cognitive-developmental stages, Kohlberg’s (1969, 1981) stages of moral reasoning, Kegan’s constructive developmental stages (1982) and ego development stages (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 1991a), see for example Brownlow (2022, p. 19). Whereas Kohlberg’s stages of moral reasoning have been shown to consist of hierarchically integrated stages of complexity (Dawson, 2002a, 2003b; Dawson et al., 2005; Dawson & Gabrielian, 2003; Dawson et al., 2003), it is unclear whether these properties have been demonstrated in Kegan’s (1982) constructive-developmental stages and Torbert’s (Fisher et al., 1987; Torbert, 1987c) stages of ego development. In sum, several adult development theories and assessments based on those theories are being used in the field of leadership to make inferences about leaders’ ability to navigate complexity without the requisite degree of empirical validation. As a result, the field risks making inaccurate and/or potentially ethically questionable inferences about leaders (Messick, 1989).

Three Types of Adult Development Theories

Kohlberg introduced an important distinction between three types of adult development theories that are often conflated in the literature: cultural ages, soft stages, and hard stages (Kohlberg & Armon, 1984; Snarey et al., 1983). Cultural ages are periods through the human lifecycle, typify the age at which certain social roles are adopted, and often vary between national cultures. Soft stages account for the accumulation of meaning, affect, impulse control, etc., through the human lifecycle. Kohlberg’s conception of hard stages is a refinement of the formal properties that Piaget (1960, 1968, 1972) originally attributed to cognitive-developmental stages (Colby & Kohlberg, 1987; Kohlberg, 1969, 1975; Kohlberg & Armon, 1984). According to Kohlberg, hard stages must satisfy four requirements: they must form an invariant sequence, be qualitatively distinct,
form structured wholes, and reflect a sequence of hierarchical integrations. These three types of developmental theories may exhaustively account for those considered in developmental psychology.

By *invariant sequence*, Kohlberg argued that “These different structures form an invariant sequence, order, or succession in individual development. While cultural factors may speed up, slow down, or stop development, they do not change its sequence” (Kohlberg & Armon, 1984, p. 384-385). Hard stages occur in a single sequence across cultures and stages cannot be skipped or reversed.

By *qualitatively distinct*, Kohlberg argued that “Stages imply a distinction or qualitative difference in structures (modes of thinking) that still serve the same basic function (for example, intelligence) at various points in development” (Kohlberg & Armon, 1984, p. 384). Stages do not merely reflect a quantitative increase in an underlying construct (i.e., more of the same). Instead, each stage results from a transformation of the previous stage and gives rise to a qualitatively distinct structure.

By *structure of the whole*, Kohlberg argued that “Each of these different and sequential modes of thought forms a ‘structured whole’. A given stage response on a task does not just represent a specific response determined by knowledge and familiarity with that task or tasks similar to it; rather, it represents an underlying thought-organization. The implication is that various aspects of stage structures should appear as a consistent cluster of responses in development” (Kohlberg & Armon, 1984, p. 385). Hard stages are not merely idiosyncratic response categories but reflect deep structures in the mind-brain that give coherence to human performance. As a result, there is a tendency for a single stage to organise performance in a particular domain rather than for performance to be broadly distributed across stages. Performance coheres in a single Stage S because performance at Stages < S has been hierarchically integrated at Stage S, and performance at Stages > S has not yet unfolded.
By *hierarchical integrations*, Kohlberg argued that “Stages are hierarchical integrations. As noted, stages form an order of increasingly differentiated and integrated *structures* for fulfilling a common function. Accordingly, higher stages displace (or, rather, integrate) the structures found at lower stages (Piaget, 1960)” (Kohlberg & Armon, 1984, p. 385). Hierarchical integration is the underlying logic determining the sequence in which hard stages occur. Stage S+1 is considered to be a hierarchical integration of its predecessor Stage S if and only if Stage S+1 is defined in terms of two or more elements of Stage S, it organises and transforms the elements of Stage S, and it does so in a non-arbitrary way (Commons & Pekker, 2008). This results in an emergent new structure, which enables performance that cannot be accomplished at Stage S alone. *Vertical development* is a colloquial term that refers to hard stage models in the practitioner literature (Petrie, 2014).

**Formal Definitions of Adult Developmental Theories and Underlying Scales**

While Kohlberg’s (1984; 1983) distinctions provide a useful heuristic, they also present two challenges. First, they provide general descriptions but do not satisfy the requirements of formal scientific definitions (MacKenzie, 2003). Definitions must specify theoretical constructs in unambiguous terms, enabling distinctions between one construct and another. Second, the descriptions are provided at the scale of theory, but they do not specify how theory is operationalised for the purpose of measurement. As a result, Kohlberg’s three types of theories are recast in terms of formal definitions and underlying scales below.

Conceptually, cultural ages are defined in terms of events that occur according to the passage of time and/or age. Therefore, cultural ages are typically operationalised via nominal scales sequenced temporally, but not according to an increase in an underlying construct. This implies that the sequence in which cultural ages occur may be culturally relative. An example includes Erickson’s (1959) stages of psychosocial development in which age predicts the existential crisis experienced at a particular point in time through the lifecycle.
Conceptually, soft stages are defined in terms of a cumulative or quantitative form of development throughout the human lifecycle. Soft stages, therefore, are operationalised via ordinal scales which increase cumulatively in an underlying construct. Soft stages may occur in an invariant sequence; however, successive categories are quantitatively but not qualitatively distinct. There is no structure of the whole because there is no requirement for a single soft stage to organise performance in a particular domain and, therefore, performance is often broadly distributed across soft stages. Examples include Likert scales applied to leadership surveys, which reflect a cumulative increase in agreement from strongly disagree to strongly agree, or a simple accumulation of the number of perspectives that can be taken on an organisational issue. Soft stages may be analogous to a more traditional form of complexity which is sometimes referred to as horizontal complexity (Shannon & Weaver, 1949). Horizontal complexity is defined by the number of unique bits of information that must be coordinated to characterise a process and may vary independently to considerations about hierarchical complexity (Commons, 2008; Commons & Pekker, 2008).

Conceptually, hard stages are defined in terms of hierarchical transformations of human performance. Therefore, hard stages are operationalised via ordinal scales comprising hierarchically sequenced stages. As described above, the logic of hierarchical integration is satisfied if and only if Stage S+1 is defined in terms of two or more elements of Stage S, it organises and transforms the elements of Stage S, and does so in a non-arbitrary way (Commons & Pekker, 2008). Hierarchical integration results in an emergent new structure, which enables performance that cannot be accomplished at Stage S alone. The logic of hierarchical integration generates stages that are invariantly sequenced, qualitatively distinct, and for which there is a tendency for a single stage to organise performance in a particular domain. Examples include Kohlberg’s (1981) stages of moral reasoning, Kitchener and King’s (2006) stages of reflective judgement, stages in The Model of Hierarchical Complexity (Commons & Pekker, 2008), and skill levels in Fischer’s Dynamic Skill Theory (1980; 2006).
All ordinal scales must be unidimensional because they reflect an increase in a single underlying construct (Rasch, 1980). Even though Kohlberg did not formally specify this, scales that underpin soft and hard stage models are also required to exhibit unidimensionality. For clarity, the distinctions between the three types of adult development theories and underlying scales are presented in Figure 1.

**Figure 1**

*Three Types of Adult Developmental Theories and Underlying Scales: Cultural Ages, Soft Stages, and Hard Stages*

**Cultural Ages**

- Cultural age 1 (10 - 20 years)
- Cultural age 2 (20 - 30 years)
- Cultural age 3 (30 - 40 years)
- Cultural age 4 (40 - 50 years)

Operationalised via nominal categories sequenced temporally

**Soft Stages**

- Soft stage 1
- Soft stage 2
- Soft stage 3
- Soft stage 4

Operationalised via ordinal categories sequenced cumulatively / quantitatively
Notes. Top panel: Cultural ages are operationalised via nominal scales sequenced temporally, but not according to an increase in an underlying construct. Middle panel: Soft stages are operationalised via ordinal scales that increase cumulatively in an underlying construct. Bottom panel: Hard stages are operationalised via ordinal scales consisting of stages sequenced hierarchically.

The Importance of Hard Stages to the Field of Leadership

There are reasons to believe that hard stages may play an important role in integrating the fields of leadership and complexity science. Hard stage theory may enable scholars to describe, explain, and predict aspects of leadership that cannot be accounted for via other theoretical conceptions. Approaches to measuring hard stages may yield scores that can be used to make inferences about leaders’ ability to navigate complexity for the purpose of empirical analysis and leadership practice. Establishing a conceptual rationale for the utility of hard stages is important because of the limited number of empirical studies that have identified relations between hierarchical complexity and leadership phenomena.

As leaders perform at higher stages, they can increasingly exhibit responses commensurate with complex organisational environments. In some contexts, this may lead to an increase in leader effectiveness. For example, abstract linear reasoning constructs linear relations between two or more abstract concepts e.g., if A therefore B. If leaders exhibit a quantitative increase in linear reasoning as described by a soft stage model, this will merely account for more linear reasoning or
linear reasoning generalised to a broader range of contexts. Arguably, linear reasoning would be inadequate if leaders were confronted with a complex organisational dilemma which required more systemic reasoning. If leaders’ performance were hierarchically transformed from abstract linear reasoning to abstract systems reasoning, they could construct multiple relations between two or more abstract concepts simultaneously, e.g., if \((A \text{ or } B) \text{ and } C\), therefore \((D \text{ and } E) \text{ or } F\). Abstract systems reasoning is the product of a hierarchical transformation of abstract linear reasoning (Commons & Pekker, 2008; Fischer & Bidell, 2006). The relation between abstract linear reasoning and abstract systems reasoning is analogous to the relation between bivariate correlations and multivariate models. Bivariate correlations account for relations between two variables at a time, whereas multivariate models provide more explanatory power by simultaneously considering multiple relations between multiple variables. For organisational leaders, the benefits of abstract systems reasoning over and above those afforded by abstract linear reasoning are likely to include more descriptive, explanatory, and predictive power to intervene in organisational complexity (Senge, 1990).

Because of their definitional attributes, it is challenging to conceive how soft stages or cultural ages could account for this phenomenon. Despite this, cumulative accounts of complexity are often used in the leadership literature. For example, in the Leaderplex Model which is reviewed in Chapter 4, cognitive complexity is partially defined in terms of “the number of dimensions and the number of categories within dimensions that are used by individuals in the perception of the physical and social environment” (Hooijberg et al., 1997, p. 377). However, according to this definition, complexity is conceived to be little more than a cumulative increase in dimensions, rather than a hierarchical transformation of leaders’ performance.

This is not to suggest, however, that hard stages are the only or best way of conceptualising and measuring complexity. Complexity science recognises many non-hierarchical types of complexity (Lloyd, 2001). Even within the field of adult development there are different conceptions of complexity. For example, Rest’s (1979) complex stage model was partially developed in response
to an earlier iteration of Kohlberg’s (1969) hard stage requirements and presents a “much messier and [more] complicated picture of development” by predicting that test-takers’ reasoning may span a broad number of stages. Notwithstanding the different conceptions of complexity, much of the contemporary adult development literature reflects an explicit commitment to the hierarchically integrated nature of developmental stages (Commons, Li, et al., 2014; Cook-Greuter, 1999; Dawson et al., 2005). Whether hierarchical, cumulative, or other manifestations of complexity are better predictors of leadership phenomena remains an open empirical question. However, it seems reasonable to suggest that hard stages may account for a unique proportion of variability in some leadership phenomena that may not necessarily be accounted for by other types of developmental theories, or manifestations of complexity.

Because of the limited integration between leadership and adult development in the published literature, relatively few empirical studies have demonstrated the relation between hard stages and leadership phenomena. A modest number of published studies have examined the relation between scores along Kegan’s scale of the evolving self and leadership-related variables. For example, scores along his scale are related to workplace outcomes including multi-rater survey results (Harris & Kuhnert, 2008; Kuhnert, 2018; Strang & Kuhnert, 2009), perceived leadership struggles (Helsing & Howell, 2014; Van Velsor & Drath, 2004), decision-making and work complexity (Lewis & Jacobs, 1992), and attainment of leadership positions, popularity, and moral reasoning (Lucius & Kuhnert, 1999). Scores along Kegan’s scale are also related to various psychological outcomes including experiencing qualitatively different types of happiness (Fossas, 2019), subjective wellbeing (Bauger et al., 2021) and work motivation (Bugenhagen & Barbuto, 2012). It has been suggested, but not demonstrated, that conceptual aspects of his stages may account for the distinctions between transactional and transformational leadership (Kuhnert & Lewis, 1987) and different approaches to delegation (Kuhnert, 1994). For a systematic literature review of published and unpublished studies on the relation between Kegan’s constructive-developmental stages and leadership refer to Brownlow (2022).
Torbert reports positive correlations between scores along his ego development scale and aspects of organisational and team performance, such as the ability to lead organisational transformation (Rooke & Torbert, 1998; Torbert, 2013), more favourable business performance (Torbert, 1994), and higher team performance (Torbert, 1987a). He also reports positive correlations between his scores and aspects of personal and interpersonal performance such as voluntarily seeking feedback on assessment results (Torbert, 1994), autonomy in organisational roles (Torbert, 1991a), more effective ways of interacting with superiors and subordinates (Fisher & Torbert, 1991), enquiring into others’ perspectives and generating second-order solutions (e.g. redefining the problem) rather than first-order solutions (e.g. accepting given definitions) (Merron et al., 1987). Other scholars have demonstrated that scores along the ego development scale are associated with a variety of other leadership attributes. Bushe and Gibbs (1990) found that scores along the ego development scale predicted ratings of organisational development consulting competence. O’Connor and Wolfe (1991) presented evidence that higher stages of ego development were associated with the ability to navigate greater and more complex types of personal change. Joiner and Joseph (2007) presented evidence to suggest that leaders that scored higher on ego development were also stronger at context setting, stakeholder agility, creative agility, and self-leadership. Brown (2012) provided evidence to suggest that leaders who score at the most advanced ego development stages exhibit unique leadership competencies that may not be found in leaders who perform at earlier stages e.g., grounding their work in transpersonal meaning, designing from a deep inner foundation. Boiral et al. (2014) demonstrated that leaders’ stage of ego development was related to their ability to take more systemic perspectives, display a longer-range focus, integrate conflicting goals, collaborate with stakeholders, engage in collaborative learning, and their stage also predicts the environmental management practices displayed by their respective organisations. James et al. (2017) demonstrated that the leadership capability of leaders who were judged to be performing at different ego development stages were experienced and valued differently.
Anecdotal evidence and a few case studies may suggest that Gravesian stages are related to various leadership practices (Beck & Cowan, 1996; Laloux, 2014). Whether Kegan’s, Torbert’s, and Gravesian stages satisfy hard stage requirements is yet to be consistently established. As a result, it may be premature to infer that relations to leadership phenomena reported above occur because scores reflect an underlying construct of hierarchical complexity.

Kohlberg’s stages of moral reasoning (Dawson et al., 2005) and the Model of Hierarchical Complexity (Commons et al., 2008; Commons, Li, et al., 2014; Dawson et al., 2010) have been shown to satisfy hard stage requirements. Kohlberg’s scale of moral reasoning has been used to predict the ethical decision-making of MBA students (Trevino & Youngblood, 1990) and the ability to influence moral judgement in teams (Dukerich et al., 1990).

Scores based on the Model of Hierarchical Complexity have been empirically shown to predict a variety of leadership-related outcomes, including qualitative observations of community culture change (Ross, 2006a), the assumption of leadership roles (Commons, Galaz-Fontes, et al., 2006), commercial business outcomes (Goodheart et al., 2015), creativity and innovation (Commons et al., 2011), the duty to report past crimes (Commons, Lee, et al., 1995), informed consent (Commons, Rodriguez, et al., 2006), the likelihood of criminality (Li et al., 2019), perceptions of bias (Commons, 2015; Commons et al., 2012), the possibility of behaviour change (in a more therapeutic context) (Commons & Tuladhar, 2014), and remuneration (Miller et al., 2015). Because Kohlberg’s stages and The Model of Hierarchical of Complexity have been shown to satisfy hard stage requirements, it can be inferred that the relations between these scores reflect a genuine relation between hierarchical complexity and leadership-related phenomena. The Model of Hierarchical Complexity has also been used to establish conceptual relations between stages of complexity and a variety of leadership-related phenomena such as empathy (Commons & Wolfsont, 2002), investment decisions (Commons & Thexton, 2015), needs and values (Harrigan & Commons, 2015), the relation between individuals and workplace ‘atmosphere’ or culture (Commons et al., 1993), social perspective taking (Commons & Rodriguez, 1990), and spirituality and positivity (J. Day, 2010).
General considerations about stages of hierarchical complexity have been conceptually and empirically associated with leaders’ ability to implement quality improvement and organisational transformation (Kjellström & Andersson, 2017). Other considerations about the relations between cognitive-developmental stages and aspects of leadership are discussed by Reams (2017).

Outside the field of leadership, a body of literature demonstrates the value of hard stages in accounting for development throughout the human lifespan (Mascolo & Fischer, 2010). There have also been a number of studies demonstrating the conceptual and empirical value of hard stages in accounting for the conceptions of evolutionary biology (Commons & Ross, 2008a), animal intelligence (Commons & Ross, 2008b), forms of artificial intelligence (Leite & Commons, 2022), the number of neurons in the brains of various species (Harrigan & Commons, 2014), and so on. Commons and Chen (2014) provide more detailed treatment of a range of phenomena that the Model of Hierarchical Complexity has been shown to predict.

**Research Questions**

Considering this literature review, this thesis is guided by three research questions. The relation among these three research questions and the four studies undertaken to address them is illustrated in Figure 2. It is worthy of note that analyses undertaken to address all three research questions depend upon a particular approach to cognitive-developmental psychology. A thorough introduction to Lectica’s approach to cognitive development and its measurement is provided in Chapter 2. Lectica’s approach was selected on the basis that a) relatively few other adult development assessments have been subject to the kind of quantitative psychometric analyses that demonstrate they are measuring hierarchically integrated constructs; b) a convenience sample was made available which was well-suited to addressing the research questions; c) Lectica’s approach to measurement is grounded in a well-regarded cognitive-developmental theory, i.e., Dynamic Skill Theory (Fischer, 1980; Fischer & Bidell, 2006); d) empirical publications based on earlier versions of Lectica’s scoring systems afforded a body of evidence in support of Lectical Assessments e.g.,
Dawson et al. (2005); e) unlike some other adult development assessments which are intended to measure the development of the general adult population, some Lectical Assessments are specifically designed to measure the cognitive development of workplace leaders; and finally f), Lectica’s scoring systems enabled scores at a lower level of granularity to be analysed, relative to the scores awarded by other adult development assessments e.g. the subject-object interview (Lahey et al., 2001). These measurement properties enabled specific types of analyses to be undertaken that could not have been performed otherwise. The three research questions that guided this thesis are outlined below.

**Research Question 1:** Do cognitive-developmental scores awarded by the Lectical Assessment System (LAS), and ego development scores awarded by the ego development scoring system, satisfy hard stage requirements? In general terms, testing whether scores satisfy hard stage requirements plays a critical role in establishing construct validity because this determines whether scores reflect hierarchically integrated stages that are sequenced in an ascending order of complexity. More specifically, a positive response to this question may have three important implications. First, it may suggest that one of the main theoretical postulates of cognitive-developmental and ego development theory—that development unfolds hierarchically—has been successfully operationalised for the purpose of measurement. Second, it may support the contention that scores reflect the underlying construct of hierarchical complexity and can legitimately be used to make inferences about leaders’ ability to navigate organisational complexity. Third, there may be significant implications which are peculiar to cognitive development and ego development, respectively. Question 1 is addressed via Studies 1A, 1B, 2A, 2B, and 2C in Chapters 2 and 3. In Chapter 2, psychometric analyses are undertaken on a large sample of cognitive-developmental scores (Study 1A) and ego development scores (Study 1B) to determine whether they satisfy hard stage requirements. Chapter 3 comprises three shorter analyses based on different samples and research methods which attempt to explain the findings reported in Chapter 2.
Research Question 2: To what extent does the hierarchical complexity of leaders’ reasoning skills satisfy the task demands of their roles? Reasoning skills are among the strongest predictors of workplace performance (Schmidt et al., 2016; Schmidt et al., 2008). As a result, a response to this question may help to determine whether leaders at multiple organisational layers exhibit the requisite cognitive abilities to perform critical aspects of their roles. Should a complexity gap be found, there may be important implications for theorising about leader complexity and how cognitive-developmental approaches may play a broader role in human capital management. Question 2 is addressed via Study 3 in Chapter 4.

Research Question 3: To what extent do leaders develop the hierarchical complexity of their reasoning skills during participation in various leader development programs? A response to this question may play a role in determining whether leader development programs foster growth in constructs related to complexity in particular. A response to this question may also play a role in justifying the use of organisational resources that are typically invested in leader development (Lamoureux, 2007; O’Leonard, 2012). Question 3 is addressed via Study 4 in Chapter 5.
Figure 2

Relation Between the Three Research Questions and Studies Undertaken

Note. Studies 3 and 4 were based on cognitive-developmental scores generated by the Computerized Lectical Assessment System (CLAS). Dashed arrows from Studies 1B, 2A, 2B, and 2C indicate the possibility of using ego development scores for Studies 3 and 4 if they are shown to satisfy hard stage requirements.
Contributions to Theory, Empirical Research, and Practice

This thesis takes an interdisciplinary approach to pure and applied research. It is interdisciplinary in the sense that it draws on insights from the fields of leadership, adult development, and complexity science. It consists of pure research in the sense that it may contribute to our understanding of fundamental aspects of adult development and leadership phenomena. It consists of applied research in the sense that it applies a particular approach to adult development (i.e., cognitive-developmental theory and an associated form of measurement) to understanding various aspects of leadership. As a result of this approach, this thesis aims to contribute to theory, empirical research, and practice pertaining to leaders’ ability to navigate organisational complexity. These contributions are summarised in Table 1.
### Table 1

**Theoretical Basis of Each Study, Summary of Current Knowledge, Identification of Key Knowledge Gaps, and Contributions to Theory, Empirical Research, and Practice**

<table>
<thead>
<tr>
<th>Study</th>
<th>Underpinning theories, models, and measures</th>
<th>Historical research</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>Dynamic Skill Theory (Fischer, 1980; Fischer &amp; Bidell, 2006), formal axioms of hierarchical complexity (Commons &amp; Pekker, 2008), the Lectical Scale and the Lectical Assessment System (Dawson, 2022b; Dawson-Tunik, 2006)</td>
<td>Sound psychometric properties and hard stage requirements have been demonstrated in scores awarded at the scale of whole complexity levels e.g., Dawson et al. (2005)</td>
<td>May contribute to our understanding of cognitive-developmental theory. May tentatively be suggested that hierarchical integrations—or processes akin to hierarchical integrations—may inhere at a lower level of granularity than whole complexity levels</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whether similar properties are demonstrated in scores at a lower level of granularity than whole complexity levels</td>
<td>Contributes to our understanding of the measurement of cognitive development. The theoretical postulate that cognition develops hierarchically may have been successfully operationalised for the purpose of measurement. Cognitive-developmental scores may be well suited to making inferences about leaders’ ability to navigate complexity. Increments of development at a lower level of granularity than whole complexity levels can be detected, and they may predict leadership phenomena</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Contributions to theory</td>
<td>Contributions to empirical research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Contributes to our understanding of how cognitive-developmental approaches may be applied to workplace contexts e.g., creating optimal conditions for leader development</td>
<td>Contributes to our understanding of how cognitive-developmental approaches may be applied to workplace contexts e.g., creating optimal conditions for leader development</td>
</tr>
<tr>
<td>Study</td>
<td>Underpinning theories, models, and measures</td>
<td>Historical research</td>
<td>Contributions to theory</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------------------------</td>
<td>---------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>1B</td>
<td>Ego development theory (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 2013), formal axioms of hierarchical complexity (Commons &amp; Pekker, 2008), and the ego development scoring system e.g., Hy and Loevinger (1996)</td>
<td>General measurement properties have been extensively evaluated including factor structures, and various types of reliability and validity e.g., Loevinger (1979), Torbert and Livne-Tarandach (2009)</td>
<td>Whether ego development scores satisfy hard stage requirements. This seems to be assumed, but not clearly demonstrated, in the ego development literature. Relatively few studies have directly addressed this critical question</td>
</tr>
<tr>
<td>2A</td>
<td>Ego development theory (Cook-Greuter, 1999), formal axioms of hierarchical complexity (Commons &amp; Pekker, 2008), and the ego development scoring system (Cook-Greuter, 1999)</td>
<td>Ego development theory consistently espouses that its stages are hierarchically integrated and that ego development scores are a reflection of hierarchical complexity (amongst various other constructs) (Cook-Greuter, 1999; Fisher et al., 1987; Loevinger, 1976)</td>
<td>Whether ego development scores satisfy hard stage requirements when analysed with contemporary and quantitative methods</td>
</tr>
<tr>
<td>Study</td>
<td>Underpinning theories, models, and measures</td>
<td>Historical research</td>
<td>Contributions</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------------------</td>
<td>---------------------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td>Summary current of knowledge</td>
<td>Identification of key knowledge gaps</td>
<td>Contributions to theory</td>
</tr>
<tr>
<td>2B</td>
<td>Ego development theory (Cook-Greuter, 1999; Loevinger, 1976), formal axioms of hierarchical complexity (Commons &amp; Pekker, 2008), scoring exemplars from the ego development scoring system (Cook-Greuter, 1999; Hy &amp; Loevinger, 1996), the Lectical Assessment System (Dawson, 2022b; Dawson-Tunik, 2006)</td>
<td>Refer Study 2A above</td>
<td>In general terms, whether the ego development scoring system yields information about hierarchical complexity. More specifically, whether scoring exemplars from successive ego development stages increase in hierarchical complexity.</td>
</tr>
<tr>
<td>2C</td>
<td>Ego development theory (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 2013), formal axioms of hierarchical complexity (Commons &amp; Pekker, 2008), the ego development scoring system e.g., Hy and Loevinger (1996), and the Computerised Lectical Assessment System (Dawson &amp; Wilson, 2004)</td>
<td>Refer Study 2A above</td>
<td>Whether ego development scores are a stronger reflection of hierarchical complexity, a cumulative increase in theoretically relevant constructs (which are proxies for soft stages), or age (which is a proxy for cultural ages)</td>
</tr>
<tr>
<td>Study</td>
<td>Underpinning theories, models, and measures</td>
<td>Historical research</td>
<td>Contributions to theory</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------------------------</td>
<td>---------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>3</td>
<td><strong>Leader complexity and related theoretical models e.g., Lord et al. (2011), the Lectical Scale (Dawson, 2022b; Dawson-Tunik, 2006), and the Computerised Lectical Assessment System (Dawson &amp; Wilson, 2004)</strong></td>
<td><strong>Summary current of knowledge</strong></td>
<td>Whether the hierarchical complexity of leaders’ reasoning skills satisfies the task demands of their roles at multiple management layers</td>
</tr>
<tr>
<td>4</td>
<td><strong>Dynamic skill theory (Fischer, 1980; Fischer &amp; Bidell, 2006), Piagetian conceptions of disequilibration and reflective abstraction (Piaget, 2000), the Lectical Scale (Dawson, 2022b; Dawson-Tunik, 2006), and the Computerised Lectical Assessment System (Dawson &amp; Wilson, 2004)</strong></td>
<td><strong>Identification of key knowledge gaps</strong></td>
<td>Whether growth in the hierarchical complexity of leaders’ reasoning skills occurs during participation in leader development processes</td>
</tr>
</tbody>
</table>

Contributes to our understanding of the pedagogical characteristics required of leader development processes that are intended to catalyse growth in hierarchical complexity. Contributes to our understanding of the social scaffolding that leaders may require to address the complexity inherent in their organisational contexts e.g., collaborative decision-making.
Stages of adult development may play a fundamental role in shaping the ways that reality is perceived (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 2004) and the way that knowledge is constructed (Commons et al., 1998; Fischer & Bidell, 2006; Piaget, 1952). As a result, stages may also influence the approaches taken by researchers, the phenomena they research, and how their findings are interpreted and applied. Within the context of these considerations, Kjellström and Stålne (2017) argue that adult development stages can be regarded as a lens through which social science can be conducted, and also offer six research perspectives that stages of adult development afford. These research perspectives are: *Introductory work* that situates stages in the context of the social sciences; *creating and refining stages for adults* which provides the definitional attributes of stages; *comparisons and/or horizontal adaptions* which examines correspondences between stages from different theories or between stages and other scientific theories; *promoting development* which enquires into the theoretical mechanisms that account for growth between successive stages; *mismatch in adult life* which examines the relations between the development demonstrated by adults and the demands of the social contexts in which they are nested; and finally *societal and organisational development* which considers the application of stages to broader social systems e.g., organisations and society. Table 1 may suggest that this thesis makes a significant contribution to all six research perspectives. Consideration of societal and organisational development along with the broader implications are discussed in Chapter 6.
Chapter 2: Two Psychometric Studies on Adult Development Scores and Hard Stage Requirements

Study 1 addressed Research Question 1: Do cognitive-developmental scores awarded by the Lectical Assessment System (LAS), and ego development scores awarded by the ego development scoring system, satisfy hard stage requirements? Study 1A evaluates a sample of cognitive-developmental scores and Study 1B evaluates a sample of ego development scores. As mentioned in Chapter 1, testing whether scores satisfy hard stage requirements plays a critical role in establishing the construct validity of both approaches to adult development.

Cognitive and ego development are the primary focus of this thesis because they are amongst the most commonly used approaches by leadership scholars and practitioners, and they are both supported by large empirical research bases. Lectica’s™ approach to cognitive development and measurement forms the basis of the cognitive-developmental approach considered (Dawson, 2010; Dawson-Tunik, 2006). The approach to ego development is based on the cumulative work of Loevinger (1976), Torbert (2013), and Cook-Greuter (1999). Other stage theories are often referred to as ego development (Erickson, 1959; Kegan, 1982; Selman, 1980), but they are out of scope in the context of this thesis. A more contemporary version of ego development in the tradition of Loevinger has recently been proposed (Murray & O’Fallon, 2020). This is also out of scope because the definitional properties of these stages have changed considerably from the original conceptions.
(O’Fallon et al., 2020), and only one study has evaluated the psychometric properties of scores based on this theory (Murray, 2020). The patterns of results obtained in this study appear somewhat different to those obtained in psychometric studies on scores that operationalise hierarchically integrated stages.

**Hard Stage Claims Made About Cognitive Development and Ego Development**

Cognitive and ego development are contended to exemplify hard stage requirements (Commons, 2007; Cook-Greuter, 1999; Dawson et al., 2005; Kohlberg, 1975; Loevinger, 1976). For example, Dynamic Skill Theory, upon which much of this thesis is based, contends to “… [treat] cognitive development as the construction of hierarchically ordered collections of specific skills, which are defined formally by means of a set-theory description” (Fischer, 1980, p. 477). Analogously, ego developmentalists claim that “This theory shares with cognitive developmentalism the Piagetian notion of stage: ego stages are conceptualized as equilibrated structures, related to each other in an invariant hierarchical sequence” (Loevinger, 1976, p. 41). A comprehensive list of hard stage claims made by Loevinger, Cook-Greuter, Torbert, and other prominent ego developmentalists is presented in Appendix A. Most of these claims are made at the scale of theory and some may consider properties at the scale of measurement.

**Contended Relation Between Cognitive Development and Ego Development**

Study 1 is situated in the context of a 50+ year dispute as to whether cognitive-developmental and ego development stages satisfy hard stage requirements. Cognitive-developmentalists have presented evidence supporting their claims (Commons et al., 2008; Dawson et al., 2005) but have been equivocal about ego development. In some parts of the literature, cognitive-developmentalists have conceded a structural isomorphism between their stages and ego stages (Snarey et al., 1983). In other parts of the literature, they have argued that “[ego development] stages are nontransformational. Rather than each stage representing an integration
or transformation of the previous stage, new aspects are added to previous stages to define higher ones. The stages appear to be increasingly inclusive, cumulative systems” (Kohlberg & Armon, 1984, p. 391). Other cognitive developmentalists have also questioned whether ego development stages are hierarchically integrated (Noam, 1993).

Ego developmentalists have generally conceded that cognitive-developmental stages satisfy hard stage requirements but have argued that their approach focuses on reasoning at the expense of a more holistic conception of human development (Cook-Greuter, 1999; Loevinger, 1986). There are a few exceptions earlier in the literature where they questioned whether cognitive-developmental stages satisfied hard stage requirements (Loevinger, 1986).

Ego developmentalists generally argue that their stages are hierarchically integrated and reflect a more holistic conception of development than cognitive-developmental stages with which they share a structural isomorphism. Various scholars have reinforced the general contention that there is a structural isomorphism between the two sets of stages by publishing tables of correspondence in which a direct mapping between one set of stages and another is assumed but not necessarily demonstrated (Commons et al., 1990; Cook-Greuter, 1990; Fisher et al., 1987; McCauley et al., 2006; Wilber, 2000).

In some contexts, ego developmentalists have speculated that the correspondence between their stages and cognitive-developmental stages may not be clear. Loevinger (2002, p. 205) regarded “Blasi (1998) [as having] the deepest understanding of ego development and its relation to Lawrence Kohlberg’s (1964) work on moral development and its relation to psychoanalysis and the work of Piaget”. However, Blasi (1998, p. 25) conceded that “the parallelism between ego development and cognitive development is only analogical”. Study 1 makes significant headway by contributing towards a resolution of this dispute.
Distinguishing Between Hard Stages, Soft Stages, and Cultural Ages

Because hard stages may play a useful role in the field of leadership, it is important to distinguish hard stages from soft stages and culture ages. Achieving this requires clear distinctions between at least five levels of analysis: *models, scales, assessments, scoring systems, and scores.*

Hard stage models relate to theory and specify the definitional attributes of stages and the processes that account for development from one stage to the next. Scales, assessments, scoring systems, and scores based on hard stage models relate to measurement. Scales based on hard stage models operationalise hard stages for the purpose of measurement. Assessments elicit observations (e.g., samples of text) from test-takers that can be scored with scoring systems. Scoring systems attribute scores to observations based on the magnitude of hierarchical complexity they exhibit. Scores are the product of scoring observations with scoring systems. As a result, scores do not reflect some generic conception of development. Instead, scores reflect the amount of hierarchical complexity attributable to observations and can be used to make inferences about leaders’ ability to navigate complexity.

Evaluating whether a set of stages satisfies hard stage requirements may take considerations at all five levels of analysis into account. These considerations may include the definitional attributes of stages, the logic underpinning the sequence of stages, the administration procedures of assessments, the scoring rules that determine how scoring systems award scores to observations, and the behaviour of scores. To ensure continuity with previous research (Commons, Li, et al., 2014; Cook-Greuter, 1999; Dawson et al., 2005), this thesis prioritises: a) the logic underpinning the sequence of stages and b) the behaviour of scores. The logic underpinning the sequence of stages must consider whether logical relations between stages reflect a succession of hierarchical integrations. This logic is likely expressed in respective theories and/or scoring systems. The expected behaviour of scores is derived deductively from the theoretical postulates of hard stage models and inductively from empirical findings on scores based on hard stage models.
It may be espoused that a set of stages satisfies hard stage requirements but fails to satisfy a) and b) above. It is also possible for a set of stages to satisfy a) but fail to operationalise theory for the purpose of measurement by failing b). This would result in a problem with measurement but not necessarily with theory. A scenario in which a set of stages fails a) but satisfies b) seems unlikely but possible.

It may be most accurate to say that a) the logic underpinning the sequence of stages satisfies hard stage requirements, whereas b) the behaviour of scores must be consistent with hard stage requirements. Considerations about theory and measurement are used throughout this thesis, so the term satisfy/satisfies hard stage requirements will be used to ensure terminological consistency.
**Study 1A: Cognitive-Developmental Scores Awarded by the Lectical Assessment System (LAS)**

The purpose of Study 1A was to determine if cognitive-developmental scores awarded by the LAS satisfy hard stage requirements.

**Approach to Cognitive Development**

**Cognitive-Developmental Theory**

The Piagetian and neo-Piagetian traditions primarily inform the theoretical orientation to cognitive development applied in this thesis. This approach to cognitive development is based on Fischer’s Dynamic Skill Theory, which is a comprehensive theory of human development throughout the lifespan (Fischer, 1980; Fischer & Bidell, 2006). The theory’s central tenet is that skills grow through successive levels of hierarchical complexity, or *complexity levels*. Skill is “the capacity to act in an organized way in a specific context” and develops in a dynamic system which incorporates the self, the interpersonal other, and the environment (Fischer, 1980; Fischer & Bidell, 2006, p. 321). As a theory of development through the lifespan, it accounts for the development of innate reflexes in infancy, sensorimotor actions in childhood, and advanced reasoning skills in adulthood. Whereas this approach emerged from cognitive-developmental tradition, reasoning skills are not narrowly conceived to focus on thought in isolation from other psychological processes. Piaget (1968) argued, “By ‘reasoning’, we mean a rather inclusive form of meaning making which reunites what a person thinks, feels, and does” (Piaget, 1968, as cited in Snarey et al., 1983, p. 309). This integrative conception of cognition as a structured whole that recruits multiple psychological processes is reflected in Piaget’s (1968) original work, is maintained by many of his successors in the cognitive-developmental tradition (Fischer & Bidell, 2006), and has attracted considerable support from psychological and neuropsychological scholars alike (LeDoux & Brown, 2017; Okon-Singer et al.,
This integrative conception of cognition has also been supported empirically as the hierarchical complexity of reasoning has been shown to vary dynamically as a function of context, knowledge domain, task, level of support and emotional state; and not merely as a function of thought considered in isolation from other psychological processes (Fischer, 1980; Fischer & Bidell, 2006; Fischer et al., 1993; Kitchener et al., 1993a; Mascolo & Fischer, 2010).

Reasoning skills, therefore, are conceived to be “integrative psychological structures” that recruit thought, emotion, motivation, meaning, and action (Mascolo & Fischer, 2010, p. 5).

Piaget (2000) referred to the process of transforming lower-order cognitive structures into qualitatively distinct higher-order structures as reflective (or reflecting) abstraction. The process of reflective abstraction leads to a state in which cognitive structures become hierarchically integrated in the mind-brain (Smith et al., 2019). This hierarchical chunking reduces the number of concepts which must be coordinated simultaneously and frees up cognitive processing space to enable increasingly complex reasoning to emerge (Burtis, 1982; Halford, 1999). Hierarchical complexity is conceived to be the behavioural manifestation of hierarchical integration and is directly observable in sensorimotor actions, verbal behaviour, and social contexts (Dawson et al., 2005; Dawson-Tunik, 2006).

**Cognitive-Developmental Scale**

Lectica is a not-for-profit organisation registered in the state of Massachusetts, USA. Its mission is to “support robust learning by developing and delivering learning tools that support the development of the skills and concepts we all need to meet the challenges of the 21st century” (Lectica, 2023, para. 1). Lectica’s aims are to: “(1) to provide free high-quality, evidence-based developmental learning tools to individual K-12 educators, (2) to advance this work by offering related research and assessment services to clients in the private and public sectors, and (3) to build knowledge about learning and its role in the future of society” (Lectica, 2023, para. 2). As a result of the mission and aims stated above, Lectica builds, administers, and conducts research on cognitive-
developmental assessments for K-12 and adult learners (Lectica, 2018c). In general terms, Lectica may be regarded as competing in the industry of cognitive-development and/or educational assessment.

Lectica has refined Fischer’s Dynamic Skill Scale (Fischer, 1980; Fischer & Bidell, 2006) in order to develop the Lectical Scale™ (Dawson, 2010, 2022b; Dawson-Tunik, 2006). It is a unidimensional ordinal scale that reflects a single underlying construct, hierarchical complexity. Scores at the scale of whole complexity levels have been demonstrated to satisfy several requirements of measurement (Dawson et al., 2005). The Lectical Scale consists of 14+ complexity levels. Each complexity level is defined in terms of two properties: the hierarchical order of abstraction of concepts and the logical structure of arguments. The hierarchical order of abstraction refers to the implicit amount of hierarchical complexity associated with the meaning of concepts. The logical structure of arguments refers to the explicit amount of hierarchical complexity required to coordinate concepts into arguments. For example, in the statement, “If you abuse dogs, they may become vicious”, the hierarchical order of abstraction accounts for the implicit amount of hierarchical complexity associated with the meaning of abuse and vicious. The logical structure of arguments accounts for the explicit ‘if A therefore B’ logical structure which coordinates the two concepts to form an argument (Dawson et al., 2005). Dawson has provided evidence to suggest that the hierarchical order of abstraction and the logical structure of arguments are definitionally identical and that distinctions are only made for heuristic and pragmatic reasons (Dawson & Gabrielian, 2003; Dawson-Tunik, 2004, 2006). The way these two properties define the complexity levels which constitute the full Lectical Scale is outlined in Appendix B. A truncated section of the scale that applies to leader development is presented in Table 2.
## Table 2

### Truncated Version of the Lectical Scale Relevant to Leader Development

<table>
<thead>
<tr>
<th>No.</th>
<th>Lectical Level</th>
<th>Hierarchical order of abstraction of concepts</th>
<th>Logical structure of arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Primary order of abstraction</td>
<td>Secondary order of abstraction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tertiary order of abstraction (i.e., Lectical Phase)</td>
<td>Underlying structure</td>
</tr>
<tr>
<td>10</td>
<td>Abstract mappings</td>
<td>The abstractions order accounts for development from approximately 8 years through to adulthood. At this order, children through adults construct abstract conceptualisations that generalise across concrete instances. These abstractions grow in complexity from single abstract ideas to entire systems of abstractions (i.e., abstract systems thinking).</td>
<td>2nd-order abstractions which coordinate 1st-order abstractions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10d highly elaborated</td>
</tr>
<tr>
<td>11</td>
<td>Abstract systems</td>
<td>3rd-order abstractions which coordinate 2nd-order abstractions</td>
<td>11a transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11b unelaborated</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11c elaborated</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11d highly elaborated</td>
</tr>
<tr>
<td>12</td>
<td>Single principles</td>
<td>The principles order accounts for the development of a minority of adults from approximately 26 years onwards (if at all). In this order, adults construct overarching principles which govern relations between multiple abstract systems.</td>
<td>1st-order principles which coordinate 3rd-order abstractions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12b unelaborated</td>
</tr>
</tbody>
</table>


Table 2 and Appendix B show that the hierarchical order of abstraction can be described at three nested layers of granularity. This resembles fractal-like patterns of repeating self-similarity and has also been observed in related models, such as The Model of Hierarchical Complexity (Commons & Richards, 2002; Ross, 2008a). At each order of abstraction, concepts hierarchically integrate the concepts from the previous order. For example, the concepts of learning and play, which occur at Lectical Level 8, are hierarchically integrated to form the new and emergent concept of learning as play which occurs at Lectical Level 9. The new concept of learning as play transforms the meaning of the two constituent elements at a higher-order of abstraction (Dawson-Tunik, 2006).

Each complexity level is also defined in terms of one of three repeating logical structures: single elements, mappings, and systems. Single elements identify one aspect of an element e.g., “A good leader is good with people”, in which being good with people is a single quality of being a good leader. Mappings coordinate one aspect of two or more elements in a linear relation e.g. “A good leader is good with people, which makes them trust her intentions”, in which being good with people and trusting intentions are organised in a linear relation. Systems coordinate multiple aspects of two or more elements in a multivariate relation e.g., “A good leader is inspiring if she is trustworthy and competent without appearing arrogant”, in which being a good leader consists of relations between multiple abstract qualities such as being inspiring, trustworthy, and competent (Dawson, 2008; Dawson-Tunik, 2006). Illustrative examples of how hierarchical complexity manifests in conceptions of leadership are provided in Table 3. These are merely illustrative examples because short utterances are not typically scored in isolation from the full arguments in which they occur.
Table 3

Examples of Conceptions of a Good Leader at Successive Complexity Levels

<table>
<thead>
<tr>
<th>Complexity Level</th>
<th>Conception of a good leader</th>
<th>Explanation of location at Complexity Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lectical Level 8</td>
<td>“A good leader is fun to have because she knows where to go, knows what to do and will show you how”</td>
<td>At this level, “a good leader is fun” represents someone who acts out a system of concrete but interrelated instances—“knows what to do,” “knows where to go”, and “will show you how”</td>
</tr>
<tr>
<td>Lectical Level 9</td>
<td>“A good leader is good with people, which means they are fun, helpful, and friendly”</td>
<td>At this level, the single abstraction “good with people” is a general quality that integrates a system of representations—“fun, helpful, and friendly”</td>
</tr>
<tr>
<td>Lectical Level 10</td>
<td>“A good leader is good with people which makes them trust her intentions”</td>
<td>At this level, two abstract qualities—“good with people” and “trust her intentions”—are related in a linear fashion that allows one quality to be inferred from the other</td>
</tr>
<tr>
<td>Lectical Level 11</td>
<td>“A good leader is inspiring if she is trustworthy and competent without appearing arrogant”</td>
<td>At this level, multiple abstractions—“inspiring,” “trustworthy,” “competent,” and “arrogant”—are related in a system with conditional and implied if/then relationships</td>
</tr>
<tr>
<td>Lectical Level 12</td>
<td>“A good leader is a highly competent servant to her organization, one who is inspiring, visionary, and deeply committed to achieving the shared goals of her organization”</td>
<td>The transition to principles allows for multiple complex abstractions “inspiring,” “visionary” and “deeply committed”—each of which could be elaborated (in a longer example) as abstract systems in their own right, to be repackaged as a single principled notion of a servant leader</td>
</tr>
</tbody>
</table>

Note. Adapted from “Lectical levels: Structure and origins,” By Dawson (2021) (https://lecticalive.org/about/skill-levels#gsc.tab=0). Copyright 2021 Lectica Inc.

As seen in Table 3, the same concept can be understood differently at different complexity levels. To provide another example that is less related to the field of leadership, there may be reasons to believe that the principle of natural selection may satisfy the requirements of a single
principle at Lectical Level 12 in the way it was originally conceived by Darwin (1859). However, there may also be reasons to believe that evolutionary processes may be conceived in a considerably simpler way by high-school and tertiary students. To provide a more contemporary example, it could be instructive to analyse the output of artificial intelligence programs such as ChatGPT (OpenAI, 2023) from the perspective of complexity levels. The hierarchical complexity of ChatGPT output may be a function of the hierarchical complexity of the questions it is asked, its specialised algorithms, its adaptive learning capacity, or interactions between these factors. This remains an open empirical question at this point in time.

In leader development, reasoning skills are accounted for within the abstractions and principles orders of abstraction. The abstract order of abstraction typically accounts for development from approximately eight years to adulthood. In this order, children through to adults construct abstract conceptualisations that generalise across concrete instances. These abstractions grow in complexity from single abstract ideas to entire systems of abstractions. At the principles order of abstraction, adults construct overarching principles that govern relations between multiple abstract systems (Dawson et al., 2005; Fischer & Bidell, 2006).

At a lower level of granularity, leader development can be accounted for within Lectical Levels 10, 11, and 12. At Lectical Level 10, reasoning is defined by 2\textsuperscript{nd}-order abstractions and linear reasoning, which coordinates one aspect of two or more abstractions. At Lectical Level 11, reasoning is defined by 3\textsuperscript{rd}-order abstractions and abstract systems reasoning, which coordinates multiple aspects of two or more abstractions. At Lectical Level 12, reasoning is defined by 1\textsuperscript{st}-order principles and definitional reasoning that coordinates one aspect of a single principle.

At an even lower level of granularity, leader development is typically accounted for by Lectical Phases 10c, 10d, 11a, 11b, 11c, 11d, 12a and 12b. Each complexity level (e.g., 11) comprises four Lectical Phases (e.g., 11a, 11b, 11c, 11d), as seen in Table 2 and Appendix B. Each phase is contended to hierarchically integrate the concepts from the preceding phase. Lectical Phase a (X.00–X.24) is called transitional because reasoning is transitioning from the previous complexity
level to the current one. Lectical Phase b (X.25–x.49) is called *unelaborated* because reasoning is yet to be fully elaborated at the current complexity level. Lectical Phase c (X.50–X.74) is called *elaborated* because reasoning is now elaborated at the current complexity level. Lectical Phase d (X.75–X.99) is called *highly elaborated* because reasoning is now highly elaborated at the current complexity level. Lectical Phases 10c and 10d account for increasingly complex linear reasoning. Lectical Phases 11a through 11d account for increasingly complex abstract systems reasoning. Lectical Phases 12a and 12b account for increasingly complex principles reasoning. Lectical Phases are discussed in some published studies (Dawson, 2006; Dawson et al., 2005) and unpublished papers (Dawson, 2008), but empirical and/or qualitative analyses of the logic underpinning Lectical Phases could not be identified.

Lectical Phases reflect a unique way of conceptualising of development within complexity levels. However, they may bear some relation to the transition steps discussed in the context of other cognitive-developmental models. In these models, hierarchical integrations—or processes akin to hierarchical integrations—have been posited to account for smaller increments of development as performance transitions from one complexity level to the next (Commons et al., 2005; Commons & Richards, 2002; Fischer, 1980; Fischer & Bidell, 2006; Lahey et al., 2001; Ross, 2008a).

**Measuring Cognitive Development**

Lectica offers a suite of cognitive-developmental assessments called Lectical Assessments. Each assessment measures the development of reasoning in a particular domain. For example, the Lectical Leadership Decision-Making Assessment™ (LDMA) assesses the development of reasoning applied to leadership decision-making. Other Lectical Assessments include the Lectical Ethical Reasoning Assessment™ (LERA), the Lectical Developmental Pedagogy Assessment™ (LDPA), the Lectical Leadership Reasoning Assessment™ (LLRA), Lectical Reflective Judgement Assessment™ (LRJA) and the Lectical Self Understanding Assessment™ (LSUA). Several assessments are used to
inform analyses undertaken in this thesis. Lectical Assessments are usually administered online and present test-takers with a semi-structured, real-life, and complex dilemma that prompts their reasoning. Dilemmas are constructed to elicit deliberation upon multiple considerations, such as psychological, behavioural, ethical, and other factors. Test-takers assume the role of the protagonist and address the dilemma through a minimum of 5 items that require short essay-type responses. Items invite test-takers to provide judgments (‘what do you think?’) and justifications (‘why do you think what you think?’) of sufficient length to score, which is typical of the approach taken in the cognitive-developmental tradition (Kohlberg, 1984). Since Lectical Assessments measure cognitive power rather than speed, time limits are not imposed.

Lectical Assessments can be used in a formative or summative manner (Dawson, 2020; Stein et al., 2010) depending on the requirements of the initiative to which Lectical Assessments are applied. In the context of this thesis, formative applications include using information about the hierarchical complexity of leaders’ reasoning skills to inform the pedagogical design of curricula and selection of developmental practices. For example, leaders with lower scores may benefit from less complex practices, such as seeking diverse perspectives. In contrast, leaders with higher scores may benefit from more complex practices, such as identifying the epistemic paradigm(s) in which diverse perspectives emerged. Summative applications include using assessments to measure change over time but not to inform design decisions or the selection of developmental practices. The summative versus formative applications of Lectical Assessments do not affect scoring procedures.

**Cognitive-Developmental Scoring Systems**

Test-takers’ responses are scored via Lectica’s two scoring systems: the Lectical Assessment System (LAS) (Dawson et al., 2005), and the Computerized Lectical Assessment System (CLAS) (Dawson & Wilson, 2004). The purpose of both systems is to locate the magnitude of hierarchical complexity exhibited in samples of reasoning along the Lectical Scale. Lectica’s scoring systems are considered to assess the core structure of development because scores are only awarded based on
two underlying structural properties i.e., hierarchical order of abstraction and logical structure of arguments. For this reason, they can assess reasoning in any domain in which reasoning can be undertaken, from leadership decision-making to understanding physics (Dawson, 2001, 2006). This stands in contrast to most scoring systems used in adult development that award scores based on conceptual content (i.e., the specific words that test-takers use) or surface structure (i.e., taking a particular social perspective). For this reason, most other scoring systems can only assess development in specific domains. LAS and CLAS are primarily informed by Piaget’s (1968) account of cognitive structures, Dynamic Skill Theory (Fischer & Bidell, 2006) and the General Stage Scoring System (Commons, Straughn, et al., 1995).

The LAS is a human-powered scoring system utilised by Lectical Analysts™ who have a minimum of four years of training. The scoring approach is particularly applicable in high-stakes settings such as executive assessment. CLAS is a scoring system powered by artificial intelligence and is particularly applicable in large-scale assessment contexts. While LAS and CLAS use somewhat different procedures to award scores, the published literature has demonstrated that the two scoring systems have an inter-rater agreement of 80% within one-third of a complexity level, which corresponds to a Kendall’s tau of .93 (Dawson & Wilson, 2004). A follow-up study demonstrated that the LAS and the CLAS have an inter-rater agreement of 96.1% within one complexity level (Dawson, 2004). As a result, LAS and CLAS scores are interchangeable for research purposes within certain constraints, as discussed in Chapters 4 and 5.

Detailed scoring procedures underpinning LAS and CLAS are the proprietary intellectual property of Lectica, however, a brief overview is provided. The LAS requires trained Lectical Analysts to pair the hierarchical order of abstraction with the logical structure of arguments. An illustrative example is provided by Dawson et al. (2005) in the following two sentences: “If you hit dogs they might bite you” and “If you abuse dogs they may become vicious”. From the perspective of the logical structure of arguments, both sentences are mappings because they employ linear if-then relations between concepts. However, from the perspective of the hierarchical order of abstraction,
hit and bite in the first sentence are representations. In contrast, abuse and vicious in the second sentence are abstractions. For this reason, the two sentences would be awarded LAS scores of Lectical Levels 7 and 10, respectively. CLAS, however, only relies upon the hierarchical order of abstraction because logical structures repeat at successive orders of abstraction, as discussed above. CLAS is based on the finding that the density of concepts at successive Lectical Phases is probabilistically related to a LAS score along the Lectical Scale. Density is defined as the proportion of unique concepts at Lectical Phase P relative to the number of unique words in a sample of reasoning. A discriminant function is applied to the density of concepts at successive Lectical Phases to yield a CLAS Score (Dawson & Wilson, 2004).

Cognitive-Developmental Scores

A score along the Lectical Scale—whether awarded by LAS or CLAS—reflects the magnitude of hierarchical complexity exhibited in a sample of reasoning. This score is used to make inferences about test-takers’ functional level of reasoning in a specific domain, such as leadership decision-making. The functional level is the level at which test-takers can perform in standard conditions without additional stress or support (Fischer & Bidell, 2006). The score cannot be used to make broad generalisations about the level at which other reasoning skills or test-takers may be located. In the dynamic cognitive-developmental approach, discrete leadership decisions but not decision-makers, are awarded Lectical Scores. Because all scores on the Lectical Scale are calibrated to the same underlying scale, two or more scores along the Lectical Scale can be directly compared. For example, one could compare the amount of hierarchical complexity exhibited in leadership decision-making between test-times 1 and 2, between leadership decision-making and ethical reasoning at test-time 1, or between the task demands of leaders’ roles and their actual reasoning skills.

To date, published psychometric studies have been undertaken on scores at the scale of whole complexity levels (Dawson et al., 2005). Other research projects have examined the relation between Lectical decimal scores and other variables (Fuhs, 2016; Van Rossum, 2013). In this thesis,
cognitive-developmental scores are operationalised by Lectical Phase Scores awarded by LAS and decimal scores awarded by CLAS. As a result, these two types of scores are used as the basis for analyses.

**Evidence that Cognitive-Developmental Scores Satisfy Hard Stage Requirements**

Previous research has demonstrated that cognitive-developmental scores at the scale of whole complexity levels satisfy hard stage requirements (Commons et al., 2008; Davison et al., 1980; Snarey, 1985). There may be reasons to believe that specific developmental theories and scores based on those theories—such as The Model of Hierarchical Complexity—may consist of stages that are even harder than those described by earlier developmentalists because of the formal mathematical axioms and theorems that define the properties of those stages (Commons, Li, et al., 2014; Commons & Pekker, 2008; Commons et al., 1998). However, there is a paucity of research on scores at a lower level of granularity and how they may be used to create optimal learning conditions for leaders. The following review is focused on the nine studies that have examined the behaviour of scores awarded by the LAS (Dawson, 2010, 2022b; Dawson-Tunik, 2006) and the scoring system which immediately preceded it, the Hierarchical Complexity Scoring System (HCSS) (Dawson, 2003a). The main arguments are summarised below, followed by a short review of the nine studies.

First, Rasch analyses have demonstrated that LAS/HCSS scores satisfy hard stage requirements. A detailed introduction to the Rasch Model is provided in the section below. In short, The Rasch Model transforms raw ordinal scales into a linear interval scale called the logit scale (Andrich, 1988; Rasch, 1980). The logit scale comprises units called logits, and estimates of person ability, \( \beta_n \), and item difficulty, \( \delta_i \), are co-located on the logit scale for direct comparison. Second, correlational analyses have demonstrated a strong relation between LAS/HCSS scores and other cognitive-developmental scores. High correlations have been used to argue that scoring systems tap a common dimension of human performance, i.e., hierarchical complexity. Third, the LAS/HCSS has
been used to score the scoring criteria from other cognitive-developmental scoring systems to
demonstrate that concepts unfold in a similar sequence. These arguments have been used to
establish that LAS Scores, at the scale of whole complexity levels, satisfy hard stage requirements.
Methodological limitations of these studies are discussed below.

**Consistency Between LAS/HCSS Scores and Hard Stage Requirements**

**Unidimensionality.** The unidimensionality of LAS/HCSS scores has been demonstrated via
Rasch fit indices for persons and items. Some studies indicated overfit to the Rasch Model, which
suggests that some samples were almost “too good to be true” (Dawson et al., 2005, p. 183).
Unidimensionality has also been supported by the high correlations between LAS HCSS scores and
other cognitive-developmental scores, which range from .86 to .92 with a median of .90 and mean of
.90 (Dawson, 2001, 2002a, 2003b; Dawson et al., 2003).

**Invariant Sequence.** Rasch analyses are typically conducted on LAS/HCSS scores awarded to
transcribed passages from cognitive-developmental interviews. Invariant sequence has been
demonstrated because items are consistently sequenced correctly along the logit scale. For
example, Item-Person Maps show that items scored at Lectical Levels 7 through 12 occur in a
theoretically predicted sequence and reflect an increase in difficulty (refer to Figure 3).

**Qualitative Distinctness.** Qualitative distinctness has been demonstrated because items at
each complexity level cluster in a narrow range of logits, and items at successive complexity levels
are distinctly separated along the logit scale. From Figure 3, it can be seen that items at Abstract
Mappings (i.e., Lectical Level 10) occur in a narrow range between 1-3 logits. There is considerable
‘white space’ between those items, and items at Single Abstractions (i.e., Lectical Level 9) and
Abstract Systems (i.e., Lectical Level 11). This suggests that items at successive complexity levels are
qualitatively distinct in terms of the magnitude of hierarchical complexity they reflect.
**Figure 3**

*Item-Person Map of Scores Awarded to Moral Reasoning Interviews (Dawson et al., 2005, p. 181)*

<table>
<thead>
<tr>
<th>Logits</th>
<th>Person estimate locations (1 = 1 person)</th>
<th>Item level estimate locations</th>
<th>Complexity Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td></td>
<td></td>
<td>SP</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes. The logit scale is shown in the middle from -22 to 24 logits, person ability, $\beta_n$, shown on the left and item difficulty, $\delta_i$, shown on the right. Items correspond to Lectical Levels 7 through 12. Adapted from “The Shape of Development,” by T. L. Dawson-Tunik, M. Commons, M. Wilson, K. W. Fischer, 2005, *European Journal of Developmental Psychology*, 2(2), p. 181. Copyright 2005 by Psychology Press Ltd. Figure produced with permission from Taylor & Francis Group (www.tandfonline.com). Representational Mappings (RM) corresponds to Lectical Level 7, representational systems (RS) corresponds to Lectical Level 8, single abstractions (SA) corresponds to Lectical Level 9, abstract mappings (AM) corresponds to Lectical Level 10, abstract systems (AS) corresponds to Lectical Level 11, and single principles (SP) corresponds to Lectical Level 12.
**Structure of the Whole.** Structure of the whole has been demonstrated because there is a tendency for a single complexity level to organise reasoning in a given domain. This tendency is demonstrated because items at each complexity level cluster in a narrow range along the logit scale (refer to Figure 4). It is also demonstrated in Category Probability Maps where, for almost any person ability along the x-axis, there is a high probability (> .80) of performance being scored at a single complexity level (refer to Figure 4). It can also be seen that the probability of convergence between adjacent complexity levels occurs with a probability of < .1, and there is a negligible probability of convergence between nonadjacent complexity levels. Convergence between adjacent complexity levels is expected as reasoning becomes consolidated at the current modal level and continues to develop to the next level.
Figure 4

Category Probability Curves of Lectical Levels 7 Through 12 Identified in Scores Awarded to Moral Reasoning Interviews (Dawson et al., 2005, p. 184)

Notes. Person ability, $\beta_{nu}$, is shown along the horizontal axis and the probability of response is shown on the vertical axis. Adapted from “The Shape of Development,” by T. L. Dawson-Tunik, M. Commons, M. Wilson, K. W. Fischer, 2005, European Journal of Developmental Psychology, 2(2), p. 184. Copyright 2005 by Psychology Press Ltd. Figure produced with permission from Taylor & Francis Group (www.tandfonline.com). Representational Mappings (RM) corresponds to Lectical Level 7, representational systems (RS) corresponds to Lectical Level 8, single abstractions (SA) corresponds to Lectical Level 9, abstract mappings (AM) corresponds to Lectical Level 10, abstract systems (AS) corresponds to Lectical Level 11, and single principles (SP) corresponds to Lectical Level 12.
Hierarchical Integrations. Hierarchical integrations have been demonstrated via the spurs and plateaus in Item Equating Maps (refer to Figure 5). Plateaus reflect periods of consolidation at a single complexity level, and spurs reflect vacillation between the current modal level and its successor. This is a pattern predicted by a multitude of developmental scholars (Andrich & Styles, 1994; Case, 1991; Dawson, 2006; Dawson et al., 2005; Dawson-Tunik, 2004; Fischer & Rose, 1994, 1999; Fischer & Silvern, 1985; Kitchener et al., 1993b; Liu & McKeough, 2005; Shultz, 2003; Thomas & Lohaus, 1993; Van der Maas & Molenaar, 1992; van Geert, 1998; Walker et al., 2001; Wilson, 1989). Hierarchical integrations have also been supported by the high correlations between LAS/HCSS scores and other cognitive-developmental scores, which occur between .86 to .92, with a median of .90, and a mean of .90 (Dawson, 2001, 2002a, 2003b; Dawson et al., 2003). This suggests that hierarchical complexity is a common dimension of performance tapped by these scoring systems.
**Figure 5**

*Item Equating Map of Scores Awarded to Moral Reasoning Interviews (Dawson et al., 2005, p. 185)*

![Figure 5](image)

**Figure 6.** Expected scores by person performance estimates.

**Notes.** Person ability, $\beta_n$, is shown along the horizontal axis, and the expected score along the Lectical Scale is shown on the vertical axis. Plateaus demonstrate consolidation of reasoning at a single complexity level, and spurts demonstrate transition from the current to the next complexity level. Adapted from “The Shape of Development,” by T. L. Dawson-Tunik, M. Commons, M. Wilson, K. W. Fischer, 2005, *European Journal of Developmental Psychology*, 2(2), p. 185. Copyright 2005 by Psychology Press Ltd. Figure produced with permission from Taylor & Francis Group (www.tandfonline.com). Representational Mappings (RM) corresponds to Lectical Level 7, representational systems (RS) corresponds to Lectical Level 8, single abstractions (SA) corresponds to Lectical Level 9, abstract mappings (AM) corresponds to Lectical Level 10, abstract systems (AS) corresponds to Lectical Level 11, and single principles (SP) corresponds to Lectical Level 12.
**Review of Empirical Studies on LAS/HCSS Scores**

In the first study, 42 participants undertook two cognitive-developmental interviews—the Moral Judgement Interview and the Good Life Interview (Dawson, 2001). As is typical in the cognitive-developmental tradition, participants were required to provide arguments that comprised a judgement (‘what do you think?’) and a justification (‘why do you think what you think?’). Interviews were scored with the HCSS and another domain-specific scoring system by two groups of scorers. Correlations between the two sets of scores ranged from .68 to .93, with a mean of .86. On average, there was 98% agreement between HCSS scores and domain-specific scores within one complexity level.

In the second study, 220 participants aged between 5 to 86 years undertook three cognitive-developmental interviews—the Moral Judgement Interview, the Good Life Interview, and a Good Education Interview (Dawson, 2002a). The Moral Judgement Interview and Good Life Interviews were scored with domain-specific scoring systems, while the Good Education Interview was scored with the HCSS. Rasch analyses demonstrated that all three scoring systems awarded scores in a manner that satisfied hard stage requirements. However, HCSS scores appeared to satisfy hard stage requirements more successfully than other scores. Disattenuated correlations between HCSS scores and domain-specific scores occurred between .90 to .92.

In the third study, 637 participants aged between 5 to 86 years undertook a Moral Judgement Interview (Dawson, 2003b). Interviews were scored with the moral judgement scoring system and with the HCSS. Statistical (i.e., non-Rasch) analyses were consistent with the summary of the main arguments presented above. Specifically, arguments scored at Lectical Levels 8 through 12 occurred in the theoretically predicted sequence. For most participants, arguments occurred at a single complexity level, and 95% confidence intervals constructed around arguments scored at complexity level L did not overlap with 95% confidence intervals constructed around arguments at complexity level L+1. Correlational analyses demonstrated a .88 correlation between the HCSS and moral judgement scores. This study also included an analysis of the moral judgement scoring...
system. The moral judgement scoring system contains criterion judgements which define how moral issues are conceived at successive stages of moral judgement. Criterion judgements determine how moral judgement scores are awarded to moral judgment interviews. Two-hundred-and-nineteen criterion judgements were scored with the HCSS, meaning that each criterion judgement was given a moral judgement stage score and a HCSS score. The correlation between the two sets of scores was .97, which suggested that the two scoring systems were tapping a common construct.

In the fourth study, 378 participants aged between 5 to 86 years undertook a Moral Judgement Interview (Dawson et al., 2003). Interviews were scored with the moral judgement scoring system and the HCSS. Rasch analyses demonstrated that both scoring systems awarded scores in a manner that satisfied hard stage requirements. However, the HCSS scores satisfied hard stage requirements more successfully than moral judgement scores. Correlational analyses demonstrated a .92 correlation between scores.

In a fifth study, 713 participants aged between 4 to 86 years undertook a Moral Judgement Interview (Dawson & Gabrielian, 2003). Interviews were scored with the moral judgement scoring system and the HCSS. Qualitative analysis of 90 moral judgement interviews at various stages demonstrated that conceptions of morality unfolded in a similar sequence between the two scoring systems.

In the sixth study, 152 undergraduate students undertook a cognitive-developmental interview on conceptions of epistemology (Dawson, 2004). Interviews were scored with different scoring systems: the Perry reflective judgement scoring system and the HCSS. All scoring systems awarded similar scores (HCSS & Perry = 91% agreement within 1 stage). Also, 95% confidence intervals constructed around arguments at Level 10 scored did not overlap with 95% confidence intervals constructed around arguments at Level 11. This supported the argument of qualitative distinctness between successive complexity levels.

In the seventh study, 246 participants aged between 5 to 86 years undertook a cognitive-developmental interview about conceptions of a good education (Dawson-Tunik, 2004). Interviews
were scored with the HCSS. Rasch analyses demonstrated that scores satisfied hard stage requirements.

In the eighth study, 747 participants aged between 5 to 86 years undertook a Moral Judgement Interview (Dawson et al., 2005). Interviews with scored with the LAS. Rasch analyses demonstrated that scores satisfied hard stage requirements. It is worthy of note that for the studies considered above, the samples consisted of participants with similar ages ranges. The author of this thesis is not aware of whether these studies were performed on different samples with similar age ranges, or if different assessments were administered to similar samples. Based on the demographic details provided by the author of these studies, it seems possible that there may have been some overlap in the samples analysed for different studies.

Finally, in the ninth study, 49 9th graders aged between 14 to 16 undertook an interview on their conceptions of energy in physics (Dawson, 2006). Interviews were scored with the LAS. While the sample was limited to conceptions of energy between Lectical Levels 9 and 10, Rasch analysis demonstrated that LAS scores satisfied hard stage requirements.

**Limitations of Previous Studies**

While Dawson makes a compelling argument that HCSS/LAS scores satisfy hard stage requirements, several limitations are noted. Most studies are limited to cognitive-developmental interviews, and it is unclear whether similar patterns would be observed in samples of reasoning acquired through written communication. The studies focus on the structural validity of HCSS/LAS scores but do not consider the predictive or convergent validity in relation to variables considered in the leadership literature e.g., leader emergence, leader role occupancy, leader performance, leader effectiveness (Messick, 1989, 1995). The studies included participants from the general population but not from leadership populations. Seven out of the nine studies were conducted on the HCSS, which is the predecessor of the LAS, so it is unclear whether LAS scores would consistently exhibit similar patterns. Some of the Rasch procedures employed in these previous studies were relatively
limited and did not explicitly test for other relevant psychometric attributes such as the magnitude of multidimensionality and response dependence. Finally, all studies were undertaken on scores at the scale of whole complex levels. As a result, it is unclear whether the reliability and validity of scores at a lower level of granularity i.e., Lexical Phase Scores, justifies their use in research and practice.

**Introduction to the Unidimensional Rasch Model**

The Unidimensional Rasch Model is an algebraic approach to evaluating the psychometric properties of measurement instruments used in the social sciences (Andrich, 1988; Rasch, 1980). As such, the algebraic expression of the model operationalises an ideal form of measurement. Rasch analysis has been described as a fundamentally different paradigm of measurement, relative to statistical approaches such as Structural Equation Modelling (SEM) (Andrich, 2004). In the context of statistical approaches, analysts typically resolve a lack of fit between a statistical model and the data by revising the model so that it accounts for a greater amount of variability in the dataset. Rasch analysts typically resolve a lack of fit between the Rasch Model and the data by revising the test design, items, or scores so that the data more closely approximates the ideal form of measurement as operationalised by the model. Rasch analysis aims to determine the extent to which a set of scores approximates the requirements of fundamental measurement by displaying unidimensionality and invariance. Unidimensionality is satisfied if items measure a single underlying construct and variability between persons can be attributed to this construct. Invariance is satisfied if estimates of person ability are independent of the particular items to which they responded, and estimates of item difficulty are independent of the particular persons who responded to them (Rasch, 1980). The Rasch Model transforms a raw ordinal scale into a linear interval scale called the logit scale. The logit scale comprises units called logits, and estimates of person ability, $\beta_n$, and item difficulty, $\delta_i$, are co-located on the logit scale for direct comparative purposes. The logit scale is based on the log-odds of the distance between person ability and item difficulty. This distance is
used to estimate the probability of a person \( n \) with ability \( \beta_n \) responding to \( i \) item with difficulty \( \delta_i \) in ordinal category \( m \).

The Rasch Model can not only be used to evaluate the general psychometric properties of developmental assessments but can also determine the extent to which scores satisfy hard stage requirements. For this reason, the Rasch Model has been employed by developmentalists to evaluate the structural validity of their scoring systems since the 1980s (Andrich & Constable, 1984; Bond, 1994; Bond & Fox, 2015; Dawson-Tunik, 2004; Muller et al., 1999). In a study on cognitive-developmental scores, Rasch parameters take on a particular meaning. The logit scale operationalises the underlying construct of hierarchical complexity. Strictly speaking, the logit scale is a scale of difficulty. However, hierarchical complexity is an instantiation of difficulty because more hierarchically complex performance is also more difficult. Person ability, \( \delta_n \), reflects persons’ functional level of reasoning; ordinal categories operationalise Lectical Phases; and thresholds, \( \tau_n \), create partitions between successive Lectical Phases, \( P \) to \( P+1 \). Given a person \( n \) who possesses functional reasoning skills at a level of hierarchical complexity \( h \) and item \( i \) in a Lectical Assessment which has a difficulty of \( \delta_i \), the Rasch Model estimates the probability of person \( n \)’s response to item \( i \) being scored at each Lectical Phase from 10c through to 12b.

Because of the complexity of the analyses presented in Study 1A, Figures 6 and 7 provide a short graphical summary of how scores are expected to behave if they do (Figure 6) and do not (Figure 7) satisfy hard stage requirements. These expectations are derived deductively from the theoretical postulates of hard stage models and inductively from the empirical findings on scores based on hard stage models, as discussed in the literature review above. This pattern of results appears to be relatively unique to scores that successfully operationalise hard stages and is not typically observed in other types of ordinal scales e.g., Andrich and Marais (2019), Hagquist et al. (2009); Steele and Day (2020).
**Figure 6**

*Illustrative Example of Scores that Do Satisfy Hard Stage Requirements in a Rasch Analysis*

Notes. A: Scores display unidimensionality by tapping a single underlying construct and this construct suffices to characterise differences between persons. B: Thresholds are correctly sequenced for all items, ensuring that hard stages are also correctly sequenced. C: Successive thresholds are spaced sufficiently apart to flank qualitatively distinct hard stages. D: Thresholds are aligned across items which ensures that hard stages are also aligned across the logit scale and therefore reflect a similar amount of development. E: Response distributions are narrowly distributed because a single stage organises performance in a domain. There is convergence between adjacent stages but not between nonadjacent stages.
Illustrative Example of Scores that Do Not Satisfy Hard Stage Requirements in a Rasch Analysis

Notes. A: Scores display multidimensionality by tapping the main construct in addition to secondary constructs and/or noise, so the main construct is insufficient to characterise differences between persons. Persons may need to be provided with 2+ scores to adequately summarise their performance. B: Thresholds are incorrectly sequenced for some items, so stages are incorrectly sequenced and not operating as intended. C: Successive thresholds are not spaced sufficiently to flank qualitatively distinct hard stages. D: Thresholds are poorly aligned across items, so stages are misaligned across the logit scale and therefore reflect differing amounts of development. This can be seen in the misaligned location for Stage 6. E: Response distributions are broadly distributed because there is not a single stage which organises performance in a domain. There is convergence between adjacent and nonadjacent stages.
Rationale for Study 1A

The patterns identified in scores that reflect hierarchically integrated stages are not typically found in scores that reflect other types of ordinal scales. Scores at the scale of whole complexity levels have been shown to satisfy hard stage requirements (Commons, Li, et al., 2014; Dawson et al., 2005), but the properties of scores at a lower level of granularity have not been consistently evaluated in the published literature. If Lectical Phase Scores are demonstrated to exhibit these patterns, there may be significant implications for cognitive-developmental theory, empirical research, and practice. From the perspective of theory, identification of these unique patterns may tentatively suggest that hierarchical integrations—or processes akin to hierarchical integrations—inhere at a lower level of granularity than whole complexity levels. If this were considered a tenable line of theorising, then this may be consistent with the postulates of related theories in which smaller increments of development have been proposed to account for the transformation of performance from one complexity level to the next (Commons et al., 2005; Commons & Richards, 2002; Fischer, 1980; Fischer & Bidell, 2006; Lahey et al., 2001; Ross, 2008a). From the perspective of measurement, identification of these unique patterns may suggest that the theoretical postulate that cognition develops hierarchically has been successfully operationalised for the purpose of measurement. Not only this, but it may suggest that Lectical Phase scores are well suited to making inferences about leaders’ ability to navigate complexity and that increments of development that are smaller than whole complexity levels can be detected. In the context of empirical research, this could be useful because it might suggest that finer within-person and between-person distinctions can be made in cognitive development and these distinctions may be predictive of other leadership phenomena and/or may be used to identify differences in interventional and longitudinal studies. From the perspective practice, identification of these unique patterns may suggest that cognitive-developmental approaches can legitimately be applied in workplace contexts to support leaders with issues that are specifically related to complexity. For example, it may be possible for practitioners to create more optimal conditions for leader development by providing leaders with
practices that are tailored to their learning needs. Lectical Assessments provide feedback and developmental practices one Lectical Phase beyond leaders’ current level of reasoning to stimulate the process of reflective abstraction. These practices may be experienced as more palatable and developmentally efficacious than practices aimed one whole complexity level above their current level of reasoning, which may be too challenging for them to grasp.

**Criteria for Study 1A**

In a Rasch analysis, a *family approach* is typically employed in which the outcomes of various analytical procedures are considered (Smith & Plackner, 2009). These analytical procedures typically include visual analysis, statistical analysis, and considering fit statistics and critical values. Unlike traditional statistical approaches, only some of these procedures are associated with significance tests. Rasch analyses conducted on developmental data do not necessarily specify hypotheses to be tested with significance tests e.g., Dawson et al. (2005). For this reason, the discussion below is framed in terms of criteria to be met for evaluation (Commons & Pekker, 2008; Dawson et al., 2005; Kohlberg & Armon, 1984; Loevinger, 1986), rather than research hypotheses.

- **Criterion 1:** Scores satisfy the requirement of unidimensionality.
  - Justification: To be consistent with the Rasch literature, a family approach is taken to evaluating unidimensionality (Andrich, 2016; Andrich & Marais, 2019; Marais & Andrich, 2008b). This approach this takes the findings from a range of analytical procedures into account. A detailed description of these procedures is provided in the Results section for Study 1A below.

- **Criterion 2:** Invariant sequence. For each item, threshold $\tau_n <$ threshold $\tau_{n+1}$.
  - Justification: Thresholds $\tau_n$ create partitions between successive Lectical Phases, P to P+1. Each item is associated with estimates for each threshold. Lectical Phases are sequenced correctly if thresholds are correctly sequenced across the logit scale.
• Criterion 3: Qualitatively distinct.
  o Criterion 3a) For all thresholds, $\tau_n$, the highest estimate for threshold $\tau_n < \text{the lowest estimate for threshold } \tau_{n+1}$.
    ▪ Justification: Thresholds $\tau_n$ create partitions between successive Lectical Phases, P to P+1. Lectical Phase P is qualitatively distinct from Lectical Phase P+1 if the thresholds flanking Phase P are sufficiently spaced apart. The highest estimate for threshold $\tau_n$ should be significantly < the lowest estimate for threshold $\tau_{n+1}$.
  o Criterion 3b) The Person Separation Index (PSI) is equal to or greater than .97, which indicates that eight statistically distinct Lectical Phases can be measured.
    ▪ Justification: Ordinal categories operationalise Lectical Phases. There must be consistency between the number of Lectical Phases the LAS claims to measure i.e., 10c through 12b, and the number of phases it can measure based on reliability indices.

• Criterion 4: Structure of the whole.
  o Criterion 4a) Category probability curves reach an apex of > .80
    ▪ Justification: Category probability curves reflect the probability of a response being scored at a certain Lectical Phase. According to the requirement of structured wholeness, a single Lectical Phase is expected to organise most reasoning in a domain. Category probability curves > .80 indicate a high likelihood of a single Lectical Phase organising responses to LDMA items.
  o Criterion 4b) Category probability curves only converge with adjacent curves, and convergence between nonadjacent curves should occur with a probability of < .10.
- Justification: If a single Lectical Phase organises most reasoning in a domain, a minority of reasoning may occur in adjacent phases. Category probability curves should only converge with nonadjacent curves with a probability of < .10.

- Criterion 5: Hierarchical integrations. Item equating graphs exhibit spurts and plateaus of development.
  - Justification: As discussed above, Lectical Phase P+1 is considered to be a hierarchical integration of its predecessor Lectical Phase P if and only if Lectical Phase P+1 is defined in terms of two or more elements of Lectical Phase P, it organises and transforms the elements of Lectical Phase P, and does so in a non-arbitrary way (Commons & Pekker, 2008). Hierarchical integrations result in reasoning that consolidates at a single Lectical Phase before vacillating between the current phase and its successor. This pattern is reflected in spurts and plateaus of growth in item equating graphs (refer to Figure 5).

Method

Participants

Lectica had previously collected the data between the years 2006 and 2016. The original sample was extracted from Lectica’s database of over 40,000 cognitive-developmental assessments and consisted of 1,913 leaders. It included persons who occupied a leadership role, had participated in a leader development initiative, and had completed at least one Lectical Leadership Decision-Making Assessment (LDMA). Leaders consented to their data being shared with third parties for legitimate research purposes via Lectica’s registration process which includes privacy and legal statements (refer to Appendix C). An optimal sample with sufficient power to identify misfit was approximated to be 525 (i.e., n = Total thresholds x 15, where five items x (8 phases - 1) x 15). A sub-sample of 750 LDMAs was randomly selected to take advantage of the available cases and to
increase the likelihood of detecting misfit. When data were collected, test-takers were provided with a binary option to select for sex. Given that data were collected from leaders in a variety of Western countries, some level of cultural representativeness is suggested. Demographic details for this subsample are presented in Table 4.

Table 4

Demographic Details of Leaders for Study 1A

<table>
<thead>
<tr>
<th>Demographics</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>452</td>
<td>60%</td>
</tr>
<tr>
<td>Female</td>
<td>260</td>
<td>35%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>38</td>
<td>5%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>497</td>
<td>66%</td>
</tr>
<tr>
<td>Non-Caucasian</td>
<td>149</td>
<td>20%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>104</td>
<td>14%</td>
</tr>
<tr>
<td>First language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>485</td>
<td>65%</td>
</tr>
<tr>
<td>Not English</td>
<td>221</td>
<td>29%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>44</td>
<td>6%</td>
</tr>
<tr>
<td>Age at test time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18–27 years</td>
<td>62</td>
<td>8%</td>
</tr>
<tr>
<td>28–37 years</td>
<td>198</td>
<td>26%</td>
</tr>
<tr>
<td>38–47 years</td>
<td>189</td>
<td>25%</td>
</tr>
<tr>
<td>48–57 years</td>
<td>188</td>
<td>25%</td>
</tr>
<tr>
<td>58–68 years</td>
<td>44</td>
<td>6%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>69</td>
<td>9%</td>
</tr>
<tr>
<td>Educational attainment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>436</td>
<td>58%</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>225</td>
<td>30%</td>
</tr>
<tr>
<td>Masters</td>
<td>49</td>
<td>7%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>40</td>
<td>5%</td>
</tr>
</tbody>
</table>

Materials

Assessments. The LDMA is a cognitive-developmental assessment that measures the hierarchical complexity of reasoning applied to leadership decision-making. The LDMA was administered online and presented leaders with a semi-structured, real-life, and complex dilemma.
that prompted their reasoning. Dilemmas were constructed to elicit deliberation upon multiple considerations, such as psychological, behavioural, ethical, and other factors. LDMA dilemmas were pre-selected from a menu of options to ensure relevance to leaders’ seniority, industry, and organisational context. An example of an LDMA dilemma for a middle management role is, “Your team has become extremely dependent on its sole contracting expert. If she left, you know that it would be nearly impossible to replace her, and your organisation would be in serious trouble. You have insider information that management believes that she is not promotable. You were in a meeting where this was discussed, and it was made very clear that this information was not to leave the room. The contracting person trusts you and comes to you for career advice”.

Test-takers assume the role of the protagonist and address the dilemma through a series of five items that require short essay-type responses. Items invite test-takers to provide judgments (‘what do you think?’) and justifications (‘why do you think what you think?’) of sufficient length to score. The same five LDMA items are used irrespective of which dilemma is selected: item 1: “What are the important things to consider in this situation? In one or two paragraphs, explain what they are and why they are important. (In your responses to all questions, feel free to include considerations that go beyond the immediate situation.)”; item 2: “Are some of the considerations you discussed in your response to question 1 more important than others? If so, what are they and why are they more important? If not, why?”; item 3: “What do you think is an appropriate response to this kind of situation? Please explain why this response is appropriate.”; item 4: “Describe another reasonable response to this kind of situation. Compare the potential risks and benefits of this response with those of your original response.”; and item 5: “What process would you recommend for deciding how to respond to situations of this kind? Please describe this decision-making process in general terms—in a way that would allow another person to use the process in a similar workplace situation—and explain why you would recommend each step in this process. Be sure to include a description of how you would actually make your decision and explain why you would make it in that way”.
**Software.** Analyses were based on the Rasch Model (Andrich, 1988; Rasch, 1980) and conducted in the Rasch Unidimensional Measurement Model (RUMM2030) software (Andrich et al., 2015).

**Procedure**

**Administration.** Lectical Assessments had been administered through various accredited providers whose identities cannot be disclosed due to contractual agreements with Lectica.

**Scoring.** As discussed above, Lectica uses two scoring systems—the LAS and the CLAS. The purpose of both systems is to locate the magnitude of hierarchical complexity exhibited in samples of reasoning along the Lectical Scale. The LAS was used to award a Lectical Phase Score to each LDMA item. The four Lectical Analysts who scored the sample had a minimum of four years training in Lectical analysis via a professional internship program and also displayed scoring skills that satisfied the requisite psychometric criteria (Lectica, 2018b).

**Data Preparation.** Eight Lectical Phases were identified in the sample (i.e. 10c through to 12b), and these were entered as categories 0 to 7 in RUMM2030 (Andrich et al., 2015). Only assessments completed at test-time 1 were included (Marais & Andrich, 2008a).

**Use of The Rasch Model to Analyse Lectical Phase Scores**

The Rasch Model for polytomous items with greater than two ordered categories is an extension of the model for dichotomous items (Andersen, 1977; Andrich, 1978; Rasch, 1961). Rating scale (Andrich, 1978) and partial credit (Masters, 1982) parameterisations of the model were considered for application. The rating scale parameterisation is typically more applicable to survey-based data for which categories are defined by identical descriptors, e.g., “strongly agree”. In the context of Study 1A, the rating scale parameterisation would assume, but not demonstrate, that Lectical Phase Scores satisfied many hard stage requirements because this parameterisation
generates equidistant thresholds. The partial credit parameterisation was selected so hard stage requirements could be explicitly tested rather than assumed (Andrich, 2005, pp. 303-304):

\[ \Pr(X_{ni} = x) = \frac{\exp(x(\beta_n - \delta_i) \cdot \sum_{k=0}^{x} \tau_k)}{\gamma_{ni}} \]

Where:

(i) \( x \in \{0, 1, 2, \ldots m_i\}\) is the integer response variable for the person \( n \) with ability \( \beta_n \) responding to item \( i \) with difficulty \( \delta_i \),

(ii) \( \tau_{1i}, \tau_{2i}, \ldots, \tau_{mi}, \sum_{x=0}^{m_i} \tau_{xi} = 0, \tau_{0i} \equiv 0 \), are the thresholds between \( m_i + 1 \) ordered categories for item \( i \), \( m_i \) is the maximum response score of item \( i \),

(iii) \( \gamma_{ni} = \sum_{x=0}^{m_i} \left[ \exp(x(\beta_n - \delta_i) \cdot \sum_{k=0}^{x} \tau_k) \right] \) is a normalising factor ensuring that \( \sum_{x=0}^{m_i} \Pr(X_{ni} = x) = 1 \),

(iv) values for \( \beta, \delta \) and \( \tau \) are located additively on the same unidimensional scale,

(v) the threshold \( \tau_{0i} \) does not exist, is included for notational convenience.

**Analytical Approach**

Analyses were undertaken in four steps, as described in Table 5.
Table 5

Steps Applied to the Rasch Analysis of the LDMA

<table>
<thead>
<tr>
<th>Step</th>
<th>Purpose</th>
<th>No.</th>
<th>Analysis name</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Confirm if the whole dataset could be analysed simultaneously</td>
<td>1a</td>
<td>Facet analysis of LDMA scorer severity</td>
<td>To determine whether the 4 LDMA scorers awarded equivalent scores to test-takers of equal ability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1b</td>
<td>Retention of Lectical Phases and LDMA items</td>
<td>To determine if the 8 Lectical Phases and 5 LDMA items could be retained for further analysis</td>
</tr>
<tr>
<td>2</td>
<td>Summary statistics</td>
<td>2</td>
<td>Summary statistics and general fit to the Rasch Model</td>
<td>To provide general summary statistics for the final dataset and present basic fit statistics</td>
</tr>
<tr>
<td>3</td>
<td>Direct tests of hard stage requirements</td>
<td>3a</td>
<td>Unidimensionality</td>
<td>To determine if there was a single underlying factor measured by LDMA items i.e., hierarchical complexity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3b</td>
<td>Invariant sequence</td>
<td>To determine if LDMA item thresholds and Lectical Phases occurred in the sequence predicted by theory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3c</td>
<td>Qualitatively distinct</td>
<td>To determine if LDMA thresholds were sufficiently spaced apart to create partitions between qualitatively distinct Lectical Phases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3d</td>
<td>Structure of the whole</td>
<td>To determine if a single Lectical Phase organised each test-taker’s response to LDMA items</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3e</td>
<td>Hierarchical integrations</td>
<td>To determine if Lectical Phases exhibited spurts and plateaus as predicted by hard stage requirements</td>
</tr>
<tr>
<td>4</td>
<td>Test for other general measurement properties</td>
<td>4</td>
<td>Differential item functioning, response dependence, measurement precision</td>
<td>Additional analyses of general measurement properties not directly related to hard stage requirements</td>
</tr>
</tbody>
</table>
Results

**Step 1: Confirm if the Whole Dataset Could be Analysed Simultaneously**

The purpose of Step 1 was to confirm if the whole dataset could be analysed simultaneously.

**Step 1a: Facet Analysis of LDMA Scorer Severity.** This analysis aimed to determine whether the four LDMA scorers awarded equivalent scores to test-takers of equal ability. If the scorers were equally severe, then the whole dataset could be analysed simultaneously to maximise statistical power. This was undertaken by a facet analysis that introduced a third facet, scorer severity, to the two facets included in the standard Rasch Model i.e., person ability and item difficulty (Linacre, 1989; Linacre & Wright, 2002; Lunz et al., 1990). Scorer identities were not available for some cases in the subsample of 750. As a result, a subsample of 800 cases was randomly selected from the original sample (n=1,913) so that it would approximate the size of the original subsample (i.e., 750 cases) and also include an equal number of cases per scorer to mitigate bias (n = 200 cases/scorer).

As seen from Table 6, scorer severity occurred with a narrow range of 0.05 logits and fit residuals occurred close to zero. A lack of fit is indicated if scorer severity for one or more scorers significantly differs from other scorers and if the fit residual exceeds the critical value $+2.5$. Findings suggest that scorers were equally severe, so the whole dataset could be analysed simultaneously.

<table>
<thead>
<tr>
<th>Scorer</th>
<th>N</th>
<th>Scorer severity ($\lambda_h$)</th>
<th>SE</th>
<th>Fit Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>0.02</td>
<td>0.20</td>
<td>1.88</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>0.00</td>
<td>0.14</td>
<td>-0.22</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>-0.03</td>
<td>0.21</td>
<td>0.81</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>0.01</td>
<td>0.19</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Step 1b: Retention of Lectical Phases and LDMA Items. This step aimed to determine whether all eight Lectical Phases and five LDMA items could be retained for further analysis. Decisions were made based on three considerations: utilisation of Lectical Phases, targeting of thresholds to person ability, and correct threshold sequencing. The response distribution across Lectical Phases is presented in Table 7.

Table 7

Response Distribution Across Lectical Phases 10c to 12b

<table>
<thead>
<tr>
<th>Variable</th>
<th>10c</th>
<th>10d</th>
<th>11a</th>
<th>11b</th>
<th>11c</th>
<th>11d</th>
<th>12a</th>
<th>12b</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>126</td>
<td>345</td>
<td>741</td>
<td>1337</td>
<td>899</td>
<td>200</td>
<td>85</td>
<td>17</td>
<td>3750</td>
</tr>
<tr>
<td>Percentage</td>
<td>3.36</td>
<td>9.20</td>
<td>19.76</td>
<td>35.65</td>
<td>23.97</td>
<td>5.33</td>
<td>2.27</td>
<td>0.45</td>
<td>100.00</td>
</tr>
<tr>
<td>Cumulative percentage</td>
<td>3.36</td>
<td>12.56</td>
<td>32.32</td>
<td>67.97</td>
<td>91.95</td>
<td>97.28</td>
<td>99.55</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

The distribution appeared to approximate a normal distribution, which is unnecessary for Rasch analysis but typical in leadership samples. Some underutilisation occurred in the highest phase (12b) with less than 1% of responses, but this alone did not suggest reducing the number of Lectical Phases included in the analysis. Uncentralised thresholds were utilised throughout Study 1 rather than centralised thresholds because uncentralised thresholds are referenced to the same origin, so they can be compared across items rather than being mean deviated about each item’s location. The person item-threshold distribution (refer to Figure 8) shows the logit scale on the horizontal axis with person ability plotted in the upper section and item thresholds in the lower section. Item thresholds occurred over a range of 31.87 logits and person ability occurred over a range of 33.14 logits (refer to Figure 8). This suggests that thresholds appropriately targeted person ability because all persons could be allocated to a Lectical Phase.
Figure 8

*Person-Item Threshold Distribution for the LDMA*

Lectical Phases are correctly sequenced if thresholds are correctly sequenced across the logit scale. No items displayed threshold disorder despite underutilisation of the highest phase 12b, as shown in Table 8.

Table 8

*Uncentralised Threshold Locations for LDMA Items*

<table>
<thead>
<tr>
<th>LDMA Item</th>
<th>Item location</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_3$</th>
<th>$\tau_4$</th>
<th>$\tau_5$</th>
<th>$\tau_6$</th>
<th>$\tau_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.37</td>
<td>-14.86</td>
<td>-10.05</td>
<td>-5.14</td>
<td>-0.06</td>
<td>5.23</td>
<td>10.78</td>
<td>16.67</td>
</tr>
<tr>
<td>2</td>
<td>0.08</td>
<td>-14.41</td>
<td>-9.84</td>
<td>-4.87</td>
<td>0.27</td>
<td>5.31</td>
<td>10.00</td>
<td>14.09</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>-14.33</td>
<td>-10.23</td>
<td>-5.20</td>
<td>0.27</td>
<td>5.71</td>
<td>10.63</td>
<td>14.55</td>
</tr>
<tr>
<td>4</td>
<td>0.33</td>
<td>-14.79</td>
<td>-10.16</td>
<td>-5.09</td>
<td>0.23</td>
<td>5.60</td>
<td>10.83</td>
<td>15.72</td>
</tr>
<tr>
<td>5</td>
<td>-0.98</td>
<td>-15.20</td>
<td>-10.72</td>
<td>-5.89</td>
<td>-0.91</td>
<td>4.04</td>
<td>8.76</td>
<td>13.06</td>
</tr>
</tbody>
</table>
The threshold probability curve for item 1 is shown in Figure 9 where it can be observed that all thresholds are correctly sequenced and spaced relatively evenly.

Figure 9

Threshold Probability Curve for LDMA Item 1

In summary, although phase 12b was underutilised the evidence provided support for retaining all eight Lectical Phases.

Decisions about item retention were based on two considerations: alignment of Lectical Phases across items and item fit. Alignment of Lectical Phases across LDMA items is important for Lectical Phases to reflect equivalent amounts of hierarchical complexity and, therefore, to have a consistent meaning. As can be seen from Figure 10, phases 10c through to 11d are relatively well aligned across items, but there is greater misalignment for phases 12a and 12b. However, this did not appear sufficiently severe to warrant rescoring categories or deleting items. At the time of writing, LDMAs at 12a and 12b have already been rescored by Lectica as a result of these analyses.
Statistical and visual analyses of item fit were also undertaken. However, since no items showed sufficient misfit to warrant deletion, findings are presented in the context of the full Rasch analysis below.

**Step 2: Summary Statistics**

This step aimed to provide general summary statistics and basic fit statistics. Summary statistics demonstrated that mean person ability was lower (M=-2.83, SD=6.05, min=-16.74, max=16.40, range=33.14) than average item difficulty (M=0.00, SD=0.56, min=-0.98, max=0.37, range=1.35) (refer Figure 11). This suggested that many leaders experienced the LDMA items to be somewhat challenging. Person abilities were broadly distributed over 33.14 logits, whereas the item difficulties were narrowly distributed over 1.35 logits. The range of item difficulties represents 4% of the range of person abilities and potential implications are explored in the discussion for Study 1A. The broad range of person abilities suggests that the LDMA distinguishes between test-takers of different abilities by distributing them along the logit scale.

---

### Figure 10

**Threshold Map of 5 LDMA Items from 10c to 12b**

<table>
<thead>
<tr>
<th>Descriptor for Item 1</th>
<th>10c</th>
<th>10d</th>
<th>11a</th>
<th>11b</th>
<th>11c</th>
<th>11d</th>
<th>12a</th>
<th>12b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Descriptor for Item 2</th>
<th>10c</th>
<th>10d</th>
<th>11a</th>
<th>11b</th>
<th>11c</th>
<th>11d</th>
<th>12a</th>
<th>12b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Descriptor for Item 3</th>
<th>10c</th>
<th>10d</th>
<th>11a</th>
<th>11b</th>
<th>11c</th>
<th>11d</th>
<th>12a</th>
<th>12b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Descriptor for Item 4</th>
<th>10c</th>
<th>10d</th>
<th>11a</th>
<th>11b</th>
<th>11c</th>
<th>11d</th>
<th>12a</th>
<th>12b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Descriptor for Item 5</th>
<th>10c</th>
<th>10d</th>
<th>11a</th>
<th>11b</th>
<th>11c</th>
<th>11d</th>
<th>12a</th>
<th>12b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>
Fit residuals for items indicated overfit to the model because they deviate from a mean of 0.00 and SD of 1.00 (M = -3.00, SD = 0.89, skewness = -0.15, kurtosis = -2.02). Fit residuals for persons also seemed to indicate slight overfit (M = -0.88, SD = 1.42, skewness = -0.22, kurtosis = -1.32).

**Item Fit.** Item fit was analysed statistically and graphically. Statistical analyses were undertaken via fit residuals, chi-square, and ANOVA statistics. Graphical analysis was undertaken via Item Characteristic Curves (ICCs). An item was considered to misfit if the fit residual exceeded the critical value of ±2.5, or if the chi-square and/or ANOVA fit statistic exceeded the critical value for alpha with a Bonferroni correction (.05/5 items = .01) (RUMM Laboratory, 2009a). Fit statistics are presented in Table 9.
Table 9

*Item Fit Statistics for LDMA Items*

<table>
<thead>
<tr>
<th>Item</th>
<th>Location</th>
<th>SE</th>
<th>Fit Residual</th>
<th>DF</th>
<th>$\chi^2$ statistic</th>
<th>DF</th>
<th>p</th>
<th>F statistic</th>
<th>DF1</th>
<th>DF2</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.37</td>
<td>0.08</td>
<td>-1.97</td>
<td>589.80</td>
<td>5.23</td>
<td>9</td>
<td>.81</td>
<td>0.86</td>
<td>9</td>
<td>732</td>
<td>.56</td>
</tr>
<tr>
<td>2</td>
<td>0.08</td>
<td>0.08</td>
<td>-2.69</td>
<td>589.80</td>
<td>9.83</td>
<td>9</td>
<td>.36</td>
<td>1.79</td>
<td>9</td>
<td>732</td>
<td>.07</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>0.08</td>
<td>-3.70</td>
<td>589.80</td>
<td>5.08</td>
<td>9</td>
<td>.83</td>
<td>0.99</td>
<td>9</td>
<td>732</td>
<td>.45</td>
</tr>
<tr>
<td>4</td>
<td>0.33</td>
<td>0.08</td>
<td>-4.12</td>
<td>589.80</td>
<td>14.19</td>
<td>9</td>
<td>.12</td>
<td>2.67</td>
<td>9</td>
<td>732</td>
<td>.00</td>
</tr>
<tr>
<td>5</td>
<td>-0.98</td>
<td>0.08</td>
<td>-2.52</td>
<td>589.80</td>
<td>14.17</td>
<td>9</td>
<td>.12</td>
<td>2.12</td>
<td>9</td>
<td>732</td>
<td>.03</td>
</tr>
</tbody>
</table>

As can be seen from Table 9, items 2, 3, 4 and 5 demonstrated overfit on at least one statistical test, but there were no instances of underfit. Item 1 demonstrated no evidence of misfit. Item 2 overfit the model according to the fit residual statistic (-2.69) but not according to the chi-square ($\chi^2(9)=9.83$, p=.36) or ANOVA fit statistic (F(9,732)=1.79, p=.07). Item 3 overfit the model according to the fit residual statistic (-3.70) but not according to the chi-square ($\chi^2(9)=5.08$, p=.83) or ANOVA fit statistic (F(9,732)=0.99, p=.45). Item 4 overfit the model according to the fit residual statistic (-4.12) and the ANOVA fit statistic (F(9,732)=2.67, p<.001), but not according to the chi-square fit statistic ($\chi^2(9)=14.19$, p=.12). Finally, item 5 overfit the model according to the fit residual statistic (-2.52), but not according to the chi-square ($\chi^2(9)=14.17$, p=.12) or ANOVA fit statistic (F(9,732)=2.12, p=.03. Visual analysis of the Item Characteristic Curves (ICCs) indicated negligible deviation between observed and predicted values. An illustrative example of an ICC for Item 1 is presented in Figure 12.
Person Fit. Eight persons with extreme response patterns were identified, representing 1.07% of the sample. Examination of the Guttman pattern for these eight persons illustrated that they consistently scored in the lowest phase for all five items (i.e., 10c). Although this pattern is uncommon in adults and may not fit the Rasch Model, it is plausible in the context of developmental theory. Relative to the critical value of ±2.5, 188 cases (25%) overfit the model which is greater than what would be expected by chance alone (>5%), but there were no cases of underfit. Overfit occurred when test-takers were awarded identical scores on all items. This pattern does not fit the Rasch Model but is consistent with the requirements of structured wholeness.

Step 3: Direct Tests of Hard Stage Requirements

The purpose of Step 3 was to directly test whether the Lectical Phase Scores satisfy hard stage requirements.

Step 3a: Unidimensionality. This analysis aimed to determine if LDMA items measured a common construct i.e., hierarchical complexity. Unidimensionality is satisfied if items measure a single construct and if a total score suffices to account for test-takers’ performance.
Unidimensionality requires that for any item of difficulty $\delta_i$, variation among responses to that item can only be attributed to person ability, $\beta_n$. Multidimensionality occurs when a single score is insufficient to account for test-takers’ performance and implies the presence of secondary constructs and/or noise (Marais & Andrich, 2008a, 2008b). Secondary constructs typically need to be measured via a subset of items and are reflected in separate scores. The analysis was undertaken in three steps: a) Principal Component Analysis (PCA); b) subtest analysis; c) item equating (RUMM Laboratory, 2009b).

**PCA.** A PCA was conducted on the residuals to determine if the remaining variability was random after extracting the Rasch factor. Secondary constructs are identified when the remaining variability forms patterns with theoretical relevance. From principal component 1 to principal component 2, there was a relatively negligible decrease in eigenvalues of 0.24 and a negligible decrease of 4.75% in the percentage variability explained. This was marginally greater than the differences between successive principal components. This suggested that principal component 1 did not reflect a secondary construct.

<table>
<thead>
<tr>
<th>Principal Component</th>
<th>Eigenvalue</th>
<th>Percentage variability explained</th>
<th>Cumulative percentage</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.50</td>
<td>30.01%</td>
<td>30.01%</td>
<td>0.19</td>
</tr>
<tr>
<td>2</td>
<td>1.26</td>
<td>25.26%</td>
<td>55.28%</td>
<td>0.14</td>
</tr>
<tr>
<td>3</td>
<td>1.17</td>
<td>23.43%</td>
<td>78.71%</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>1.05</td>
<td>20.90%</td>
<td>99.61%</td>
<td>0.13</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>0.39%</td>
<td>100.00%</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Items were sorted according to their loadings on principal component 1 to determine whether they revealed a theoretically relevant pattern. As can be seen from Table 11, items 1 and 2 loaded positively and these items related to general considerations about the LDMA dilemma. Items 3, 4, and 5 loaded negatively, and these items related to the provision of solutions to the LDMA
dilemma. These loadings appeared meaningful, so a subtest analysis was undertaken to estimate the amount of multidimensionality in the data.

**Table 11**

*LDMA Item Loadings on Principal Component 1 (Organised by Strength of Loading)*

<table>
<thead>
<tr>
<th>LDMA item</th>
<th>Loading on principal component 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.656</td>
</tr>
<tr>
<td>1</td>
<td>0.641</td>
</tr>
<tr>
<td>3</td>
<td>-0.318</td>
</tr>
<tr>
<td>5</td>
<td>-0.489</td>
</tr>
<tr>
<td>4</td>
<td>-0.565</td>
</tr>
</tbody>
</table>

**Subtest Analysis.** A subtest analysis approximates the amount of multidimensionality in the data and, therefore, the extent to which a set of scores deviates from unidimensionality. In a subtest analysis, two sets of estimates are compared. All five LDMA items were used for the first estimate, and they were assumed to be statistically independent. For the second estimate, items 1 and 2 were combined to form a higher-order item based on loadings on principal component 1, and items 3, 4, and 5 were combined to form a second higher-order item based on loadings on principal component 1. Therefore, the second estimate only contained these two higher-order items. Higher-order items are typically constructed to have an equivalent number of items +1 to ensure the stability of estimates.

Multidimensionality is detected in four ways (RUMM Laboratory, 2009b). First, a significant drop in coefficient alpha and/or the Personal Separation Index (PSI) between the first and second estimates indicates that the reliability of the first estimate was inflated because of multidimensionality. PSI is an index of reliability analogous to coefficient alpha and estimates the amount of measurement error associated with a set of scores. PSI and alpha are determined by the ratio of true score variance to total score variance. Unlike coefficient alpha, the PSI is based on Rasch-transformed interval scores rather than raw ordinal scores and provides a more stringent
estimate (Andrich, 1982; Dawson et al., 2005; Guilford, 1965; Kubiszyn & Borich, 1993; Wright, 1996). Second, $c$ is an index of the amount of multidimensionality in a dataset. Strictly speaking, unidimensionality is violated if $c > 0$, however scores that approach 0 may be regarded as acceptable when considered in the context of other analyses of unidimensionality (Andrich, 2016; Andrich & Marais, 2019; Marais & Andrich, 2008b). Third, $\rho$, is an index of the correlation between the two higher-order items from the second estimate. A high correlation suggests that the two higher-order items tap the same unidimensional construct. Fourth, $A$, is the proportion of non-error variance common to the two higher-order items. A high value reflects a high amount of true common variance between the two higher-order items from the second estimate. As shown in Table 12, the reliability indices from the first estimate did not seem to be inflated. The amount of multidimensionality appeared to be negligible, given that $c = .01$. The correlation between the two higher-order items from the second estimate, $\rho$, was very high between .99–1.0. Finally, the non-error variance between the two higher-order items, $A$, was very high between .95–1.0.

Table 12

Subtest Analysis on the Magnitude of Multidimensionality in LDMA Items

<table>
<thead>
<tr>
<th>Reliability index</th>
<th>First estimate</th>
<th>Second estimate</th>
<th>Magnitude of multidimensionality ($c$)</th>
<th>Correlation between subscales ($\rho$)</th>
<th>Common non-error variance ($A$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person Separation Index (PSI)</td>
<td>.97</td>
<td>.97</td>
<td>Imaginary number (i)</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Coefficient alpha ($\alpha$)</td>
<td>.98</td>
<td>.93</td>
<td>.01</td>
<td>.99</td>
<td>.95</td>
</tr>
</tbody>
</table>

**Item Equating.** As a final way of examining multidimensionality, an item equating procedure was undertaken (RUMM Laboratory, 2009b). For scores to be unidimensional, all items must tap a common construct and different sets of items should award equivalent scores to persons of equal ability. Item equating is a Rasch procedure that identifies whether different sets of items give persons of equal ability equivalent scores. Based on the item loadings on principal component 1,
items 1 and 2 were included in the first set, and items 4 and 5 were included in the second. A paired sample t-test for persons demonstrated that the scores on the first set of items were not significantly different to scores on the second set of items where $t(749)=1.17, p=.24$. The correlation between the two sets of scores was significant and high ($r=.93, r^2=.87, n=750, p<.001$). Considered collectively, analyses support the conclusion that Lectical Phase Scores satisfy the requirement of unidimensionality.

**Step 3b: Invariant Sequence.** This analysis aimed to determine if item thresholds and, therefore, Lectical Phases occurred in the sequence predicted by theory. In the context of developmental data, thresholds take on a particular meaning because they create the partitions between successive Lectical Phases e.g., where 11a terminates and where 11b begins. Correct sequencing of item thresholds was already established in step 1b, so no further analyses were undertaken.

**Step 3c: Qualitatively Distinct.** This analysis aimed to determine if item thresholds were sufficiently spaced apart to create partitions between qualitatively distinct Lectical Phases. The person item-threshold distribution shows the logit scale on the horizontal axis with person ability plotted in the upper section and threshold estimates in the lower section. As can be seen from Figure 13, thresholds 1 through 5 appeared relatively well aligned and occurred in a narrow range of logits. For thresholds 6 and 7, estimates were more broadly distributed. For all thresholds, the highest estimate for threshold $\tau_n$ appeared to be lower than the lowest estimate for threshold $\tau_{n+1}$. This created the visible ‘white space’ between successive thresholds where Lectical Phases were located. For example, Lectical Phase 11a can be located between thresholds 2 and 3.
Descriptive statistics for thresholds are presented in Table 13.

### Table 13

**Descriptive Statistics for Uncentralised Thresholds (LDMA Items)**

<table>
<thead>
<tr>
<th>Threshold $\tau_n$</th>
<th>Partition between Lectical Phases</th>
<th>M</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
<th>Range</th>
<th>M of SEs</th>
<th>Separation between adjacent thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10c/10d</td>
<td>-14.72</td>
<td>0.36</td>
<td>-15.20</td>
<td>-14.33</td>
<td>0.87</td>
<td>0.18</td>
<td>3.61</td>
</tr>
<tr>
<td>2</td>
<td>10d/11a</td>
<td>-10.20</td>
<td>0.32</td>
<td>-10.72</td>
<td>-9.84</td>
<td>0.88</td>
<td>0.13</td>
<td>3.95</td>
</tr>
<tr>
<td>3</td>
<td>11a/11b</td>
<td>-5.24</td>
<td>0.39</td>
<td>-5.89</td>
<td>-4.87</td>
<td>1.02</td>
<td>0.11</td>
<td>3.96</td>
</tr>
<tr>
<td>4</td>
<td>11b/11c</td>
<td>-0.04</td>
<td>0.50</td>
<td>-0.91</td>
<td>0.27</td>
<td>1.18</td>
<td>0.13</td>
<td>3.76</td>
</tr>
<tr>
<td>5</td>
<td>11c/11d</td>
<td>5.18</td>
<td>0.67</td>
<td>4.04</td>
<td>5.71</td>
<td>1.68</td>
<td>0.22</td>
<td>3.04</td>
</tr>
<tr>
<td>6</td>
<td>11d/12a</td>
<td>10.20</td>
<td>0.87</td>
<td>8.76</td>
<td>10.83</td>
<td>2.07</td>
<td>0.39</td>
<td>2.23</td>
</tr>
<tr>
<td>7</td>
<td>12a/12b</td>
<td>14.82</td>
<td>1.41</td>
<td>13.06</td>
<td>16.67</td>
<td>3.61</td>
<td>0.82</td>
<td>NA</td>
</tr>
</tbody>
</table>

Statistical analyses confirmed the findings from the visual analysis. For thresholds 1 through 5, estimates were well aligned and occurred in a narrow range of logits from 0.87 to 1.68. For thresholds 6 and 7, estimates were more broadly distributed from 2.07 to 3.61. For all thresholds, the highest estimate for threshold $\tau_n$ was lower than the lowest estimate for threshold $\tau_{n+1}$. Also,
95% confidence intervals constructed around the highest estimate for threshold $\tau_n$ never overlapped with 95% confidence intervals constructed around the lowest estimate for threshold $\tau_{n+1}$ (i.e., $2 \times 1.96 \times SE$).

**Reliability and Distinct Strata of Performance.** Reliability indices can estimate the number of statistically distinct strata of performance that an assessment can measure. In the context of developmental data, strata of performance operationalise Lectical Phases. To distinguish between eight Lectical Phases, the LDMA must exhibit a reliability index of at least .97 (Fisher, 1992; Wright, 1996; Wright & Masters, 2002). The LDMA displayed a coefficient alpha of .98 and a PSI of .97. As shown in the unidimensionality analyses, reliability estimates were unlikely to be inflated because of the negligible amount of multidimensionality identified.

Considered collectively, these analyses suggest that thresholds create clear partitions between Lectical Phases and the location of Lectical Phases can be identified along the logit scale in terms of the amount of hierarchical complexity they reflect. The number of Lectical Phases the LAS claims to measure is consistent with the number of Lectical Phases which can be statistically distinguished.

**Step 3d: Structure of the Whole.** This analysis aimed to determine whether a single Lectical Phase organised each leader’s responses to LDMA items. Category probability curves show the logit scale on the horizontal axis and the probability of scoring in successive Lectical Phases on the vertical axis. For example, a person with an ability of -3 would receive a score of 11b (refer to Figure 14). Two patterns emerge if the requirement of structured wholeness is satisfied. First, each curve should reach an apex of > .80. Second, each curve should only converge with adjacent curves, and convergence between nonadjacent curves should occur with a probability of < .10. Category probably curves for all items were analysed and displayed similar patterns. Item 5 is presented in Figure 14.
For item 5, apexes for all Lectical Phases occurred between .81 and .86. Using phase 11b as an example, the probability of convergence with the two adjacent phases, 11a and 11c, was .49. However, the probability of convergence with two nonadjacent phases, 10d and 11d, was between .07-.08. Considered collectively, the evidence suggests that Lectical Phase Scores satisfy the requirement of structured wholeness.

**Step 3d – Hierarchical Integrations.** This analysis aimed to determine if Lectical Phase Scores exhibited spurts and plateaus of development. Item equating is a Rasch procedure which examines the relation between person ability and total score on an assessment. All five items were entered simultaneously into an item equating procedure and item 1 was selected at random for a separate procedure. The item equating graph shows person ability on the horizontal axis and total score on the vertical axis to reveal the ‘shape of development’ (refer to Figures 15 and 16). Both figures demonstrate spurts and plateaus which may suggest a succession of hierarchical integrations between Lectical Phases.
Tests for additional general measurement properties unrelated to hard stage requirements are reported in Appendix D.
Evaluation of Criteria for Study 1A

- Criterion 1: Unidimensionality. The family approach suggested that scores contained a negligible amount of multidimensionality. More specifically, the PCA revealed no additional constructs after the Rasch factor was extracted. There was a negligible difference between the PSIs from the subtest analysis. The magnitude of multidimensionality, $c$, approached 0. There was a very high correlation between the two higher-order items, $\rho$. The true common variance between the two higher-order items, $A$, was very high. The item equating analysis demonstrated that two sets of LDMA items did not award significantly different scores to persons of equal ability.

- Criterion 2: Invariant sequence. This was supported because threshold estimates occurred in the theoretically predicted sequence for all items. This suggest that Lectical Phases are correctly sequenced.

- Criterion 3: Qualitatively distinct.
  - Criterion 3a) This was supported because 95% confidence intervals around the highest estimate for threshold $\tau_n$ were consistently < than 95% confidence intervals around the lowest estimate for threshold $\tau_{n+1}$. As a result, Lectical Phases can be considered qualitatively distinct.
  - Criterion 3b) This was supported because the PSI was .97. This suggests that the LAS can distinguish between eight statistically distinct Lectical Phases.

- Criterion 4: Structure of the whole.
  - Criterion 4a) This was supported because category probability curves reached an apex > .80. This suggests that a single Lectical Phase organised each leader’s responses to LDMA items.
Criterion 4b) This was supported because convergence between nonadjacent curves occurred with a probability of < 0.10. This suggests that reasoning was not broadly distributed across multiple Lectical Phases.

- Criterion 5: Hierarchical integrations. This was supported because spurts and plateaus of development were observed in the item equating graph for all items considered collectively and, in each item, considered individually.

Discussion

All criteria were supported. Lectical Phase Scores awarded by the LAS were shown to satisfy the requirements of unidimensionality, invariant sequence, qualitative distinctness, structured wholeness, and hierarchical integrations. These findings are consistent with Dawson’s (2001, 2002b, 2003b, 2004, 2006; 2005; 2003; 2004) studies on scores at the scale of whole complexity levels and earlier studies on other cognitive-developmental scores (Commons et al., 2008; Davison et al., 1980; Snarey, 1985).

Four explanations may account for these findings. The first is that Lectical Phase Scores satisfy hard stage requirements, which is reflected in how LAS awards scores to LDMA items. However, three competing explanations cannot be ruled out, so conclusions are regarded as provisional.

The second explanation is that these findings are an artefact of Lectica’s scoring procedures. When Lectical Analysts score an LDMA, they score responses to all items from the same LDMA consecutively. As a result, the finding of structured wholeness may result from a particular type of response dependence called the halo effect in which the score given to the first item influences the scores given to other items and/or scorers give similar scores to all items based on a general impression of the test-taker rather than on the basis of their unique response to each item (Humphry & Heldsinger, 2014; Marais & Andrich, 2011). If the halo effect were unique to only some scorers, it would be detectable through the facet analysis of scorer severity in Step 1a above. If the
halo effect were common to all scorers, then it would not be detectable through these standard fit indices. Psychometric analyses have demonstrated that when the magnitude of the halo effect is high, the person-item threshold distribution resembles the person-item threshold distribution reported for the LDMA (refer Figure 13). For an example, refer to Figure 1c which is reported in Marais and Andrich (2011, p. 205). Ruling out the halo effect presents some difficulties because in the context of measuring cognitive-development, it is expected that most scores will cohere within a single stage (or Lectical Phase) because stages / phase are contended to reflect deep structures in the mind-brain. As a result, it is difficult to determine whether the finding of structured wholeness in the LMDA may be a function of response dependence / the halo effect or because of the peculiarities of the construct that is being measured. Nevertheless, ruling out the halo effect is important given that its presence has been shown to result in inflated reliabilities, artificially stretched logit scales, and several other consequences (Marais & Andrich, 2011). One way of ruling out the halo effect might be to determine whether the requirement of structured wholeness is satisfied in the context of different scoring procedures. For example, if Lectical Analysts scored different items from different LDMAs in a randomised sequence rather than reviewing responses to all items from the same LDMA consecutively.

The third explanation is that some of the findings reported for the LDMA may be an artefact of the narrow distribution of average item difficulties. In the context of some educational assessments, dichotomous items (i.e., items with correct and incorrect responses) are constructed to progressively increase in difficulty. This is undertaken to ensure that the distribution of item difficulties (and also item thresholds) effectively target the distribution of person abilities. For these types of educational assessments, estimates of person ability are informed by the most difficult items that can be answered correctly. When item difficulties are broadly distributed, this typically mitigates against floor and ceiling effects i.e., when a significant proportion of test-takers obtain the minimum and maximum score on an assessment and their true ability cannot be determined. Dichotomous items are sometimes constructed to reflect similar average difficulties when they are
administered to test-takers who are expected to have similar abilities as discussed by Andrich and Marais (2019). However, the LDMA is based on a fundamentally different premise in which Lectical Phase scores function as polytomous response categories that reflect the hierarchical complexity of test-takers’ reasoning rather than the correctness of their responses. The narrow range of item difficulties obtained for the LDMA is a peculiarity of some cognitive-developmental assessments in which items are deliberately constructed to reflect an equivalent amount of difficulty. Such assessments are typically informed by Kohlberg’s (1983) argument that test-takers should be provided with an equal opportunity on all items. While a narrow range of item difficulties may be consistent with the expectations of cognitive-developmentalists, further research is required to determine whether a narrow range of difficulties impacts the measurement of cognition. In some contexts, assessments that consist of items with similar difficulties may be regarded as consisting of one item and this may result in compromised measurement properties (Andrich, 1988; Rasch, 1980).

Several psychometric studies have demonstrated that the distribution of item difficulties may impact a range of measurement properties (Feldt, 1993; Robitzsch, 2021; Seo & Weiss, 2013). For example, Robitzsch (2021) analysed a simulated dataset in which item difficulties and a number of other parameters were systemically varied. Findings indicate that the distribution of item difficulties influenced reliability estimates, bias, and root mean square error. However, these studies have not been conducted on assessments such as the LDMA for which there is an expectation that most responses will cohere within a single stage (or Lectical Phase). As a result, the applicability of these studies to the LDMA, is unclear.

As discussed earlier in Chapter 2, the hierarchical complexity of test-takers’ performance has been shown to vary as a function of the hierarchical complexity of tasks administered to test-takers (Fischer & Bidell, 2006). As a result, the hierarchical complexity of test-takers’ responses may be influenced by the hierarchical complexity of the LDMA items. If all LDMA items are constructed to reflect a similar amount of hierarchical complexity, then test-takers’ responses may also reflect a
similar amount of hierarchical complexity, and this may result in similar scores being awarded to all LDMA items. If similar scores are awarded to all LDMA items, this may result in higher reliability estimates than what might be expected if there was more variability in scores. Lectica does not attribute Lectical Scores to LDMA items, but from Figures 11 and 13, the hierarchical complexity of LDMA items appears to be commensurate with the transition between 11b and 11c. Confirmatory evidence for the hierarchical complexity of LDMA items could be provided by scoring each item via a procedure similar to that demonstrated in Study 2B. From Table 7, it can be seen that 20% of leaders responded at 11a, 36% responded at 11b, 24% responded at 11c, and 5% at 11d. This means that 60% of leaders reasoned at the same level of hierarchical complexity as the LDMA items plus or minus one Lectical Phase, and 85% of leaders reasoned at the same level of hierarchical complexity as LDMA items plus or minus two Lectical Phases. As a result, the majority of leaders provided responses that were somewhat similar in hierarchical complexity to that of the items. It is possible that this occurred because of an interaction between item difficulty and leaders’ reasoning. While it remains an open empirical question as to if, and how, the measurement of cognitive processes may be impacted, Lectica may consider revisiting its rationale for using items with similar difficulties and still achieve alignment of Lectical Phases across the logit scale.

Fourth, while Lectical Phase Scores behave in a manner that satisfies hard stage requirements, the logic underpinning the sequence of Lectical Phases was not explicitly tested. It is plausible that a Rasch analysis would yield similar findings if there were a tendency for a single Lectical Phase to organise responses to LDMA items but if Lectical Phases reflected a cumulative increase in an underlying construct. For example, if the transition from Lectical Phase P to P+1 reflected a cumulative increase in elaboration rather than a hierarchical increase in complexity. In some of her earlier publications, Dawson (2004) presented concept maps in which concepts at successive complexity levels were shown to hierarchically integrate concepts from earlier complexity levels. A similar procedure could be applied at the scale of Lectical Phases by showing that concepts at Lectical Phase P+1 hierarchically integrate concepts at Lectical Phase P. Demonstrating that a
common logic determines the sequence of complexity levels and Lectical Phases may further validate the unidimensionality of the scale.

The findings reported in Study 1A are limited to Lectical Phases 10c through 12b, which means that inferences cannot necessarily be made about Lectical Phases below 10c and above 12b. While this may not present a challenge for leader development, it may have implications for adolescents and/or the general adult population who may perform at, or below, 10c. The findings are also limited to how LAS awards scores to responses about leadership decision-making. Whether similar patterns are observed in other domains of reasoning is yet to be determined. Broader implications, limitations, and future research directions are discussed in Chapter 6.
Study 1B: Ego Development Scores Awarded by the Ego Development Scoring System

The purpose of Study 1B was to determine if ego development scores awarded by the ego development scoring system satisfy hard stage requirements. Appendix A provides a comprehensive list of hard stage claims about ego development theory and measurement.

The Ego Development Approach

As discussed earlier in Chapter 2, the ego development approach considered in this thesis is based on the cumulative work of Loevinger (1976), Torbert (2013), and Cook-Greuter (1999). While there are important distinctions between their perspectives, the following section presents the commonalities that most ego developmentalists would accept as adequately reflective of their approach. Distinctions between approaches are highlighted where relevant. Torbert’s empirical studies have focused on “testing the external validity of the instrument, on learning whether later [stage] leaders were in fact better, as predicted, at helping organizations transform, and on how to support leadership development” (Torbert, 2017, The Appendix, para. 3). Because Torbert’s work is less focused on addressing hard stage requirements, it will be draw on where relevant.

Ego Development Theory

The central tenet of ego development theory is that the ego develops through approximately nine hierarchically integrated stages. The ego is a central personality structure that strives for coherent meaning and organises our perception of reality (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 2013). Contrary to popular belief, this conception of the ego is not borrowed from Freud’s tripartite personality structure in which the ego mediates between the id's instinctual desires and the superego's social expectations (Loevinger, 1976). Instead, the ego is conceived to be a synthetic process analogous to a psychologically oriented central processing unit, which is more
consistent with other theoretical conceptions (Erikson, 1950; Fingarette, 1963; Funk, 1994; Nunberg, 1948). The synthetic process results in constructing a worldview that structures one’s perception and experience of internal and external reality. This worldview typically includes a representation of the self, which takes the form of a solid sense of self-identity (one’s sense of ‘I’ and ‘me’) to mitigate feelings of existential anxiety (Cook-Greuter, 1999). As a theory of development through the lifespan, it accounts for how the ego—or the worldview constructed by the ego—develops through a sequence of stages resulting from the interaction between the self and the environment. Ego development is a multifaceted process that accounts for growth in thought, emotion, motivation, meaning and action (Cook-Greuter, 1999; Loevinger, 1976).

Torbert (2008) refers to stages as action-logics to highlight how they structure one’s logic and action in the world. For consistency, the term stage will be used throughout this thesis. However, the use of this term is not intended to presuppose that stages satisfy hard stage requirements. While different nomenclature is used to label stages in the literature, Torbert’s stage names are primarily used in this thesis due to their relevance to workplace leadership (Rooke & Torbert, 2005). Cook-Greuter’s (2000) names are only used for the highest two stages (i.e., Construct-Aware and Unitive) because of her extensive research on these later stages of development.

Loevinger (1976) has been criticised for not providing an account of the logic underpinning the sequence of stages, even though she espoused that they were hierarchically integrated (Broughton & Zahaykevich, 1988; Cook-Greuter, 1999; Habermas, 1979; Noam, 1993). As a result, her original conception of ego development may be better considered to approximate a descriptive framework rather than an explanatory scientific theory. While Torbert (1994) and Cook-Greuter (1999) provide somewhat different logics, they share an explanatory process which focuses on the acquisition and deepening of perspectives between successive stages (refer to Table 14).

For (1994, 2004), the underpinning logic focuses on the acquisition and coordination of four phenomenological perspectives which arise in our awareness called territories of experience. The
first territory refers to outside events, including results, assessments, observed behavioural consequences and environmental effects. The second territory refers to one’s sensed performance, including behaviour, skills, patterns of activity, and enactment processes. The third territory refers to internal psychological events and includes strategies, schema, ploys, game plans, and modes of reflection. The fourth territory refers to attention and includes a sense of presence, awareness, vision, intuition, and aims (Torbert, 1972, 2004). Torbert contends that transformation between successive stages typically “[involves] successive concentration on developing reliable operational awareness of an additional territory of experience, or on the interplay among several” (Torbert, 1994, p. 183).

In some of his work, (Torbert, 1991a; 1992) describes a second underpinning logic that determines the sequence of his stages. According to this logic, each stage is defined in terms of a state of affairs in which the governing frame of reference of Stage S+1 ‘rules’ the governing frame of reference of Stage S. For example, ‘norms rule needs’ at the Diplomat stage and ‘craft logic rules norms’ at the Expert stage. This logic may illuminate the relations between successive stages, but it has not been conceptually or empirically demonstrated that this logic results in a succession of hierarchical integrations. It is possible that this logic may be more akin to the psychoanalytic process of sublimation, which is a defence mechanism via which the ego subsumes (or replaces) less acceptable psychological processes with more acceptable psychological processes in order to mediate between the superego and the id (Freud, 1989). As a result, this second underpinning logic is acknowledged but it remains unclear whether it results in a set of stages that are organised in an ascending order of hierarchical complexity.

For Cook-Greuter (1999), the underpinning logic focuses on the acquisition and deepening of more cognitively oriented grammatical perspectives called perspectives on the self. According to Cook-Greuter, transformation between successive stages involves the acquisition of a new perspective or the deepening of an existing perspective. For example, growth to the Expert stage involves the acquisition of a third-person perspective and growth to the Achiever stage involves the
expansion of the third-person perspective that was acquired in the Expert stage. Cook-Greuter contends that her logic “[moves] ego development theory from a “soft” stage theory to one more akin to a “hard” stage type according to Kohlberg’s distinction” (Cook-Greuter, 1999, p. 52).

As discussed earlier in Chapter 2, determining whether stages satisfy hard stage requirements must consider whether stages are sequenced according to the logic of hierarchical integration. Given that the theory primarily focuses on the acquisition and deepening of perspectives (i.e., an accumulation of perspectives), it is questionable whether this can account for the espousal that stages are hierarchically integrated.

Compared to several other theories in the field of adult development, ego development has a relatively unique relation to theory because the theory was primarily developed through a combination of inductive and bootstrapping methodologies. “The most distinctive feature of this theory, and of the instrument, is that the first evolved out of the second” (Cook-Greuter, 1999, p. 38). For example, Loevinger’s original conception of ego development stages was informed by the Interpersonal Theory of Psychiatry (Sullivan, 1953) which consists of stages of interpersonal maturity. Loevinger revised the conception of these stages through an iterative bootstrapping process in which observations from the measure were used to develop the descriptions of the stages and the scoring system, and the descriptions of the stages and the scoring system were used to make more precise distinctions in the measure (Loevinger, 1993b; Loevinger & Wessler, 1970). Loevinger clarified that these methods resulted in a conception of stages that was primarily informed by “whatever is measured by the SCT” (Loevinger, 1983, p. 346). Torbert’s version of ego development theory followed in a similar vein because his theoretical conception of ego development stages seems to be primarily based on biographies of historical characters and qualitative accounts of senior leaders (Torbert, 1987c, 2004). Furthermore, the territories of experience that he uses to account for relations between ego development stages seem to have been derived from first-person phenomenological experiences (Torbert, 1972, 2004). Over time, he fleshed out a more fully formed theoretical conception as reflected in Collaborative Developmental
Action Inquiry (CDAI) Theory (Torbert, 2013). Cook-Greuter (1994, 1999) defined the attributes of her two highest stages of ego development through a combination of abductive and inductive approaches. The abductive approach consisted of identifying essential features of advanced stages in specific developmental models and using these features to iterate her conception of the highest stages of development. The inductive approach consisted of identifying late-stage individuals whose development could not be measured through earlier ego development scoring systems and then using the themes that emerged from these individuals to develop scoring criteria for the highest two stages. It may be reasonable to conclude that ego development theory and measurement are so tightly coupled that an evaluation of measurement properties may have significant implications for the central tenets of the theory.

It is not in scope to conduct a detailed review of all critiques levelled against ego development theory because this thesis is explicitly focused on the hierarchically integrated nature of the stages. In general terms, some critiques have focused on the ambiguity of the construct being measured (Noam, 1993), the manner in which the ego performs its synthetic function (Thorne, 1993), the potential conflation between structure and content in the definitional attributes of stages (Noam, 1993), and the relationship between ego development stages and the big five personality factors (Costa & McCrae, 1993).

**The Ego Development Scale**

The ego development scale is contended to be a unidimensional ordinal scale consisting of nine hierarchically integrated stages which satisfy hard stage requirements (refer to Appendix A). Because stages are considered to be hierarchically integrated, an increase in hierarchical complexity is typically considered necessary but not sufficient to account for development from one stage to the next (e.g., Cook-Greuter, 1999, pp. 90-93). While various scholars have outlined the definitional attributes of stages, some variability between scholars is apparent, particularly at the highest stages (Cook-Greuter, 1990, 1999, 2002, 2013; Loevinger, 1976; Rooke & Torbert, 2005; Torbert, 2004).
Cook-Greuter argued that her conception converges with Loevinger’s up to the Strategist stage (Cook-Greuter, 1999) and is similar to Torbert’s conception up to his highest stages (Cook-Greuter, 2013).

The scale is contended to account for ego development from birth to the most advanced stages of development which are currently measurable via conventional approaches to metrology. Some of the essential properties of the nine stages, from Impulsive to Unitive, are summarised in Table 14. In general terms, however, the scale describes an ego that develops through the preconventional stages of Impulsive and Opportunistic in which social mores are yet to be internalised; through to the conventional stages of Diplomat, Expert, and Achiever in which social conventions are increasingly integrated and mastered; through to the postconventional stages of Individualist, Strategist, and Construct-Aware in which the ego becomes increasingly aware of, and subsequently transcends social conventions; through to the Unitive stage in which the ego is believed to transcend arbitrary social constructs including itself. The scale has been described as accounting for the “acquisition of conventional liberal individualism” until the Strategist stage (Broughton & Zahaykevich, 1988, p. 195), and then the stepwise deconstruction of arbitrary social constructs—including the construct of the self—in favour of direct phenomenological experience of reality in the Construct-Aware and Unitive stages (Cook-Greuter, 1999, 2000). Recent research on smaller samples suggests that the majority of workplace leaders tend to score at approximately the Expert stage of development (Hagström & Backström, 2017).
<table>
<thead>
<tr>
<th>No.</th>
<th>Stage</th>
<th>Relative % in adult population</th>
<th>Cook-Greuter’s underpinning logic</th>
<th>Torbert’s underpinning logic</th>
<th>Definitional attributes of ego development stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Impulsive</td>
<td>0.4</td>
<td>1st person perspective</td>
<td></td>
<td>Impulses govern actions, affirmation of separate identity</td>
</tr>
<tr>
<td>2</td>
<td>Opportunist</td>
<td>4.6</td>
<td>Expanded 1st person perspective</td>
<td>1st territory as primary reality</td>
<td>Some self-control, anticipation of short-term rewards, attack oriented, win at any expense, self-oriented, manipulative and externalizes blame, hostile humour, rejects feedback</td>
</tr>
<tr>
<td>3</td>
<td>Diplomat</td>
<td>5.5</td>
<td>2nd person perspective &amp; expanded 2nd person perspective</td>
<td>2nd territory as primary reality</td>
<td>Socially expected behaviour, approval oriented, identify with needs of the group, observes rules, avoids overt conflict, wants to belong</td>
</tr>
<tr>
<td>4</td>
<td>Expert</td>
<td>29.7</td>
<td>3rd person perspective</td>
<td>3rd territory as primary reality</td>
<td>Relies on logic and expertise, focus on procedure and efficiency over effectiveness, self-reflection is possible, develops multiple solutions to problems, perfectionistic</td>
</tr>
<tr>
<td>5</td>
<td>Achiever</td>
<td>33.5</td>
<td>Expanded 3rd person perspective</td>
<td>Single-loop feedback on 2nd territory</td>
<td>Focus on delivery of results, effectiveness, long-term goals, juggles multiple responsibilities, self-criticism, can reflect on past and future and step into others’ shoes</td>
</tr>
<tr>
<td>6</td>
<td>Individualist</td>
<td>15.3</td>
<td>4th person perspective</td>
<td>Relativism trumps any territory</td>
<td>Perceives self in relationship to system, relativistic thinking, awareness of contexts, decisions based on own view of reality, values independence, creates unique solutions, increased tolerance of self and other</td>
</tr>
<tr>
<td>7</td>
<td>Strategist</td>
<td>8.0</td>
<td>Expanded 4th person perspective</td>
<td>Double-loop feedback on 3rd territory</td>
<td>Linking theory and principles with practice, perceives interaction in dynamic systems, address conflicting needs and duties, strategic perspective, engages in complex relationships, values self-fulfilment, generates organisational transformation</td>
</tr>
<tr>
<td>8</td>
<td>Construct-Aware</td>
<td>2.3</td>
<td>5th person perspective</td>
<td>Triple-loop feedback</td>
<td>Interplay of awareness, thought, action, and effects; transforming</td>
</tr>
<tr>
<td>No.</td>
<td>Stage</td>
<td>Relative % in adult population</td>
<td>Cook-Greuter’s underpinning logic</td>
<td>Torbert’s underpinning logic</td>
<td>Definitional attributes of ego development stages</td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
<td>--------------------------------</td>
<td>-----------------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>9</td>
<td>Unitive</td>
<td>0.6 Cosmic perspective</td>
<td>Coordinates single, double, and triple loop feedback</td>
<td>on 4th territory</td>
<td>self and others, awareness of the ego that constructs an identity, communicates metaphorically, generates social transformation</td>
</tr>
</tbody>
</table>

Notes.  

* Relative percentage in the adult population is based on a large sample (N=6895) presented by Cook-Greuter (2010).  

* Cook-Greuter’s underpinning logic is based on descriptions provided by Cook-Greuter (1990, 1999, 2013).  

* Torbert’s underpinning logic is based on descriptions provided by Rooke and Torbert (1999); Torbert (1994, 2004, 2013).  

* Definitional attributes of stages are based on definitions provided by Cook-Greuter (2004); Cook-Greuter and Soulen (2007); Ingersoll and Cook-Greuter (2007); Rooke and Torbert (2005).

**Measuring Ego Development**

Ego development is typically measured via a semi-projective test called the Sentence Completion Test (SCT), which is typically administered in paper-and-pencil or electronic format. The SCT is administered to locate the development of test-takers’ egos along the scale of ego development which ranges from Impulsive to Unitive. The SCT is claimed to be “the most extensively validated projective technique” (Lilienfeld et al., 2000, p. 56), which may account for why it is one of the most frequently used developmental assessments employed in the adult development literature. Test-takers are provided minimal instructions, such as “Complete the following sentences”, to avoid influencing their responses (Hy & Loevinger, 1996, p. 26). The SCT usually consists of 18 to 36 incomplete SCT sentence stems (or ‘SCT stems’ or simply ‘stems’ for short), which test-takers are invited to complete. When test-takers have responded to the stems, the completed assessment is referred to as an SCT protocol, or simply protocol for short. Examples
of stems include “When a child will not join in group activities ...”, “What gets me into trouble is ...”, “Rules are ...”. Test-takers’ responses are assumed to be a manifestation of their primary stage, which they are believed to project onto each stem. This assumption is premised on the notion that test-takers’ use of language reflects their underlying worldview (Cook-Greuter, 1999). Different versions of the SCT are available for different assessment contexts, several of which are used to inform analyses undertaken in this thesis. The term ‘SCT’ is used to refer to this general class of semi-projective assessments. Since the SCT measures development rather than speed, time limits are typically not imposed.

**The Ego Development Scoring System**

Test-takers’ responses to stems are scored via the ego development scoring system (Hy & Loevinger, 1996). Ego developmentalists often refer to the ‘scoring manual’ rather than the ‘scoring system’. However, the term scoring system is explicitly used by Hy and Loevinger (1996) and Cook-Greuter (1994) because the manual consists of a system of considerations and rules about scoring ego development. For this reason, and also to ensure terminological similarity to the discussion on cognitive development, the term scoring system will be used throughout this thesis. There are multiple versions of the ego development scoring system currently in circulation. The purpose of the scoring system is to determine the primary stage from which stems were completed. The scoring system is domain-specific because it can only assess ego development stages but no other dimensions of development, such as leadership decision-making. Notwithstanding this limitation, some ego developmentalists use ego development scores to make inferences about dimensions of development that may not be strictly related to the ego (Cook-Greuter, 2013). Scoring is undertaken by matching test-takers’ responses to examples and exemplars in the scoring system and also by applying a series of scoring rules. For example, when scoring the stem “When a child will not join in group activities ...”, the response “check for possible signs of illness” is scored at the Diplomat stage while the response “he’s probably marching to a different drummer” is scored at the Individualist
stage (Hy & Loevinger, 1996). Scoring rules enable scorers to deal with situations in which responses fit a single stage, lower-level responses combine to form responses at higher stages, and in which responses do not fit any particular category. For example, scoring rule 2 specifies that “where the combination of two or more elements in a compound response generates a more complex level of conception, rate the response one step higher than the highest elements” (Hy & Loevinger, 1996, p. 34). This types of scoring rule may lend support to the notion that ego development scores are awarded on the basis of considerations that may be related to the axioms of hierarchical complexity (Commons & Pekker, 2008). The distribution of scores across all stems is compared to a series of cut-scores to provide a total score for the whole protocol, called a Total Protocol Rating (TPR) (Hy & Loevinger, 1996).

Whereas the scale is claimed to reflect a unidimensional construct i.e., ego development, the scoring system explicitly acknowledges the multifaceted ways in which ego development manifests in test-takers’ responses. Loevinger’s original conception considered ego development to manifest in impulse control, interpersonal style, conscious preoccupations, and cognitive style (Loevinger, 1976; Loevinger & Wessler, 1970). Somewhat later in the literature, Cook-Greuter (1990, p. 100) elaborated on ten indicators of ego development which included “evolving perspectives on the self; semantic distinctions in vocabulary such as vague/differentiated, physical/psychological, static/dynamic; number of elements combined in a response; tone of voice; themes and preoccupations; qualifiers, conditions, and contingencies; contrast; time frame; internality; and cognition as evidenced by linear, systemic, or unitary concepts”. More recently, (Cook-Greuter, 2002, 2013) enumerated additional manifestations of ego development, including feelings, cognitive style, conscious preoccupations, internal dimensions, problem-solving, decision-making style, organisational type, defence mechanisms, counselling style, anxiety, depression, character, language clues, impulse control, interpersonal style, and conceptions of truth. It remains to be fully explored whether each of these indicators develops hierarchically, if trained scorers consistently consider this multitude of manifestations, and how this may impact the unidimensionality of the scale.
**Ego Development Scores (Including SCT Stem Scores and TPRs)**

The ego development scoring system awards scores at two scales: *SCT stem scores* awarded to individual stems and *TPRs* awarded to whole protocols. Total Weighted Scores (TWS) are sometimes utilised in the ego development literature, but since they were not consistently calculated for the samples analysed in this thesis, they were not considered further. The term *SCT score* is a general term used to refer to both SCT stem scores and TPRs in the context of this thesis. A TPR reflects the primary stage from which test-takers responded to the SCT. This score is used to make inferences about the ‘centre of gravity’ of test-takers’ worldviews. This tendency to summarise the development of the human ego or worldview via a single score may be regarded as somewhat reductionistic in some contexts. The centre of gravity is typically conceived to be “that level of ego maturity to which a person has consistent access under ordinary circumstances without special support conditions or under unusual stress. In other words, the measured stage is the one that a person reliably and routinely uses in response to everyday life” (Cook-Greuter, 2013, pp. 7-8). It is estimated that it takes approximately five years to develop to a new stage as a function of life circumstances or approximately one year of deliberate practice through a well-designed development program (Cook-Greuter, 2013; Manners et al., 2004). Formative scores inform the design of developmental curricula and summative scores are used to evaluate growth over time. Because all scores are calibrated to the same underlying ordinal scale, two or more scores along the scale can be directly compared e.g., test-times 1 and 2. In this thesis, ego development scores are operationalised by SCT scores, which are used as the basis for analyses.

**Evidence that Ego Development Scores Satisfy Hard Stage Requirements**

Ego developmentalists consistently espouse the hard stage properties of their stages as outlined in Appendix A, but their evidence is equivocal. While the evidence suggests that stages are invariantly sequenced, clear evidence in support of unidimensionality, qualitative distinctness, structured wholeness, and hierarchical integrations is more challenging to identify. A considerable
proportion of the ego development literature has focused on establishing various forms of reliability and validity (e.g., criterion, discriminant, predictive, and external) rather than explicitly testing hard stage requirements (Blasi, 1993; Broughton & Zahaykevich, 1988; Hauser, 1976, 1993; Loevinger, 1979, 1993b; Manners & Durkin, 2001). As a result, the literature review below draws upon a broad range of research to construct an argument on behalf of ego developmentalists. Based on a literature review of over 500 publications, Cook-Greuter’s (1999) dissertation is one of the few attempts to directly demonstrate that SCT scores satisfy hard stage requirements. As a result, her data set is reanalysed using more contemporary methods in Study 2, described in Chapter 3.

**Unidimensionality**

The evidence on the dimensionality of SCT scores converges on two main findings. First, scores reflect a broad unidimensional construct, i.e., ego development. Second, this construct only accounts for a minor amount of variability in scores (approximately 20%). This suggests that most of the variability (approximately 80%) is attributable to unidentified secondary constructs and/or noise (Blasi, 1971; Kishton et al., 1984; Lambert, 1972a, 1972b; Loevinger & Wessler, 1970; Lorr & Manning, 1978). This may indicate compromised internal consistency and stands in contrast to other scores based on hard stage models for which the first principal component has been shown to account for 92 - 99% of variability in various samples (Commons & Chen, 2014). The underlying factor structure of SCT scores is an important point to which this literature review will return shortly. If most of the variability in SCT scores is attributable to unidentified secondary constructs and/or noise, then this may challenge the conclusions that have been drawn via correlational studies that seek to establish that a total score on the SCT converges with, or predicts, other leadership phenomena.

Loevinger and Wessler (1970) undertook seminal research on the SCT, which is consistently referred to in the literature to justify its dimensionality. A PCA was performed on a sample of 543 women and girls who had undertaken the 36-stem version of the SCT. The PCA demonstrated that
the first component had an eigenvalue of 8.8 and accounted for approximately 20% of variability, while the second component had an eigenvalue of 1.2 and accounted for 5.6% of variability. Loevinger predicted that stems which loaded on secondary factors would be unrelated to the target construct and could be removed from the SCT. Secondary factors were uninterpretable, and most, if not all, of the original stems, were retained. She concluded that the SCT was tapping a single underlying construct and that ego development was an indivisible structural unity.

To test Loevinger’s conclusion, earlier researchers hypothesised that some stems measured subconstructs of ego development. For example, Blasi (1971) hypothesised that some stems tapped a subconstruct of responsibility, and Lambert (1972a, 1972b) hypothesised that some stems tapped a subconstruct of moral responsibility. Neither found sufficient evidence to support the contention that additional subconstructs could be identified beyond the primary construct of ego development.

Several other studies support the general direction of these findings. Lorr and Manning (1978) conducted a PCA on a sample of 648 SCTs, revealing only one interpretable construct that accounted for 20% of variability for girls and 19% for boys. Kishton, Starrett and Lucas (1984) examined the dimensionality of SCT for 89 early adolescents and 82 late adolescents. A parcel analysis was undertaken by grouping stems into parcels and then entering the parcels into a factor analysis to increase the stability of the estimates. For each adolescent, the 36 stems were scored individually and were grouped into three parcels e.g., stems 1, 4, 7 were allocated to parcel one and stems 2, 5, 8 were allocated to parcel two, and so on. For the early adolescents, correlations between parcels ranged from .82 to .84. However, for the late adolescents, correlations between parcels ranged from .55 to .57. This suggested that, while scores may reflect a unidimensional construct early in life, secondary constructs play a more substantial role in later life. When parcels were entered into a PCA, results demonstrated that the first factor loaded on ego development for early adolescents. However, the first factor loaded on ego development and word count for late adolescents. This suggests that word count may play an increasingly important role or become increasingly conflated with ego development for older test-takers.
One of the most sophisticated analyses of the dimensionality of the SCT was undertaken by Novy et al. (1994). The researchers administered a battery of assessments to 267 adults, including the 36-stem version of the SCT and various surveys conceptually related to ego development theory. This included measures of impulse control, interpersonal style, conscious preoccupations, and cognitive style. Structural equation modelling (SEM) was used to determine how ego development was related to the other four factors. The most tenable model was one in which ego development covaried with the four subconstructs, and another higher-order construct subsumed all five constructs. This higher-order construct was strongly related to interpersonal style ($r=.91$) and moderately related to conscious preoccupations and ego development ($r=.48$). These findings suggest that even if ego development emerges as a unidimensional construct, it may not play a superordinate role in personality organisation, as ego developmentalists typically suggest.

The most recent study on dimensionality was undertaken on the Harthill Leadership Development Profile, a version of the SCT for workplace and leadership applications (Torbert & Livne-Tarandach, 2009). The authors describe their analytical techniques as based on cluster analysis in one publication (Torbert & Livne-Tarandach, 2009) and factor analysis in another (Torbert et al., 2010), despite the important distinctions between these procedures. The researchers explain their findings by appealing to correlations between SCT stem scores, so it is assumed that an Exploratory Factor Analysis (EFA) was undertaken because EFA generates factors via intercorrelations between item scores. Two samples were analysed: 830 SCTs at Achiever or earlier and 61 SCTs at Individualist or later. Eight distinct constructs were identified for the less developed sample and stems displayed minimal cross-loadings. Eleven constructs were identified for the more developed sample and stems displayed significant cross-loadings. The researchers suggest that simple mental maps underpinned less developed SCTs, but more systems-oriented mental maps underpinned the more developed SCTs.

Several limitations are noted because the researchers provide no interpretation of the constructs identified, no description of the amount of variability accounted for by each construct or
the general amount of variability accounted for in their dataset. Their interpretation also appears to be inconsistent with ego development theory. For the SCT to be unidimensional, one would expect a single factor to be identified. From a statistical perspective, these findings suggest the presence of secondary constructs and/or noise—particularly in the more developed protocols, which may challenge the scale’s unidimensionality. Alternatively, their findings may be more consistent with an assessment that measures multiple constructs and awards separate scores to each construct. Similarly to the literature discussed earlier, this study appears to highlight the presence of multidimensionality, particularly in samples with higher scores.

**Invariant Sequence**

The evidence for invariance sequence is relatively sound and is supported by cross-sectional, longitudinal, interventional, and cross-cultural studies.

**Cross-sectional Evidence.** The cross-sectional evidence demonstrates that TPRs are awarded in the theoretically predicted sequence as a curvilinear function of age. The most significant study on the relation between TPRs and age was Cohn’s (1998) meta-analysis of 30 studies involving 4,330 test-takers from childhood through to late adulthood. For children and adolescents aged 11-20, the weighted estimate of the population correlation between age and TPR was .40. For adults aged 20-65+, the estimate was .04. For the cross-generational sample, the estimate was .27. As the correlations for all three cohorts were significant and positive, this provided general support that scores are given in the theoretically predicted sequence in increasingly older test-takers. Additional cross-sectional studies not included in Cohn’s meta-analyses provide general support for higher scores in older cohorts (Foster & Sprinthall, 1992; Kishton et al., 1984; Loevinger & Wessler, 1970).

Davison et al. (1980) reasoned that if stages occur in a theoretically predicted sequence and if a protocol has a modal score of \( S \), then adjacent scores \( (S+1) \) should be the second most frequent. In a sample of 171 high school and university students, 72-81% of the sample was consistent with
this prediction. While the researchers claimed their evidence was supportive of invariance sequence, 19-28% of protocols violated this pattern which may be considered greater than the percentage expected by random chance.

Lanning et al. (2018) analysed a large sample of protocols based on computational linguistics. For the total sample of 10,670 words and a curated sample of 1,811 words, there was a strong relation in word use between adjacent stages (S, S+1) but a weaker relation in word use between nonadjacent stages. This pattern was observed in most pairwise comparisons and used to support the requirement of invariant sequence.

**Longitudinal Evidence.** The longitudinal evidence also demonstrates that TPRs unfold in the theoretically predicted sequence as a curvilinear function of age. Cohn’s (1998) meta-analysis also included 16 longitudinal studies involving 39 samples and 1862 test-takers. Cohn found a significant increase in TPRs over one to nine years, and the size of this correlation varied as a function of age. There was a significant increase in scores for 11 of the 16 samples for children and adolescents. The estimated population correlation between age and stage was .41. However, for the college age group, there was only a significant increase in scores for six of the 21 samples and an estimated population correlation between age and stage of .13. The authors concluded that, while longitudinal growth was present, it was less pronounced later in life. Additional longitudinal studies published after Cohn’s meta-analyses support growth over time in the theoretically predicted sequence for adolescents, college students, and adults over time intervals of two to 18 years (Allen et al., 1994; Bauer & McAdams, 2010; Hennighausen et al., 2004; Lilgendahl et al., 2013; Syed & Seiffge-Krenke, 2013, 2015; Westenberg & Gjerde, 1999). More recently, Anagnostakis (2022) demonstrated longitudinal growth in the theoretically predicted sequence over the course of a 5-month leader development program.

While growth is apparent when evaluating the average change in groups, growth trajectories are more variable when considered at the scale of individuals. During adolescence and early adulthood, patterns of growth, arrest, and even regression are often observed in individuals (Adams
Redmore (1983) found an overall increase of 0.5 in SCT scores in a sample of college students tested over four years, even though 41% increased, 49% arrested, and 10% regressed. Westenberg and Gjerde (1999) tracked ego development over nine years, from age 14 to 23. They found that while there was an overall increase of 1.5 stages in the group, there were considerable differences in the timing and extent of development for individuals. Of the participants who entered the study at the Opportunist or Diplomat stages, 10% arrested, 23-27% developed by one stage and 60-68% developed by two or more stages. Of the participants who entered the study at the Expert stage, 29% developed by one stage, 58% arrested, and only 13% developed by two stages.

**Interventional Evidence.** The interventional evidence has demonstrated that when people are exposed to optimal developmental conditions, SCT scores increase in the theoretically predicted sequence. While the longitudinal evidence discussed above suggests that children and young adults may have grown in the absence of intervention, higher SCT scores have been fostered in adult populations that may not have occurred as a function of life conditions alone.

In younger populations, significant growth in TPRs has been identified in high school students during attendance at semester-long counselling classes (Moscher et al., 1971), in MBA students at an action-research program (Torbert, 1994), among young adults who engaged in practices based on autobiographical awareness (Torbert & Fisher, 1992) and also in students who engaged in specific contemplative practices (Chandler et al., 2005). Many of these studies found a higher growth rate relative to control groups. However, causal inferences are challenging because ego development is likely to occur in adolescence and early adulthood without additional support (Cohn, 1998).

Interventional studies on adult populations are more persuasive as they suggest that growth can be deliberately catalysed during a phase of life that may not have otherwise occurred. Significant growth in TPRs has been identified in various adult samples over various interventional
durations. Growth has been measured in teachers after an 11-week empathy training program (Hurt, 1977), female nurses after two years of training and experience (White, 1985), prison inmates after a seven-month moral education program (MacPhail, 1989) and after one year of meditative practice (Alexander et al., 1990). Growth has also been measured in adults after a 10-week program (Manners et al., 2004), a 30-week program based on deliberate psychological education (Cannon & Frank, 2009), a 10-month community leadership program (Vincent et al., 2015), a 5-month executive leadership program (Anagnostakis, 2022), an 18 month intensive training program in the Enneagram personality typologies (Daniels et al., 2018), and following exposure to multiple perspectives on the attribution of meaning to personal experiences (Pillay & April, 2022). Although direct causal attributions cannot be made because of methodological limitations (e.g., limited control conditions), at least some of these interventions may be associated with growth in TPRs in the theoretically predicted direction. Notwithstanding these positive results, some studies have failed to identify significant growth even after attendance at targeted development programs (Lambie & Ieva, 2012; Manners et al., 2004; Takei et al., 2000). A variety of other unpublished interventional studies on ego development are reviewed by Brownlow (2022).

**Cross-cultural Evidence.** The cross-cultural evidence provides general support for the requirement of invariant sequence. For example, exploratory research has supported the applicability of the SCT to spiritual subcultures in Western countries (Haan et al., 1973; Rosén & Nordquist, 1980), immigrants to the United States (Arredondo, 1984; Arredondo-Dowd, 1981), various ethnic groups suffering from substance abuse (Wurzman et al., 1982), black identities in US colleges (Dunston & Roberts, 1987; Looney, 1988), and comparative analyses between Western and Asian cultures (Ravinder, 1986).

The applied research has provided tentative support for the relevance of non-English translations of the SCT in Germany (Kapfhammer et al., 1993), Israel (Snarey & Blasi, 1980; Zlotogorski, 1983), Japan (Kusatsu, 1977, 1978; Sasaki, 1980, 1981a, 1981b; Tochio & Akiba, 1988; Tochio & Hanada, 1991; Watanabe & Yamamoto, 1989), the Netherlands (Westenberg et al., 1998),

**Structure of Whole**

The evidence for structured wholeness is weakly supported in the ego development literature. According to theory, each stage is contended to be an internally consistent system of meaning-making. Ego developmentalists operationalise the requirement of structured wholeness by using TPRs to make inferences about the ego’s centre of gravity. The centre of gravity is the stage at which the ego is located for several years in normal conditions without support or stress (Cook-Greuter, 2013; Loevinger, 1976; Palus et al., 2020). Nonetheless, ego developmentalists simultaneously recognise that SCT stem scores are broadly distributed across the ego development scale and infer that people do not consistently function at a single stage in all contexts (Cook-Greuter, 1999, 2013; Davison et al., 1980; Loevinger, 1993a; Torbert, 2004). This raises an implicit contradiction between ego development theory and measurement. How can the centre of gravity be a sufficiently powerful attractor to organise the entirety of one’s worldview and yet be insufficiently powerful to organise responses to 18–36 stems? As Noam (1993, p. 46) argues, “considerable stage scatter is not infrequent on Loevinger tests, but ends up disappearing in the final stage assignment”. Inferences about the centre of gravity appear somewhat tenuous in the context of Loevinger’s admission that “An implicit assumption underlies our technique: every person, although exhibiting variability in his responses and other behavior, belongs in principle to one and only one point on the [ego development] scale. This is a working assumption, not subject to any kind of test except the long range one of the fruitfulness of the whole technique” (Loevinger & Wessler, 1970, p. 22). Little of the post-Loevinger literature has attempted to directly address the requirement of structured wholeness. However, three potential explanations can be inferred from the literature: measurement error, décalage (Piaget, 1952), and growth patterns over time.
The first explanation is found in Loevinger’s work, where she asserted that “Probably variability is more a function of the instrument than of the person” (Loevinger, 1976, pp. 239-240). While the ego may be stably located at a single stage, the SCT lacks the measurement precision to identify its centre of gravity. Given Loevinger’s professional background as a psychometrician and the number of studies that have attempted to validate the measurement properties of the SCT, this explanation does not seem to be tenable.

The second explanation is décalage which is a term coined by Piaget (1952, 1985) to describe the tendency for people to perform at different stages in different domains. The broad distribution of scores in protocols is often attributed to décalage (Cook-Greuter, 1999; Holt, 1998; Loevinger, 1993a, 1998c; Noam, 1993; Novy, 1993). Although this explanation may be conceptually appealing, it does not explain why a single TPR is contended to reflect the worldview. If décalage were responsible for the broad distribution of scores, it would seem more reasonable to provide test-takers with a developmental range or separate scores to reflect different magnitudes of development in different facets of life.

The third explanation is attributed to growth over time. As explained by Cook-Greuter (1999, p. 36), “Reasoning from earlier stages, because it is part of one’s own past experience, is to some degree still available. Upward scatter, on the other hand, is interpreted as intimations of more complex ways of meaning making in an SCT, although not sufficiently numerous to show that a person has shifted into the adjacent higher level of meaning making”. Although this explanation is somewhat plausible, it is hard to explain why “Often responses span over four, five and rarely over the whole range of possibilities” (Cook-Greuter, 2013, p. 8). This explanation is inconsistent with the requirement of structured wholeness and stands in contrast to the findings on cognitive-developmental scores reviewed in Study 1A. Hierarchical integrations between successive stages typically result in reasoning from previous stages being consolidated at the current modal stage and negligible access to higher stages. In sum, there are only relatively weak reasons to accept that SCT scores satisfy the requirement of structured wholeness.
Qualitatively Distinct

Claims about the requirement of qualitative distinctness are made consistently throughout the literature (Cook-Greuter, 2013; Loevinger, 1976; Torbert, 2010), but the evidence does not clearly support this conclusion. Qualitative distinctness appears tenuous in the context of Loevinger’s (1993b, p. 11) reflections later in her career, “Are there discrete stages of development? Or are there designated stages arbitrary points along a continuum? ... I began by leaning toward distinct stages but have increasingly recognized that there is no evidence for that premise and some evidence for the contrary premise of gradual transitions.” The four main arguments used to establish the requirement of qualitative distinctness relate to thematic differences, SCT distributions, factors structures, and distinct strata of performance.

The ego development scoring system is claimed to distinguish between qualitatively distinct stages. For example, the opportunistic hedonism of the Opportunist stage (“I just can’t stand people who ... bug you to death to finish their damn survey”) is relatively distinct from the rule-boundedness of the Diplomat stage (“A man should always ... be on his best behaviour”) (Hy & Loevinger, 1996, pp. 10-12). This argument is supported by a qualitative analysis of a sample of 44,000 stem completions and the 250 most characteristic terms used at each stage (Lanning et al., 2018). Findings indicated that SCT stem scores were associated with a distinct conceptual focus. For example, the stems scored at the Individualist stage were associated with intellectual growth, whereas those scored at the Strategist stage were associated with renegotiating life goals. Although this suggests thematic distinctions between SCT stem scores, it does not establish that distinctions emerge because of qualitative transformations between successive stages.

Distributions scored at successive TPRs occupy different locations along the ego development scale. Thus, some scholars argue that the ego development scoring system makes qualitative distinctions between stages (Cook-Greuter, 1999; Loevinger & Wessler, 1970). However, this pattern would be expected for all three types of developmental theories—cultural ages, soft stages, and hard stages. Many Likert scales organised according to a cumulative increase in an
underlying construct have been shown to exhibit a similar pattern (Bond & Fox, 2015; Hagquist et al., 2009; Klein et al., 2022; Pallant & Tennant, 2007; Steele & Day, 2020). Although this argument suggests that SCT scores reflect some type of distinction, it does not establish that distinctions emerge because of qualitative transformations.

As discussed in the preceding section on unidimensionality, Torbert and Livne-Tarandach’s (2009) analysis demonstrated that eight distinct constructs were identified for protocols with lower TPRs, and 11 interdependent constructs for protocols with higher TPRs. This has been used to argue that stages reflect qualitatively distinct structures (Torbert & Livne-Tarandach, 2009). However, this explanation is not consistent with standard interpretations of factor analyses. For SCT scores to satisfy the requirement of unidimensionality, there should be a single underlying construct identified. A mere increase in the number of constructs may constitute a quantitative, not a qualitative, increase in constructs between protocols with different TPRs. Stems that cross-load on multiple constructs for protocols with higher TPRs may indicate poorly defined stems, secondary constructs, and/or noise in the data. This cannot be used as evidence to support qualitative distinctions between successive stages.

Reliability indices estimate the precision attributable to a set of scores, and can be used to identify the number of statistically distinct strata of performance which can be distinguished by a scoring system (Fisher, 1992; Wright, 1996; Wright & Masters, 2002). In the context of hard stage requirements, the number of statistically distinct strata is conceptually equivalent to the number of statistically distinct development stages that a scoring system can distinguish. The higher the reliability, the higher the precision of the estimates, the lower the amount of measurement error, and the more developmental stages that can be distinguished. In order to claim that the ego development scoring system can distinguish between nine qualitatively distinct stages, the scoring system would need to award scores with a reliability of .98. A multitude of studies have examined the reliability of the SCT and typically reported estimates in the range .70 to .91 depending on the number of stems included in the analysis (refer to Appendix E). Reliabilities for the 18-stem version
of the SCT ranged from .70 to .91 with a median of .82. However, assessments with a reliability of .82 can only measure approximately three distinct stages. Reliabilities for the 36-stem version of SCT ranged from .76 to .91 with a median of .88. However, assessments with a reliability of .88 can only measure approximately four distinct stages. As a result, there is an inconsistency between the nine stages that the scoring system claims to distinguish between, and the three or four stages that can be distinguished.

Hierarchical Integrations

Claims about the hierarchically integrated nature of stages are consistently made throughout the literature, as presented in Appendix A. However, these claims are only weakly supported by the evidence. This may have led some ego developmentalists to speculate that their stages may not be hierarchically integrated, despite earlier claims about them (Blasi, 1998; Loevinger, 1986).

Cook-Greuter (1999, p. 52) claimed to have resolved these questions by asserting that her perspectives on the self “[moved] ego development theory from a “soft” stage theory to one more akin to a “hard” stage type according to Kohlberg’s distinction”. The three main arguments to support this conclusion relate to the asymmetry of comprehension, correspondence to hard stages, and convergent correlations.

The first argument, asymmetry of comprehension, is that people can only minimally comprehend stages beyond their centre of gravity because they reflect hierarchical integrations beyond their current worldview. In a series of experiments, Redmore (1976) demonstrated that university students and staff could reduce but not increase their TPRs by more than half a stage. Participants were administered the SCT at test-time 1 under standard conditions and provided alternative administration instructions at time-time 2. Findings indicated that TPRs typically decreased when test-takers were asked to fake a lower stage or fake a bad impression, and scores either stayed the same or only increased half a stage when they were asked to fake a good
impression. Jurich and Holt (1987) administered the SCT to undergraduate psychology students in a modified design. The control group was provided with standard SCT instructions. In contrast, the experimental group was provided with modified instructions (“This is a test of personal maturity: Please complete each sentence in as an adult and mature a manner as you can”). Findings suggested that the experimental group only scored one-half a stage higher than the control group.

In a more carefully controlled study, Blumentritt et al. (1996) administered the SCT to 90 adults with standard instructions. One week later, the SCT was readministered to participants who were divided into three groups: the control group (standard instructions), a role-play group (“Please complete the following sentence stems as if you were a [Construct-Aware] person”); and a best effort group (“Please complete the following sentence stems in the most complex, thought-provoking way that you can. Give your best responses). At test-time 2, the role-play and best-effort groups only increased their scores by half a stage.

Although these studies suggest that test-takers struggle to fake an increase to their TPRs, they do not demonstrate that this phenomenon occurs because stages are hierarchically integrated. It is conceptually plausible for similar phenomena to occur for cultural ages in which people may find it hard to grasp a later phase of life that they have not yet experienced; or soft stages, in which people find it hard to display a cumulative increase in an underlying construct which may exceed their current level of capability (e.g., memorising a larger number of items than what they are accustomed to). These alternative explanations have not been discussed or ruled out in the ego development literature. Several studies have demonstrated that test-takers can ‘fake bad’ more successfully than they can ‘fake good’ on a variety of psychological instruments such Conditional Reasoning Tests (Schoen et al., 2022) and emotional intelligence surveys (Whitman et al., 2008). These findings have not led scholars to conclude that the constructs measured by these tests are hierarchically integrated because test-takers struggle to fake an increase in their performance.

The second argument in support of hierarchical integrations is that there is a structural isomorphism between ego development and cognitive-developmental stages (Cook-Greuter, 1990,
Loevinger’s original work in which she mapped ego development stages to a wide variety of other stage models—see for example Loevinger (1976, pp. 68-135). This general line of argumentation was perpetuated by ego developmentalists who elaborated on Loevinger’s work. If there were a direct correspondence, we would expect to find a high (or very high) correlation between SCT and cognitive-developmental scores in a similar range to that found between scores based on other hard stage models. As discussed in Study 1A, Dawson’s studies (2001, 2002a, 2003b; 2003) consistently found correlations between cognitive-developmental scores between .86 to .92, with a median of .90, and a mean of .90. These results suggest that scores were tapping the same unidimensional construct, i.e., hierarchical complexity. Approximately 31 studies have examined the correlation between SCT scores and cognitive-developmental scores. Lee and Snarey (1988) meta-analysed nine studies and 525 cases to examine the relation between TPRs and scores on the moral judgement interview. Correlations between SCT scores and moral judgement scores ranged between .29 to .43, depending on the scoring procedures applied. These correlations appear to fall considerably outside the range of .86 to .92 and do not support the contention that there is a direct correspondence between the two sets of stages.

Cohn and Westenberg (2004) included the results from five studies and 383 test-takers in a second meta-analysis. The unadjusted correlation between TPRs and scores on Piagetian assessments was .34, with a 95% confidence interval ranging from .25 to .43. However, the adjusted correlation between scores was .44, with a 95% confidence interval ranging from .35 to .52. This range is similar to that reported by Lee and Snarey (1988) above. Seventeen additional studies on the correlation between SCT scores and cognitive-developmental scores were identified and are presented in Appendix F. Many of these studies obtained correlations which fall within the range of .35 to .52, as reported by Cohn and Westenberg (2004). Correlations between .35 to .52 may be interpreted as moderate or large in general conventions (Cohen, 1988), but these correlations need to be interpreted in context. First, the range of .35 to .52 appears to fall outside the range of .86 to
.92 which has been found between scores based on hard stage models. Based on a hypothetical scenario in which a correlation of .52 is found between cognitive-developmental and SCT scores (N=200), and a correlation of .86 is found between two sets of cognitive-developmental scores (N=200), .52 would be significantly less than a correlation of .86 (N=400, z=-7.12, p<.001). Second, correlations between .35 and .52 suggest that scores exhibit only 12% to 27% shared variance. This is insufficient to suggest they tap a common construct because 73% to 88% of the variance in TPRs cannot be directly attributed to hierarchical complexity.

The third argument in support of hierarchical integrations is that TPRs are significantly correlated with increasingly complex leadership attributes. Rooke and Torbert (1998) examined the relation between 10 CEOs’ TPRs and their ability to lead organisational transformations. The independent variable was the CEO’s stage, and the dependent variable was whether their organisations transformed. The researchers found that five out of five CEOs who scored at the Strategist stage or above successfully led organisational transformation. However, only two CEOs who scored below the Strategist stage successfully led organisational transformation. The remaining three organisations did not transform at all. The correlation between TPRs and a measure of organisational transformation was significant (ρ=.65, ρ^2=.42, p<.05). In further analysis, it was reported that the combination of the CEOs’ and the consultants’ TPRs had an even stronger relation to organisational transformation (ρ=.77, ρ^2=.59, p<.01) (Torbert, 2004, 2013). The authors concluded that leaders’ and advisors’ stage of development is a primary predictor of organisational transformation. Although Torbert does not appear to make this argument in support of hard stage requirements, this evidence could be used to mount a case that leaders who score at higher stages can lead increasingly complex processes.

Although the study undertaken by Rooke and Torbert (1998) affords a unique opportunity to evaluate the development of senior organisational leaders, it also presents several methodological challenges. The small sample presents challenges to representativeness and generalisability as Torbert (2013) acknowledges. Only four organisational development consultants worked with the
10 CEOs, so the design may violate the assumption of independent observations. Finally, based on
the data provided in the publications, there are even higher correlations between organisational
transformation and the length of the consulting project ($\rho=.71, \rho^2=.51, p<.05$) and between
organisational transformation and the degree of consultant influence ($\rho=.72, \rho^2=.53, p<.05$).
Calculations on consultant influence were based on data reported by Torbert (2013) and Rooke and
Torbert (1998) in which 1=consultant only, 2= consultant and member of the management team, 3=
consultant and Board member. A sample size of 10 CEOs is too small to perform more sophisticated
analyses. However, in a larger sample a regression model may reveal that the relation between TPRs
and organisational transformation is weaker than that between organisational transformation and
other predictors. Furthermore, if the majority of variability in SCT scores is attributable to
unidentified secondary constructs and/or noise (Blasi, 1971; Kishton et al., 1984; Lambert, 1972a,
1972b; Loevinger & Wessler, 1970; Lorr & Manning, 1978), then it is unclear whether these
correlations suggest an empirical relation between ego development and leadership phenomena.
These results may be interpreted as suggesting randomness predicts leaders’ ability to foster
organisational transformation instead. Given that ego development has been shown to account for
only 20% of variability in SCT scores (Loevinger & Wessler, 1970; Lorr & Manning, 1978), it may have
been more instructive to examine the correlation between ego development factor scores and other
variables, if the sample size had been greater.

Torbert and others have found relations between TPRs and more favourable business
performance (Torbert, 1994), voluntarily seeking feedback on SCT results (Torbert, 1994), autonomy
in organisational roles (Torbert, 1991a), more effective ways of interacting with superiors and
subordinates (Fisher & Torbert, 1991), enquiring into others’ perspectives and generating second-
order solutions (e.g. redefining the problem) rather than first-order solutions (e.g. accepting given
definitions) (Merron et al., 1987), and higher team performance (Torbert, 1987a). Although these
attributes are considered desirable, it is challenging to determine if and how relations between TPRs
and these attributes relate to hierarchical integrations between stages. Also, given that ego
development only accounts for a minor amount of variability in SCT scores (Loevinger & Wessler, 1970; Lorr & Manning, 1978), it is difficult to establish whether these findings support the conclusion that ego development or randomness is related to complex leadership attributes.

Finally, in a recent linguistic analysis of approximately 44,000 stem completions, Lanning et al. (2018) found a significant relation between SCT stem scores and various aspects of cognitive complexity. The number of unique individuals who produced these stem completions was not reported. For example, they found a positive relation between SCT stem scores and linguistic differentiation, identification of cognitive processes, word length, insight, and comparative use of language. However, these cognitive aspects only accounted for approximately 1-2% of the variability in scores. This is less than what might be expected if the ego development scoring system were awarding scores based on considerations relating to hierarchical complexity.

**Rationale for Study 1B**

Specifying testable criteria for ego development presents a unique challenge. On one hand, ego developmentalists consistently claim their stages satisfy hard stage requirements. On the other, evidence to support this claim is considerably limited. In deference to ego developmentalists and to ensure symmetry between studies, Study 1B applied the same requirements as those applied to Study 1A.

**Criteria for Study 1B**

- Criterion 1: Scores satisfy the requirement of unidimensionality.
  - Justification: To be consistent with the Rasch literature, a family approach is taken to evaluating unidimensionality (Andrich, 2016; Andrich & Marais, 2019; Marais & Andrich, 2008b). This approach this takes the findings from a range of analytical procedures into account. A detailed description of these procedures was provided in the Results section for Study 1A above.
• Criterion 2: Invariant sequence. For each stem, threshold $\tau_n < \text{threshold } \tau_{n+1}$.
  
  o Justification: Thresholds, $\tau_n$, create partitions between successive stages, S to S+1. Each stem is associated with estimates for each threshold. Stages are sequenced correctly if thresholds are correctly sequenced across the logit scale.

• Criterion 3: Qualitatively distinct.
  
  o Criterion 3a) For all thresholds, the highest estimate for threshold $\tau_n < \text{the lowest estimate for threshold } \tau_{n+1}$.
    
    ▪ Justification: Thresholds, $\tau_n$, create partitions between successive stages, S to S+1. Stage S is qualitatively distinct from Stage S+1 if the thresholds which flank Stage S are sufficiently spaced apart. The highest estimate for threshold $\tau_n$ should be significantly < the lowest estimate for threshold $\tau_{n+1}$.

  o Criterion 3b) The Person Separation Index (PSI) is equal to or greater than .96, which indicates that seven statistically distinct stages can be measured.
    
    ▪ Justification: Ordinal categories operationalise stages. There must be consistency between the number of stages the ego development scoring system claims to measure, i.e., Impulsive through Unitive, and the number of stages it can measure based on reliability indices.

• Criterion 4: Structure of the whole.
  
  o Criterion 4a) Category probability curves reach an apex of > .80
    
    ▪ Justification: Category probability curves reflect the probability of a response being scored at a certain stage. According to the requirement of structured wholeness, a single stage is expected to organise the worldview. Category probability curves > .80 indicate a high likelihood of a single stage organising responses to the stems.
Criterion 4b) Category probability curves only converge with adjacent curves, and convergence between nonadjacent curves should occur with a probability of < .10.

- Justification: If a single stage organises the worldview, a minority of responses may occur in adjacent stages. Category probability curves should only converge with nonadjacent curves with a probability of < .10.

- Criterion 5: Hierarchical integrations. Item equating graphs exhibit spurts and plateaus of development.

- Justification: As discussed above, stage S+1 is considered to be a hierarchical integration of its predecessor stage S if and only if stage S+1 is defined in terms of two or more elements of stage S, it organises and transforms the elements of stage S, and does so in a non-arbitrary way (Commons & Pekker, 2008). Hierarchical integrations result in a worldview that consolidates at a single stage before vacillating between the current stage and its successor. This pattern is reflected in spurts and plateaus of growth in item equating graphs (refer to Figure 5).

Methods

Participants

The data were collected by three leadership providers based in various global locations between the years 2009 and 2015. The sample consisted of 666 leaders in a variety of roles and organisations who had completed at least one SCT during participation in various leader development initiatives. Leaders consented to their data being used for legitimate research purposes (refer to Appendix H). For some leaders, consent was provided directly to the author of this thesis. In other cases, consent was provided to the leadership providers, and in collaboration with The University of Western Australia, consent was extended to the author of this thesis. An optimal sample with sufficient power to identify misfit was approximated to be 6,600 (i.e., n = Total thresholds x 15, where 55 stems x (9 stages - 1) x15). Given the limited power of the current sample
to identify misfit, aggregated data were used wherever possible. When data were collected, test-takers were provided with a binary option to select for sex. Demographic details of the leaders are presented in Table 15.

**Table 15**

*Demographic Details of Leaders for Study 1B*

<table>
<thead>
<tr>
<th>Demographics</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>224</td>
<td>34%</td>
</tr>
<tr>
<td>Female</td>
<td>194</td>
<td>29%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>248</td>
<td>37%</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18–35 years</td>
<td>105</td>
<td>16%</td>
</tr>
<tr>
<td>36–45 years</td>
<td>138</td>
<td>21%</td>
</tr>
<tr>
<td>46–71 years</td>
<td>118</td>
<td>18%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>305</td>
<td>46%</td>
</tr>
<tr>
<td><strong>Birth year</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1940–1959</td>
<td>64</td>
<td>10%</td>
</tr>
<tr>
<td>1960–1979</td>
<td>259</td>
<td>39%</td>
</tr>
<tr>
<td>1980–1999</td>
<td>38</td>
<td>6%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>305</td>
<td>46%</td>
</tr>
<tr>
<td><strong>First language</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>471</td>
<td>71%</td>
</tr>
<tr>
<td>Not English</td>
<td>104</td>
<td>16%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>91</td>
<td>14%</td>
</tr>
</tbody>
</table>

Data pertaining to education, occupation, and culture of origin were not gathered with sufficient consistency to enable analysis. Inspection of role titles indicated some representativeness across management layers. Given the data were collected from leaders in the USA, UK, and Australia, some level of representativeness across western cultures is suggested.

**Materials**

**Assessments.** Five different versions of the SCT were administered to leaders: SCT Professional Sentence Completion Form, SCT-Integral Life Studio (SCT-ILS), Global Leadership Profile (GLP), Washington University Sentence Completion Test (WUSCT) Form 81 Parts A and B. Refer to Appendix G for examples of the SCT in the public domain. Each version represented a particular
adaptation of the SCT for different contexts. SCTi-MAP Professional Sentence Completion Form consists of 36 stems and is intended for professional and clinical applications (Cook-Greuter, 2006). The SCT-Integral Life Studio (SCT-ILS) consists of 36 stems and is an adaptation of the SCTi-MAP Professional for leaders who participate in development processes informed by Integral Theory (Wilber, 2000). The GLP is a version of the SCT that includes 30 stems and uses terminology that better reflects workplace leadership (Herdman-Barker & Torbert, 2012). Forms 81 Parts A and B each contain 18 stems and are often used for pre- and post-assessments or split-half reliability studies (Loevinger, 1985; Loevinger & Cohn, 1998; Redmore & Waldman, 1975).

Scoring is undertaken in several steps. In step one, stem completions are scored without regard to context by matching test-takers’ responses to examples in the ego development scoring system. In step two, the whole protocol is reviewed, and an impressionistic score is recorded. In step three, the cumulative frequency distribution of scores from step one is compared to cut-scores to determine an automatic TPR. Finally, when the impressionistic score from step two coincides with the automatic TPR from step three, a final TPR is awarded. When the impressionistic score from step two does not coincide with the TPR from step three, a final TPR is awarded by appealing to ego development theory and nonpsychometric signs e.g., repetition of words, phrases, or ideas (Hy & Loevinger, 1996). The TPR is used to make inferences about test-takers’ centres of gravity.

**Software.** Analyses were based on the Rasch Model (Andrich, 1988, 2005; Rasch, 1980) and conducted in the Rasch Unidimensional Measurement Model (RUMM2030) software (Andrich et al., 2015).

**Procedure**

**Administration.** The data were collected by three leadership providers based in various global locations. The data structure is presented in Table 16, and short descriptions of the development initiatives are presented in Table 17. The SCTs were administered in a formative and summative manner depending on the requirements of each leader development initiative.
Table 16

Data Structure of the SCT Sample

<table>
<thead>
<tr>
<th>Provider</th>
<th>Program</th>
<th>No. of leaders</th>
<th>SCT version</th>
<th>No. of stems</th>
<th>SCT scorer</th>
<th>No. of SCTs per test-time</th>
<th>Total SCTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>94</td>
<td>SCTi-ILS</td>
<td>36</td>
<td>1</td>
<td>94</td>
<td>139</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>80</td>
<td>SCTi-Professional</td>
<td>36</td>
<td>1</td>
<td>26</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GLP</td>
<td>30</td>
<td>2</td>
<td>53</td>
<td>67</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>253</td>
<td>Form 81 Part A</td>
<td>18</td>
<td>3</td>
<td>252</td>
<td>252</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Form 81 Part B</td>
<td>18</td>
<td>3</td>
<td>0</td>
<td>251</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>239</td>
<td>GLP</td>
<td>30</td>
<td>2</td>
<td>239</td>
<td>239</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>666</td>
<td>5</td>
<td>55</td>
<td>3</td>
<td>664</td>
<td>996</td>
</tr>
</tbody>
</table>

leaders | versions | stems | scorers | SCTs | SCTs | SCTs | SCTs |
<table>
<thead>
<tr>
<th>Program</th>
<th>Delivery locations</th>
<th>Focus</th>
<th>Pedagogical structure</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>On campus, online options</td>
<td>Theory and practice of Integral Theory</td>
<td>10 quarter-long modules</td>
<td>2 years</td>
</tr>
<tr>
<td>2</td>
<td>Face-to-face facilitation in various global locations</td>
<td>Navigating 21st century complexity.</td>
<td>15 days of facilitation, individual coaching, communities of practice, developmental practices, application to personal or professional project</td>
<td>9 months</td>
</tr>
<tr>
<td>3</td>
<td>Face-to-face facilitation in Australia</td>
<td>Adaptive challenges faced in an increasingly VUCA world.</td>
<td>2-3 half or full day modules per month, reading, reflection, action learning, 3 field trips</td>
<td>10 months</td>
</tr>
<tr>
<td>4</td>
<td>Various delivery locations</td>
<td>Applications of Collaborative Developmental Action Enquiry</td>
<td>Various</td>
<td>Various</td>
</tr>
</tbody>
</table>

**Scoring.** As can be seen from Table 16, there were three scorers involved in the scoring of this sample. Scorers had different levels of training and experience. Scorers 1 and 2 had received formal training and had multiple years of experience. Scorer 3 was self-taught through the process recommended by Hy and Loevinger (1996).

**Data Preparation.** Nine ego development stages were identified in the sample and they were entered as categories 0 to 8 in RUMM2030 (Andrich et al., 2015). As a result, SCT stem scores rather than TPRs formed the basis for analysis. The data structure is considered to be a linked design because different versions of the SCT shared some common stems and also had some unique stems (Styles & Andrich, 1993). The full dataset was relatively complex and consisted of 996 protocols produced by 666 leaders. It included 55 stems and five versions of the SCT. A list of stems can be found in Appendix I.
Including multiple test-times in an analysis can introduce dependence into the data such that responses to a stem at time t+1 may depend on responses at time t, thereby causing inflated parameters (Marais & Andrich, 2008a). Using multiple test-times as independent observations is justified because consecutive administrations were spaced by a minimum of seven months (Willett, 1989), artificial dependence introduced to the data as a function of analytical procedures can be detected through Rasch analyses, and it is common practice in the analysis of developmental and non-developmental data alike (Armon & Dawson, 1997; Bond & Fox, 2015; Dawson, 2000b, 2002b; Puddey et al., 2014).

**Use of the Rasch Model to Analyse SCT Scores**

The Rasch Model cannot only be used to evaluate general psychometric properties of developmental assessments but can also determine the extent to which scores satisfy hard stage requirements. In the context of ego development stages, Rasch parameters take on a particular meaning: the logit scale operationalises the underlying construct of ego development which is contended to increase hierarchically. Strictly speaking, the logit scale is a scale of difficulty. However, hierarchical complexity is understood to be an instantiation of difficulty because more hierarchically complex performance is also more difficult. Person ability reflects the centre of gravity of the ego; ordinal categories operationalise stages; and thresholds create partitions between successive stages, S to S+1. Given a person n who possesses an ego which is developed to stage s, and a SCT sentence stem i which has a difficulty of δi, the Rasch model estimates the probability of person n’s response to stem i being scored at each stage from Impulsive through to Unitive.

**Analytical Approach**

Analyses were undertaken in a similar sequence to Study 1A and are described in Table 18.
### Table 18

**Steps Applied to the Rasch Analysis of the SCT**

<table>
<thead>
<tr>
<th>Step</th>
<th>Purpose</th>
<th>No.</th>
<th>Analysis name</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Confirm if the whole dataset could be analysed simultaneously</td>
<td>1a</td>
<td>Facet analysis of SCT scorer severity</td>
<td>To determine whether the 3 SCT scorers awarded equivalent scores to test-takers of equal ability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1b</td>
<td>Invariance between versions of the SCT</td>
<td>To determine whether the 5 versions of the SCT awarded sufficiently similar scores to test-takers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1c</td>
<td>Retention of stages and stems</td>
<td>To determine if the 9 stages and 55 stems could be retained for further analysis</td>
</tr>
<tr>
<td>2</td>
<td>Summary statistics</td>
<td>2</td>
<td>Summary statistics</td>
<td>To provide general summary statistics for the final dataset and present basic fit statistics</td>
</tr>
<tr>
<td>3</td>
<td>Direct tests of hard stage requirements</td>
<td>3a</td>
<td>Unidimensionality</td>
<td>To determine if there was a single underlying factor measured by the stems i.e., ego development</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3b</td>
<td>Invariant sequence</td>
<td>To determine if stem thresholds and stages occurred in the sequence predicted by theory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3c</td>
<td>Qualitatively distinct</td>
<td>To determine if SCT thresholds were sufficiently spaced apart to create partitions between qualitatively distinct stages</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3d</td>
<td>Structure of the whole</td>
<td>To determine if a single stage organised each test-taker’s responses to the stems i.e., their centre of gravity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3e</td>
<td>Hierarchical integrations</td>
<td>To determine if stages exhibited spurts and plateaus as predicted by hard stage requirements</td>
</tr>
<tr>
<td>4</td>
<td>Test for other general measurement properties</td>
<td>4</td>
<td>Differential item functioning, response dependence, measurement precision, guessing/cheating analysis</td>
<td>Additional analyses of general measurement properties directly related to hard stage requirements</td>
</tr>
</tbody>
</table>
Results

**Step 1: Confirm if the whole dataset could be analysed simultaneously**

The purpose of Step 1 was to confirm if the whole dataset could be analysed simultaneously.

**Step 1a: Facet Analysis of SCT Scorer Severity.** This analysis aimed to determine whether the three SCT scorers awarded equivalent scores to test-takers of equal ability. If the scorers were equally severe, then the whole dataset could be analysed simultaneously to maximise statistical power. This was undertaken by a facet analysis which introduced a third facet, scorer severity, to the two facets included in the standard Rasch model i.e., person ability and stem difficulty (Linacre, 1989; Linacre & Wright, 2002; Lunz et al., 1990). A subset of 11 stems scored by all three scorers were analysed (i.e., stems 1, 2, 5, 6, 7, 9, 10, 11, 14, 18, 20). Ten stems displayed threshold disorder, possibly because of the underutilisation of extreme categories. This could not be resolved by rescoring, so the total number of stages was reduced from nine to seven across all stems. Stems which still displayed disorder were deleted, so that only six stems with correctly ordered thresholds were retained, i.e., stems 2, 6, 10, 11, 14, 18. To mitigate potential bias, an equal number of cases per scorer was identified through random sampling. As can be seen from Table 19, scorer severity occurred with a narrow range (min = -0.05, max = 0.08) and fit residuals occurred close to zero (min = 0.00, max = 0.41). A lack of fit is indicated if scorer severity for one or more scorers significantly differs from other scorers and/or if fit residual exceeds the critical value + 2.5. Findings suggest that scorers were equally severe, so the whole dataset could be analysed simultaneously.

**Table 19**

*Facet Analysis on Three SCT Scorers*

<table>
<thead>
<tr>
<th>Scorer</th>
<th>n</th>
<th>Scorer severity ($\lambda_h$)</th>
<th>SE</th>
<th>Fit Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>187</td>
<td>-0.05</td>
<td>0.08</td>
<td>0.39</td>
</tr>
<tr>
<td>2</td>
<td>187</td>
<td>-0.04</td>
<td>0.09</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>187</td>
<td>0.08</td>
<td>0.11</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Step 1b: Invariance Between Versions of the SCT. This analysis aimed to determine whether the five versions of the SCT awarded similar scores to test-takers of equivalent ability. This was undertaken via an item equating procedure which examines the relation between person abilities and SCT scores. Of the 55 stems in the original sample, threshold disorder was observed in six stems (stems 1, 7, 9, 50, 52, 55), possibly because of underutilisation of extreme categories. Before analysing invariance between versions of the SCT, a methodical data cleaning processes was undertaken to resolve all instances of disorder. The data cleaning processes consisted of rescoring all stems to reduce the number of stages one at a time i.e., from nine in the original dataset to eight, from eight to seven, and so on. The overarching goal was to resolve disorder in a manner that did not reduce the reliability estimates or unnecessarily reduce the total number of ego development stages that could be retained for analysis. As discussed earlier, Rasch analysis typically involves making iterative revisions to the dataset to better approximate the ideal form of measurement, as operationalised by the model. As a result, the rescoring processes applied to Step 1b are standard practice in the context of a Rasch analysis (Andrich & Marais, 2019). The best model resulted from rescoring the total number of stages from nine to seven across all stems and resolving outstanding instances of disorder on a stem-by-stem basis. Stems were entered into the equating analysis based on whether they were included in each version of the SCT (refer to Appendix I). A line graph informed visual analysis of person ability (x-axis) plotted against the mean SCT score (y-axis). Inspection of Figure 17 showed that the mean scores were closely related for all person abilities from -7 to +13 logits.
Figure 17

*Common Item Equating Conducted on Five Versions of the SCT Showing Person Ability (Vertical) and Mean SCT Stem (Horizontal)*

This was further examined via a correlation matrix on a sub-sample of 200 randomly chosen persons to ensure an appropriate level of statistical power. An identity relation was observed between all pairwise comparisons \( n=200, r=1.00, r^2=1.00, p<.001 \) as reported in Table 20. These analyses suggested that different versions of the SCT were awarding similar scores to test-takers of equal ability.

Table 20

*Intercorrelations Between Five Versions of the SCT Included in Study 1B*

<table>
<thead>
<tr>
<th>SCT version</th>
<th>M</th>
<th>SD</th>
<th>SCTi-Professional</th>
<th>SCTi-ILS</th>
<th>GLP</th>
<th>Form 81 Part A</th>
<th>Form 81 Part B</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCT-Professional</td>
<td>4.62</td>
<td>2.21</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCT-ILS</td>
<td>4.57</td>
<td>2.19</td>
<td>1.00*</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLP</td>
<td>4.60</td>
<td>2.16</td>
<td>1.00*</td>
<td>1.00*</td>
<td>1.00*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form 81 Part A</td>
<td>4.67</td>
<td>2.22</td>
<td>1.00*</td>
<td>1.00*</td>
<td>1.00*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form 81 Part B</td>
<td>4.43</td>
<td>2.15</td>
<td>1.00*</td>
<td>1.00*</td>
<td>1.00*</td>
<td>1.00*</td>
<td></td>
</tr>
</tbody>
</table>

*Note. * p<.01
**Step 1c: Retention of Ego Development Stages and SCT Stems.** This step aimed to determine whether all nine stages and 55 stems could be retained for further analysis. Stems discussed in the following section are presented in Appendix I. Decisions about stage retention were based on three considerations: utilisation of stages, targeting of thresholds to person ability, and correct threshold sequencing. The response distribution across stages is presented in Table 21.

**Table 21**

*Response Distribution Across Nine Ego Development Stages*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
<th>Stage 5</th>
<th>Stage 6</th>
<th>Stage 7</th>
<th>Stage 8</th>
<th>Stage 9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>31</td>
<td>219</td>
<td>2061</td>
<td>4986</td>
<td>9122</td>
<td>5808</td>
<td>2196</td>
<td>501</td>
<td>39</td>
<td>24,963</td>
</tr>
<tr>
<td>Percentage</td>
<td>0.12</td>
<td>0.88</td>
<td>8.26</td>
<td>19.97</td>
<td>36.54</td>
<td>23.27</td>
<td>8.80</td>
<td>2.01</td>
<td>0.16</td>
<td>100.00</td>
</tr>
<tr>
<td>Cumulative percentage</td>
<td>0.12</td>
<td>1.00</td>
<td>9.26</td>
<td>29.23</td>
<td>65.77</td>
<td>89.04</td>
<td>97.84</td>
<td>99.84</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

The distribution appeared to approximate a normal distribution, which is unnecessary for Rasch analysis but typical in leadership samples. Some underutilisation occurred in the lowest and highest stages. This may be a function of the narrow range of stages over which workplace leaders typically function. This may suggest that a more parsimonious model could account for the response distribution. Uncentralised thresholds were used throughout Study 1B rather than centralised thresholds because uncentralised thresholds are referenced to the same origin so they can be compared across stems rather than being mean deviated about each stem’s location. The person item-threshold distribution shows the logit scale on the horizontal axis with person ability plotted in the upper section and stem thresholds in the lower section. Stem thresholds occurred over a broad range of 31.47 logits, while person abilities were more narrowly distributed over 5.55 logits (refer to Figure 18). All persons could be allocated to a stage, but persons were not well separated. Many of the thresholds at the lower and upper end of the scale are redundant because there are no persons in range to measure.
Stages are correctly sequenced if thresholds are correctly sequenced across the logit scale.

Six stems displayed threshold disorder (refer to Table 22). Stems 1, 9, 50 and 52 displayed disorder between the lowest two thresholds. Stem 55 displayed disorder between the highest two thresholds. Stem 7 displayed disorder between the lowest and highest two thresholds. This may result from underutilisation at the upper and lower end of the ego development scale, as discussed above.

Table 22

Uncentralised Threshold Locations for Stems with Disordered Thresholds

<table>
<thead>
<tr>
<th>SCT Stem</th>
<th>Stem Location</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_3$</th>
<th>$\tau_4$</th>
<th>$\tau_5$</th>
<th>$\tau_6$</th>
<th>$\tau_7$</th>
<th>$\tau_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.23</td>
<td>-2.33*</td>
<td>-2.84*</td>
<td>-2.45</td>
<td>-1.45</td>
<td>-0.09</td>
<td>1.35</td>
<td>2.60</td>
<td>3.39</td>
</tr>
<tr>
<td>7</td>
<td>-0.32</td>
<td>-2.47*</td>
<td>-3.23*</td>
<td>-2.70</td>
<td>-1.37</td>
<td>0.30</td>
<td>1.81</td>
<td>2.68*</td>
<td>2.44*</td>
</tr>
<tr>
<td>9</td>
<td>-0.27</td>
<td>-3.08*</td>
<td>-3.46*</td>
<td>-2.76</td>
<td>-1.36</td>
<td>0.34</td>
<td>1.93</td>
<td>3.02</td>
<td>3.21</td>
</tr>
<tr>
<td>50</td>
<td>0.85</td>
<td>-2.57*</td>
<td>-2.68*</td>
<td>-2.15</td>
<td>-1.05</td>
<td>0.53</td>
<td>2.51</td>
<td>4.82</td>
<td>7.36</td>
</tr>
<tr>
<td>52</td>
<td>0.23</td>
<td>-2.63*</td>
<td>-2.84*</td>
<td>-2.08</td>
<td>-0.72</td>
<td>0.88</td>
<td>2.36</td>
<td>3.36</td>
<td>3.53</td>
</tr>
<tr>
<td>55</td>
<td>-0.16</td>
<td>-3.43</td>
<td>-3.21</td>
<td>-2.11</td>
<td>-0.55</td>
<td>1.06</td>
<td>2.29</td>
<td>2.73*</td>
<td>1.97*</td>
</tr>
</tbody>
</table>

Note. * Indicates disorder with at least one adjacent threshold
The threshold probability curve for stem 1 is shown in Figure 19, where it can be seen that thresholds 1 and 2 are reversed.

**Figure 19**

*Illustrative Example of Threshold Probability Curve for SCT Stem 1*

Three tests suggested that the full Rasch analysis should be based on a more parsimonious model with less than nine stages. As discussed earlier, a methodical data cleaning processes was undertaken to resolve all instances of disorder. The data cleaning processes consisted of rescoring all stems to reduce the number of stages one at a time i.e., from nine in the original dataset to eight, from eight to seven, and so on. The overarching goal was to resolve disorder in a manner that did not reduce the reliability estimates or unnecessarily reduce the total number of ego development stages that could be retained for analysis. The model that fit best was one in which the total number of stages was reduced from nine to seven by combining the two lowest stages (i.e., Impulsive and Opportunist) and the two highest stages (i.e., Construct-Aware and Unitive). There was no change to the Personal Separation Index (PSI) of .95, and there was still a sufficiently broad range of stem thresholds (min=-6.46, max=11.32, range=17.78), to assess persons of all abilities (min=-2.59, max=3.11, range=5.70). However, this model resulted in a larger number of stems with threshold disorder (stems 1, 3, 7, 9, 15, 23, 27, 28, 54, 55) (refer to Table 23). Disorder was successfully
resolved for these individual stems by rescoring categories for which disorder was present. These
data cleaning processes are standard practice in the context of a Rasch analysis (Andrich & Marais, 2019).

Table 23

Uncentralised Threshold Locations for Stems with Disordered Thresholds in the Seven-Stage Model

<table>
<thead>
<tr>
<th>Stem</th>
<th>Location</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_3$</th>
<th>$\tau_4$</th>
<th>$\tau_5$</th>
<th>$\tau_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.28</td>
<td>-2.06*</td>
<td>-2.29*</td>
<td>-1.44</td>
<td>-0.01</td>
<td>1.50</td>
<td>2.59</td>
</tr>
<tr>
<td>3</td>
<td>-0.41</td>
<td>-2.21*</td>
<td>-2.39*</td>
<td>-1.31</td>
<td>0.25</td>
<td>1.51</td>
<td>1.69</td>
</tr>
<tr>
<td>7</td>
<td>-0.28</td>
<td>-2.22*</td>
<td>-2.67*</td>
<td>-1.41</td>
<td>0.50</td>
<td>2.01</td>
<td>2.07</td>
</tr>
<tr>
<td>9</td>
<td>-0.26</td>
<td>-2.35*</td>
<td>-2.79*</td>
<td>-1.43</td>
<td>0.61</td>
<td>2.19</td>
<td>2.19</td>
</tr>
<tr>
<td>15</td>
<td>-0.25</td>
<td>-2.24</td>
<td>-2.09</td>
<td>-0.84</td>
<td>0.68</td>
<td>1.67*</td>
<td>1.30*</td>
</tr>
<tr>
<td>23</td>
<td>0.17</td>
<td>-1.82*</td>
<td>-1.90*</td>
<td>-0.85</td>
<td>0.69</td>
<td>2.12</td>
<td>2.80</td>
</tr>
<tr>
<td>27</td>
<td>-0.19</td>
<td>-2.23*</td>
<td>-2.48*</td>
<td>-1.16</td>
<td>0.71</td>
<td>2.09*</td>
<td>1.95*</td>
</tr>
<tr>
<td>28</td>
<td>0.17</td>
<td>-1.58*</td>
<td>-1.62*</td>
<td>-0.97</td>
<td>0.20</td>
<td>1.68</td>
<td>3.29</td>
</tr>
<tr>
<td>54</td>
<td>0.79</td>
<td>-1.74*</td>
<td>-1.86*</td>
<td>-0.97</td>
<td>0.72</td>
<td>2.99</td>
<td>5.61</td>
</tr>
<tr>
<td>55</td>
<td>0.09</td>
<td>-2.60</td>
<td>-2.08</td>
<td>-0.50</td>
<td>1.27</td>
<td>2.40*</td>
<td>2.06*</td>
</tr>
</tbody>
</table>

Note. * Indicates disorder with at least one adjacent threshold

Decisions about stem retention were based on two considerations: alignment of stages across stems and stem fit. The threshold map shows the location of stages across the logit scale. Alignment of stages across stems is important for stages to reflect equivalent amounts of development and, therefore, to have a consistent meaning. As can be seen from Figure 20, there is a considerable misalignment of stages across stems. For example, for stem 7, stage 3 (Diplomat) represents a similar amount of development as stage 4 (Expert) for most other stems. Visual analysis indicated that 12 stems contained at least one stage whose location appeared to be significantly misaligned i.e., stems 1, 3, 7, 9, 13, 15, 23, 27, 28, 51, 54, and 55. This was cross-validated against the threshold locations to mitigate visual bias. These 12 stems included 10 of the stems which initially displayed threshold disorder from Analysis 1c above. While rescoring successfully resolved disorder, it appeared to create stage misalignment between stems. In order to complete the data cleaning processes as described above (Andrich & Marais, 2019), these 12 stems were deleted and the analysis was re-run with the remaining items.
Figure 20

Threshold Map for the 55 Stems and the Location of Stages in the Seven-Stage Model

Note. 1/blue=Combined Impulsive and Opportunist. 2/red=Diplomat. 3/green=Expert.
4/Mauve=Achiever. 5/pink=Individualist. 6/brown=Strategist. 7/yellow=Combined Construct-Aware-Unitive.

Stem fit was analysed for the remaining 43 stems through statistical and visual procedures.

Stems were considered to misfit if the Fit Residual statistic exceeded the critical value of +2.5. Stem
2 (Fit Residual=-3.01), stem 40 (Fit Residual=-2.93), stem 44 (Fit Residual=-2.84) and stem 22 (Fit Residual=2.92) misfit the model. However, no stems misfit according to the chi-square statistic. Only stem 44 misfits according to the ANOVA statistic with a Bonferroni correction for 43 stems (i.e., .05/43 = .0011) (F(9,296)=3.84, p<.001). Of the four stems that misfit according to the Fit Residual statistic, only stem 44 appeared to overdiscriminate in the lower- and upper-class intervals according to the Item Characteristic Curves (ICC). It was considered to overdiscriminate because the actual values for the lower-class intervals were lower than the expected values and the actual values for the upper-class intervals were higher than the expected values (refer to Figure 21). As a result, stem 44 was deleted.

Figure 21

Illustrative Example of Item Characteristic Curve for Stem 44 from the SCT

In summary, 24% of stems (n=13) appear to be inconsistent with other stems. This is considerably higher than what would be expected through random chance alone (e.g., 5%). This finding may warrant a thorough content analysis of these stems. Alternatively, they should be removed from future versions of the SCT. The full Rasch analysis was based on a reduced dataset that retained seven stages, 42 stems, and 996 cases.
**Step 2: Summary statistics**

This step aimed to provide general summary statistics and basic fit statistics. Summary statistics demonstrated that mean person ability was relatively consistent (M= -0.17, SD=1.03, min= -2.55, max=3.24, range =5.79) with average stem difficulty (M = 0.00, SD= 0.32, min= -0.69, max=0.84, range =1.53). This suggests that many leaders experienced stem difficulty to be well matched to their ability.

Person abilities were distributed over 5.79 logits, and stem difficulties were distributed over 1.35 logits. The range of stem difficulties represents 23% of the range of person abilities. The relatively narrow range of person abilities suggests that the SCT is less successful at distinguishing between test-takers of different abilities. The range of stem difficulties is narrow, but relative to the distribution of person abilities, this is relatively broad. As mentioned in Study 1A, stems are often constructed to reflect an equivalent amount of difficulty to provide test-takers with an equal opportunity on all stems. The ego development literature explicitly states that stems are interchangeable (Loevinger, 1993a, 1998c), but this may only be possible if stems reflect a similar amount of difficulty. However, as discussed in Study 1A, assessments that consist of items with similar difficulties may be regarded as consisting of only one item, and this may result in compromised measurement properties in some contexts (Andrich, 1988; Rasch, 1980).
Figure 22

Person-Item Location Distribution for the SCT Showing Person Ability (top) and Average Stem Location (Bottom) Along the Logit Scale

Fit residuals for stems indicated slight overfit to the model because they deviate from a mean of 0.00 and SD of 0.00 (M = -0.36, SD = 1.47, skewness = 0.02, kurtosis = 0.95). Fit residuals for persons also seemed to indicate slight overfit (M = -0.42, SD = 1.32, skewness = 0.20, kurtosis = 0.77).

Stem Fit. Stem fit was analysed statistically and graphically. Statistical analysis was undertaken via fit residuals, chi-square, and ANOVA statistics. Graphical analysis was undertaken via Characteristic Curves (ICCs). A stem was considered to misfit if the fit residual exceeded the critical values of ±2.5, or if the chi-square or ANOVA fit statistic exceeded a critical value of alpha with a Bonferroni correction (.05/42 stems = .001). Stems 2, 14, 31, and 40 misfit on at least one statistical test (refer to Table 24).
### Table 24

*Stem Fit Statistics for the Stems which Misfit on at Least One Statistical Test*

<table>
<thead>
<tr>
<th>Stem</th>
<th>Location</th>
<th>SE</th>
<th>Fit residual</th>
<th>DF</th>
<th>$\chi^2$ statistic</th>
<th>DF</th>
<th>p</th>
<th>F</th>
<th>DF1</th>
<th>DF2</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.69</td>
<td>0.04</td>
<td>-3.13</td>
<td>699.93</td>
<td>13.74</td>
<td>9</td>
<td>.13</td>
<td>1.91</td>
<td>9</td>
<td>735</td>
<td>.05</td>
</tr>
<tr>
<td>14</td>
<td>0.25</td>
<td>0.04</td>
<td>1.72</td>
<td>699.93</td>
<td>29.81</td>
<td>9</td>
<td>.00</td>
<td>3.26</td>
<td>9</td>
<td>735</td>
<td>.00</td>
</tr>
<tr>
<td>31</td>
<td>0.38</td>
<td>0.05</td>
<td>2.75</td>
<td>463.18</td>
<td>15.73</td>
<td>9</td>
<td>.07</td>
<td>1.55</td>
<td>9</td>
<td>483</td>
<td>.13</td>
</tr>
<tr>
<td>40</td>
<td>-0.43</td>
<td>0.07</td>
<td>-2.99</td>
<td>287.49</td>
<td>12.61</td>
<td>9</td>
<td>.18</td>
<td>1.99</td>
<td>9</td>
<td>296</td>
<td>.04</td>
</tr>
</tbody>
</table>

Stem 2 overfit the model according to the fit residual statistic (-3.13) but not according to the chi-square ($\chi^2(9)=13.74$, $p=.13$) or ANOVA fit statistics ($F(9,735)=1.91$, $p=.05$). Stem 14 significantly misfit the model according to the chi-square ($\chi^2(9)=28.81$, $p<.001$) and the ANOVA fit statistics ($F(9,735)=3.26$, $p<.001$) but not the fit residual (1.72). Stem 31 significantly underfit the model according to the fit residual statistic (2.75) but did not misfit according to the chi-square ($\chi^2(9)=15.73$, $p=.07$) or ANOVA fit statistics ($F(9,483)=1.55$, $p=.13$). Stem 40 significantly overfit the model according to the fit residual statistic (-2.99) but did not misfit according to the chi-square ($\chi^2(9)=12.61$, $p=.18$) or ANOVA fit statistics ($F(9,296)=1.99$, $p=.04$). Visual analysis of ICCs indicated mild deviation between observed and predicted values. Stem 14 displayed the greatest misfit according to the ICCs because of underdiscrimination in the lower- and upper-class intervals (refer to Figure 23).
Since all instances of misfit appeared to be relatively moderate, there was insufficient
evidence to warrant further stem deletion.

**Person Fit.** No persons with extreme response patterns were identified. Eighty-two cases (8.2%) had fit residuals that exceeded the critical value of $\pm 2.5$, of which 60 cases (6.0%) overfit the model (i.e., < -2.5) and 22 cases (2.2%) underfit the model (i.e., > +2.5). Guttman patterns for misfitting persons suggested that overfit generally occurred when persons were awarded scores within one or two stages across all stems and that underfit occurred when persons were awarded scores across a broad range of stages across stems. In summary, a significant number of persons misfit the model but for reasons that may be consistent with ego development theory.

**Step 3: Direct Tests of Hard Stage Requirements**

The purpose of Step 3 was to directly test whether SCT scores satisfied hard stage
requirements.

**Step 3a: Unidimensionality.** This analysis aimed to determine if stems were measuring a
single underlying construct, i.e., ego development. While ego development is a broad construct
related to how test-takers conceive of themselves and reality in general, the SCT only provides a
single score to reflect test-takers’ development. Unidimensionality is satisfied if stems measure a single construct and if a total score on the assessment suffices to account for test-takers’ performance. Unidimensionality requires that for any stem of difficulty $\delta_i$, variation among responses to that stem can only be attributed to person ability, $\beta_n$. Multidimensionality occurs when a single score is insufficient to account for test-takers’ performance and implies the presence of secondary constructs and/or noise (Marais & Andrich, 2008a, 2008b). Secondary constructs typically need to be measured via a subset of stems and are reflected in separate scores. Because of the linked design, unidimensionality analyses could not be undertaken on the aggregated sample. Instead, analyses were conducted on the five subsamples, and a summary of the results is presented in Table 25. The analysis was undertaken following steps: PCA, subtest analysis, and item equating (RUMM Laboratory, 2009b).
### Table 25

**Analyses Conducted on the Unidimensionality of the SCT**

<table>
<thead>
<tr>
<th>#</th>
<th>Subsample</th>
<th>N</th>
<th>N</th>
<th>Eigen</th>
<th>%</th>
<th>Eigen</th>
<th>%</th>
<th>Subtest 1 stems</th>
<th>PSI</th>
<th>Subtest 2 stems</th>
<th>PSI</th>
<th>Subtest Analysis</th>
<th>c</th>
<th>p</th>
<th>A</th>
<th>Stem block</th>
<th>Stem block</th>
<th>r</th>
<th>r²</th>
<th>Paired t-test</th>
<th>5% critical alpha</th>
<th>1% critical alpha</th>
<th>Highest Corr. Between Stems</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sub-</td>
<td>N</td>
<td>N</td>
<td>Eigen</td>
<td>%</td>
<td>Eigen</td>
<td>%</td>
<td>Subtest</td>
<td>PSI</td>
<td>Subtest</td>
<td>PSI</td>
<td>Item Equating and Paired T-Tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sample</td>
<td>stems</td>
<td>cases</td>
<td>value</td>
<td>var</td>
<td>value</td>
<td>var</td>
<td>1 stems</td>
<td></td>
<td>2 stems</td>
<td></td>
<td>Residual Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>11</td>
<td>745</td>
<td>1.5</td>
<td>13.2%</td>
<td>1.4</td>
<td>12.3</td>
<td>6, 8, 9, 1, 2, 4, 5, 7</td>
<td>.92</td>
<td>.89</td>
<td>.20</td>
<td>.96</td>
<td>.97</td>
<td>6, 8, 9, 1, 2, 4</td>
<td>.79</td>
<td>.63</td>
<td>t(744)</td>
<td>6.98%</td>
<td>2.01%</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>common</td>
<td>%</td>
<td></td>
<td>10, 11</td>
<td></td>
<td></td>
<td></td>
<td>10, 11</td>
<td></td>
<td>5, 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>SCTI-LS</td>
<td>34</td>
<td>94</td>
<td>2.9</td>
<td>8.5%</td>
<td>2.6</td>
<td>7.7%</td>
<td>7, 9, 10, 1, 2, 3, 4, 5</td>
<td>.93</td>
<td>.85</td>
<td>.45</td>
<td>.83</td>
<td>.91</td>
<td>7, 9, 1, 2, 3</td>
<td>.74</td>
<td>.55</td>
<td>t(93)</td>
<td>14.89</td>
<td>6.38%</td>
<td>.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11, 18, 6, 12, 13,</td>
<td></td>
<td>10, 11, 4, 5, 6,</td>
<td></td>
<td>-0.34</td>
<td>%</td>
<td>stems 34 &amp; 35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23, 24, 14, 15, 16,</td>
<td>18, 23, 12, 13,</td>
<td>24, 27, 14, 15,</td>
<td>28, 29, 16, 17,</td>
<td>30, 31, 19, 20,</td>
<td>32, 33, 21, 22, 26</td>
<td>34, 35, 26</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The table includes descriptive statistics, PCA on residuals, subtest analysis, item equating, and paired t-tests for the SCT analyses. The table entries are structured to show eigenvalues, percentage of variance explained, subtest analysis results, and paired t-tests with their respective critical values. The highest correlation between stems is also indicated.
<table>
<thead>
<tr>
<th>#</th>
<th>Sub-</th>
<th>N</th>
<th>N</th>
<th>Eigen %</th>
<th>Eigen %</th>
<th>Subtest 1 stems</th>
<th>Subtest 2 stems</th>
<th>Original PSI</th>
<th>Subtest PSI</th>
<th>c</th>
<th>p</th>
<th>A</th>
<th>Stem 1 block</th>
<th>Stem 2 block</th>
<th>r</th>
<th>r²</th>
<th>Paired t-test</th>
<th>5%</th>
<th>1%</th>
<th>Highest Corr. Between Stems</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>WUSCT</td>
<td>17</td>
<td>252</td>
<td>2.0</td>
<td>11.7%</td>
<td>2, 3, 4, 5,</td>
<td>10, 11,</td>
<td>.86</td>
<td>.77</td>
<td>.50</td>
<td>.80</td>
<td>.89</td>
<td>2, 3, 4,</td>
<td>10, 11,</td>
<td>.60</td>
<td>.36</td>
<td>t(251)=12.30</td>
<td>5.56%</td>
<td>.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Form 81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6, 7, 8, 9,</td>
<td>12, 13, 14,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5, 6, 7,</td>
<td>12, 13,</td>
<td></td>
<td></td>
<td>p=.24,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>15, 16, 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9, 17</td>
<td>14, 15,</td>
<td></td>
<td></td>
<td>p=.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WUSCT</td>
<td>17</td>
<td>251</td>
<td>1.6</td>
<td>9.4%</td>
<td>1, 2, 3, 4,</td>
<td>6, 7, 8, 9,</td>
<td>.89</td>
<td>.85</td>
<td>.31</td>
<td>.91</td>
<td>.95</td>
<td>1, 2, 3,</td>
<td>6, 7, 8,</td>
<td>.71</td>
<td>.51</td>
<td>t(250)=9.15%</td>
<td>2.39%</td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Form 81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>11, 12, 13,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4, 5,</td>
<td>9, 11,</td>
<td></td>
<td></td>
<td>p=.04,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10, 14</td>
<td>16, 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10, 14</td>
<td>12, 13,</td>
<td></td>
<td></td>
<td>p=.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GLP</td>
<td>29</td>
<td>306</td>
<td>1.8</td>
<td>6.4%</td>
<td>1, 2, 3, 4,</td>
<td>11, 13,</td>
<td>.96</td>
<td>.95</td>
<td>.19</td>
<td>.96</td>
<td>.98</td>
<td>1, 2, 3,</td>
<td>11, 13,</td>
<td>.89</td>
<td>.80</td>
<td>t(305)=9.80%</td>
<td>2.29%</td>
<td>.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5, 6, 8, 9,</td>
<td>14, 15, 16,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4, 5, 6,</td>
<td>14, 15,</td>
<td></td>
<td></td>
<td>p=.25,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10, 12</td>
<td>17, 18, 20,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8, 9,</td>
<td>16, 17,</td>
<td></td>
<td></td>
<td>p=.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19, 22</td>
<td>21, 23, 25,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10, 12,</td>
<td>18, 20,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24, 28</td>
<td>26, 27, 29,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19, 22,</td>
<td>21, 25,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24, 28</td>
<td>26, 27,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**PCA.** A PCA was conducted on the residuals to determine if the remaining variability was random after extracting the Rasch factor. Secondary constructs are identified when the remaining variability forms patterns with theoretical relevance. For some subsamples, the difference between principal components 1 and 2 was greater than the difference between subsequent principal components. For example, for subsample 2, there was a decrease in eigenvalues of 0.54 (2.00 – 1.46) and a decrease in the percentage of variability explained by 3.13% (11.74% - 8.61%). This may suggest that the remaining variability formed secondary constructs in the data after the Rasch factor was extracted, but further analyses were required to determine this. Stems were sorted according to their loadings on principal component 1 to determine whether loadings indicated meaningful secondary factors. There did not appear to be any theoretically relevant pattern to the content of stems from subsamples 1 and 2. However, for subsamples 3, 4, and 5, one group of stems appeared to be somewhat related to ‘understanding others’. In contrast, the other stems seemed to reflect a broad range of considerations relating to the worldview. This may suggest the presence of a secondary factor, but further analyses were required to confirm this.

**Subtest Analysis.** A subtest analysis approximates the amount of multidimensionality in the data and, therefore, the extent to which an assessment deviates from unidimensionality (RUMM Laboratory, 2009b). The procedure for conducting a subtest analysis is described on in Study 1A. The presence of multidimensionality is detected through four indices as described in Study 1A: PSI, c, ρ, and A. Reliability indices from the first estimate did seem to be inflated for some subsamples. For example, the PSI appeared to drop significantly for subsamples 2 and 3 but not for subsamples 1, 4 and 5. There appeared to be a moderate amount of multidimensionality in the data because c ranged from .19 to .50 with a median of .31, which did not appear to approach 0. The correlation between the two higher-order items from the second estimate, ρ, was high and ranged between .80 - .96 with a median score of .91. Finally, the amount of non-error variance between the two higher-order items from the second estimate, A, was high and ranged between .89 - .98 with a median score of .95.
**Item Equating.** An item equating procedure was undertaken as a final way of examining the amount of multidimensionality in the SCT scores. For an assessment to be unidimensional, all stems must tap the same underlying factor, and a different set of stems should award equivalent scores to persons of equal ability. Item equating is a Rasch procedure which identifies whether different sets of stems give persons of equal ability equivalent scores. Two identically sized sets of stems were created based on the stem loadings on principal component 1. A paired sample t-test demonstrated that the scores on the first set of stems were not significantly different from those on the second set of stems. However, the correlations between the two sets of scores ranged from moderate to high (i.e., \(0.60 \leq r \leq 0.89\)). Person ability was better reflected in two scores on the SCT for a significant percentage of persons. This effect was particularly pronounced for subsamples 2, 3, 4, and 5. For example, ability is better reflected by two separate scores for 15% of test-takers for subsample 2. Considered collectively, analyses suggest that some subsamples display a moderate amount of multidimensionality. This suggests that the SCT is an impure measure of ego development and is likely to tap secondary constructs such as understanding others and/or random noise.

**Step 3b: Invariant Sequence.** This analysis aimed to determine if stem thresholds and, therefore, stages occurred in the sequence predicted by theory. In the context of developmental data, thresholds take on a particular meaning because they create the partitions between successive stages e.g., where stage S terminates and where stage S+1 begins. Correct sequencing of stem thresholds was established in Step 1c. In this step, significant data cleaning was undertaken, and the decision was made to reduce the dataset from nine stages and 55 stems to seven stages and 42 stems. Thresholds were only correctly sequenced after stages were combined and 24% of stems were deleted (n=13), so no further analyses were undertaken.

**Step 3c: Qualitatively Distinct.** This analysis aimed to determine if stem thresholds were sufficiently spaced apart to create partitions between qualitatively distinct stages. The person item-threshold distribution shows the logit scale on the horizontal axis with person ability plotted in the
upper section and threshold estimates in the lower section. As can be seen from Figure 24, thresholds did not have a consistent location, and this failed to create visible ‘white space’ between successive thresholds. As such, the location of stages could not be clearly identified along the logit scale. For example, a person with an ability of -0.13 would likely receive a score of Individualist on stem 2, Achiever on stem 25, and Expert on stem 26.

**Figure 24**

*Person-Item Threshold Distribution for the SCT*

Descriptive statistics for thresholds are presented in Table 26.
Table 26

**Descriptive Statistics for Uncentralised Thresholds Between Ego Development Stages**

<table>
<thead>
<tr>
<th>Threshold $\tau_n$</th>
<th>Partition between stages</th>
<th>M</th>
<th>SD</th>
<th>MIN</th>
<th>MAX</th>
<th>Range</th>
<th>M of SEs</th>
<th>Separation between adjacent thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Impulsive &amp; Opportunist /Diplomat</td>
<td>-3.28</td>
<td>0.68</td>
<td>-5.16</td>
<td>-2.14</td>
<td>3.01</td>
<td>0.12</td>
<td>-1.02</td>
</tr>
<tr>
<td>2</td>
<td>Diplomat/Expert</td>
<td>-2.00</td>
<td>0.45</td>
<td>-3.16</td>
<td>-1.16</td>
<td>2.00</td>
<td>0.08</td>
<td>-0.66</td>
</tr>
<tr>
<td>3</td>
<td>Expert/Achiever</td>
<td>-0.76</td>
<td>0.41</td>
<td>-1.82</td>
<td>-0.07</td>
<td>1.76</td>
<td>0.07</td>
<td>-0.18</td>
</tr>
<tr>
<td>4</td>
<td>Achiever/Individualist</td>
<td>0.53</td>
<td>0.36</td>
<td>-0.25</td>
<td>1.44</td>
<td>1.69</td>
<td>0.08</td>
<td>-0.20</td>
</tr>
<tr>
<td>5</td>
<td>Individualist/Strategist</td>
<td>1.95</td>
<td>0.40</td>
<td>1.24</td>
<td>2.67</td>
<td>1.43</td>
<td>0.14</td>
<td>-0.96</td>
</tr>
<tr>
<td>6</td>
<td>Strategist &amp; Alchemist/Ironist</td>
<td>3.56</td>
<td>0.99</td>
<td>1.71</td>
<td>7.54</td>
<td>5.83</td>
<td>0.40</td>
<td>-</td>
</tr>
</tbody>
</table>

Statistical analyses confirmed the findings from the visual analysis. For all thresholds, estimates were broadly distributed over 1.43 to 5.83 logits. For all thresholds, several estimates for threshold $\tau_n$ were higher than the lowest estimate for threshold $\tau_{n+1}$. This explains the absence of white space between successive thresholds and the inability to identify the location of stages along the logit scale.

**Reliability and Distinct Strata of Performance.** The rationale for using reliability indices to determine the number of statistically distinct strata of performance is provided in Study 1A. In the context of developmental data, strata of performance operationalise distinct stages. To distinguish between seven statistically distinct stages, the SCT must exhibit a reliability of at least .96 (Fisher, 1992; Wright, 1996; Wright & Masters, 2002). The SCT displayed a PSI of .93, and coefficient alpha could not be calculated because of the linked data structure. However, an assessment with this level of reliability can only distinguish between four or five stages. As shown in the unidimensionality...
analyses, some reliability estimates were likely to be inflated and the PSIs from the subtest analyses ranged from .77 to .95. This suggests that the ego development scoring system may only distinguish between as few as three stages in some subsamples.

Considered collectively, these analyses suggest that thresholds do not create clear partitions between stages, and that the location of stages cannot be identified along the logit scale in terms of the amount of development they reflect. Finally, there is an inconsistency between the number of stages the scoring system claims to measure and the actual number of stages that can be distinguished.

**Step 3d: Structure of the Whole.** This analysis aimed to determine whether a single stage organised each leader’s responses to the stems. Category probability curves show the logit scale on the horizontal axis and the probability of scoring in each stage on the vertical axis. For example, a person with an ability of 0 on the logit scale would receive a score of Achiever (refer to Figure 25).

Two patterns emerge if the requirement of structured wholeness is satisfied. First, each curve should reach an apex of > .80. Second, each curve should only converge with adjacent curves, and convergence between nonadjacent curves should occur with a probability of < .10. Category probably curves for all 42 stems were analysed. Stem 43 had the most broadly spaced thresholds, while stem 8 had the most narrowly spaced thresholds. They are presented in Figures 25 and 26, respectively.
For stem 43, the maximum probability occurs for the Achiever stage at .57. For stem 43, the probability of convergence between Achiever and the two adjacent stages, Expert, and Individualist, was .43. The probability of convergence with two nonadjacent stages, Diplomat and Strategist, was
.21. For stem 8, the maximum probability occurs for the Strategist stage at .55. Using Achiever as an illustrative example, the probability of convergence between Achiever and the two adjacent stages, Expert, and individualist, was .34 - .35. The probability of convergence with two nonadjacent stages, Diplomat and Strategist, was .24. Considered collectively, stages cannot clearly be considered structured wholes because a single stage does not organise responses to the stems. Considered collectively, the evidence suggests that scores may not satisfy the requirement of structured wholeness.

**Step 3d: Hierarchical Integrations.** This analysis aimed to determine if stage scores exhibit spurts and plateaus of development. Item equating is a Rasch procedure which examines the relation between person ability and total score on an assessment. All 42 stems were entered simultaneously into an item equating procedure. The item equating graph shows person ability on the horizontal axis and total score on the SCT to reveal the ‘shape of development’ (refer to Figure 27).

**Figure 27**

Item Equating Graph for all 42 Stems
As can be seen from Figure 27, item equating graphs fail to exhibit spurts and plateaus. The graph suggests a monotonic pattern of growth rather than stage-like discontinuities observed in scores based on hard stage models. There are no statistical procedures to test the shape of development, so the possibility of micro plateaus that cannot be detected through visual analysis cannot be ruled out.

Tests for additional general measurement properties unrelated to hard stage requirements are reported in Appendix J.

**Evaluation of Requirements for Study 1B**

- **Criterion 1: Unidimensionality.** The family approach suggested that scores contained a moderate amount of multidimensionality for at least some subsamples. More specifically, the PCA suggested the presence of additional constructs after the Rasch factor was extracted. There PSI appeared to be inflated for at least two subsamples, which suggested the presence of multidimensionality. The magnitude of multidimensionality, $c$, ranged between .19 and .50 with a median of .31 which may be interpreted to suggest a moderate amount of multidimensionality. There was a relatively high correlation between the two higher-order items, $\rho$. The true common variance between the two higher-order items, $A$, was relatively high. However, the item equating analysis suggested that performance on the SCT is better reflected by more than one score for a statistically significant proportion of test-takers.

- **Criterion 2: Invariant sequence.** This was partially supported. The initial dataset consisted of nine stages and 55 stems but exhibited threshold disorder. Thresholds were correctly sequenced after nine stages were collapsed to seven stages, and 24% of stems ($n=13$) were deleted. This suggests that a more parsimonious model may exhibit correct threshold sequencing.
• Criterion 3: Qualitatively distinct.
  o Criterion 3a) This was not supported because there was a consistent pattern across six thresholds such that the highest estimate for threshold $\tau_n > \text{the lower estimate for threshold } \tau_{n+1}$. Thresholds did not have a consistent location which meant that the location of stages could not be identified. As a result, stages may not be considered qualitatively distinct.
  o Criterion 3b) This was not supported because the PSI was .93, which suggested that the ego development scoring system can only distinguish between approximately four or five stages. Given the inflated PSIs identified in the subtest analysis, the scoring system may only distinguish between three stages in some subsamples.

• Criterion 4: Structure of the whole.
  o Criterion 4a) This was not supported because category probability curves reached an apex < .80, which suggested that a single stage did not organise most responses to the stems.
  o Criterion 4b) This was not supported because convergence between nonadjacent curves occurred with a probability > .10, which suggested that reasoning was broadly distributed across multiple stages.

• Criterion 5: Hierarchical integrations. This was not supported because the item equating graph suggested a monotonic pattern of growth rather than spurts and plateaus. However, the possibility of micro plateaus that cannot be observed through visual analysis cannot be ruled out.

Discussion

Few criteria were supported. SCT scores awarded by the ego development scoring system were shown to satisfy the requirement of invariant sequence but not the requirements of unidimensionality, qualitative distinctness, structured wholeness, and hierarchical integrations. This
pattern of results is inconsistent with the claims made by ego developmentalists but relatively consistent with the evidence discussed in the literature review. As discussed in the literature review, previous empirical studies provide strong support for the requirement of invariant sequence but not necessarily for other requirements. As acknowledged earlier, this led to the unique challenge of testing criteria based on ego developmentalists’ espousals without supporting evidence.

Several explanations may account for these findings. The first explanation is that these findings are an artefact of the dataset and the linked design which included three scorers, five versions of the SCT, and a multitude of stems which would not typically be included in other samples. This explanation does not seem tenable, considering the evidence presented. Step 1a confirmed that the three scorers were equally severe in their scoring. Step 1b confirmed that different versions of the SCT awarded equivalent scores to persons of equal ability. Rasch analyses on subsamples revealed similar patterns to those obtained for the whole sample. For example, the person-item threshold indicated stages could not be located along the logit scale for any version of the SCT. An illustrative example is presented in Appendix K.

The second explanation is that findings are an artefact of how hard stage requirements were operationalised and the Rasch procedures employed. This explanation may be more tenable, particularly considering the ambiguous relation that ego development has with the requirement of structured wholeness. As discussed in the literature review, ego developmentalists simultaneously assert that the worldview has a centre of gravity and that protocols consist of broadly distributed stem scores. But how can the centre of gravity be so powerful as to organise the entirety of one’s worldview when the findings reported in Study 1B suggest that the centre of gravity only organises approximately 18 out of 36 stem completions? Holding these positions simultaneously may constitute a contradiction between theory and measurement. Given the broad distribution of SCT stem scores, hard stage requirements, as operationalised in Study 1B, would be nearly impossible to satisfy. First, broad response distributions are more likely to contain stems that are only weakly or moderately correlated and, therefore, unlikely to yield a unidimensional construct in data reduction.
procedures. Second, broad distributions likely display considerable overlap and rarely satisfy qualitative distinctness. Third, broad distributions suggest that a single stage does not organise most responses; therefore, distributions will rarely satisfy structured wholeness. Finally, it would be challenging for scores to satisfy the requirement of hierarchical integrations because broad distributions may suggest that responses have not been hierarchically integrated at a single stage.

The third explanation is that SCT scores operationalise a soft stage model. As discussed earlier in Chapter 2, this thesis prioritises two considerations when distinguishing hard stages from soft stages and cultural ages: a) the logic underpinning the sequence of stages, and b) the behaviour of scores. In relation to a), the literature review for Study 1B suggested that the logic underpinning the sequence of stages reflected an accumulation and deepening of perspectives. This raised a suspicion that ego development theory defines stages in a manner that more closely approximates the definitional attributes of a soft stage model. In relation to b), SCT scores failed to satisfy hard stage requirements, except for the requirement of invariant sequence. The general pattern of results resembles those typically found in cumulative ordinal scales for which there is often variability in the location of ordinal categories, broadly distributed category probability curves, and monotonic patterns of growth (Bond & Fox, 2015; Hagquist et al., 2009; Klein et al., 2022; Pallant & Tennant, 2007; Steele & Day, 2020). If this third explanation were tenable, it would need to be supported by the weight of additional evidence and account for why ego developmentalists make hard stage claims. However, it would be consistent with Kohlberg’s (1984) claim that ego developmental stages are nontransformational and merely reflect an additive increase in elements.

Should ego development fail to satisfy hard stage requirements, this may impact how the theory and the SCT are applied in leadership science (Torbert, 2013) and practice (Petrie, 2014). Given the prolific use of ego development, this warrants further investigation. Chapter 3 determines whether scores might be shown to satisfy hard stage requirements in other samples and via alternate research methods that circumvent the need to account for broad response distributions.
Irrespective of hard stage requirements, there appear to be several opportunities to improve the measurement properties of the SCT. Approximately 24% of stems needed to be deleted, which suggests that some stems are inconsistent with the functioning of other stems. A content analysis of the deleted stems is required to identify which stems are less likely to elicit responses related to the targeted construct. Given that unidimensionality is compromised, some stems may tap a construct other than ego development and/or there is considerable noise in the data. The scale’s dimensionality seems a relatively pressing issue because the SCT awards a single TPR to test-takers. This score should reflect a unidimensional construct that accounts for most variability in scores. If the ego operates across a range of stages as a function of décalage (i.e., the tendency for people to perform at different stages in different domains), it may be more sensible to award separate scores to reflect test-takers’ development in different facets of life. The inability to locate stages along the logit scale is problematic because it suggests that stages reflect different amounts of development for different stems. This implies that ego development stages have different meanings in different contexts e.g., the Expert stage reflects a lower amount of development when responding to a stem about education, but the Expert stage reflects a higher amount of development when responding to a stem about family. The ego development scoring system may need to be revisited to ensure that SCT stem scores are awarded more consistently. Finally, the ego development scoring system may only distinguish between half as many stages as it claims to measure. This is important to resolve because some pairwise comparisons may not yield meaningful distinctions between Stage S and Stage S+1. Broader implications, limitations, and future research directions are discussed in Chapter 6.
Chapter 3: Additional Analyses Undertaken on Ego Development

Chapter 2 concluded by questioning the hard stage claims made about ego development given that they were not clearly supported by earlier research, or the findings presented in Study 1B. It was also acknowledged that SCT stem scores are broadly distributed, which may prevent Rasch analyses from identifying patterns that would satisfy hard stage requirements. Given the prolific use of ego development, this warrants further investigation. The purpose of Chapter 3 was to undertake three relatively simple studies based on different samples and research methods. Studies 2B and 2C circumvented the need to account for broad distributions in protocols. The goal was to find an explanation for the findings reported in Study 1B.
Study 2A: Reanalysis of Earlier Ego Development Data

Cook-Greuter’s Argument About Hard Stage Requirements

As mentioned in Chapter 2, Cook-Greuter’s (1999) unpublished dissertation is one of the few research projects that attempted to directly address hard stage requirements. The purpose of Study 2A was to reanalyse Cook-Greuter’s original data using more contemporary methods, to determine whether her conclusions are justified.

With reference to Figure 28, which is reproduced from Cook-Greuter’s original data, her arguments in support of hard stage requirements are quoted to ensure they are represented accurately.

The following numbers and the corresponding graph of the distribution curves for each ego stage (type of TPR) shows how, with increasing ego stage, the percentage of responses at a given stage shifts to the right. With each shift, fewer responses occur at the lower levels and more at increasingly higher ones in a clearly clustered pattern.

The data make vivid the wide range of completions seen in many high-stage protocols. When the distributions for each ego stage type are aggregated as they are here, the effect is even more noticeable. The widest spread occurred at the [Strategist] stage where the twelve protocols had completions ranging from [Impulsive to Unitive]. The narrow spread for [Expert] could be an artefact of the N=1, but it also represents a general tendency of the conventional personality to adhere to the narrow parameters of conventionality. Not unusually, one encounters a protocol that is entirely located at the conventional tier.

As can be seen in [Figure 28], the overall distributions are consistent with predictions as well as the gradual shift of the center of the distributions toward the right with increasing TPR. This reflects the shifting center of meaning making at successively higher ego stages. The graph shows that the distribution curves for the different personality types (ego stages) are clearly discontinuous and distinct from each other as each type of TPR peaks in a different area of the distribution space. The prediction that the distribution patterns for different ego stages are distinct is supported by the evidence.

For a stage theory to be hierarchical, the stages have to be distinct and sequential. [Strategist, Construct-Aware, Unitive] follow each other in a sequential order and are discontinuous according to the percentage frequency distributions for the different groups of TPRs. Therefore, the Construct-aware and the Unitive stage seem to fulfil the stage requirements for being higher-order hierarchical integrations (Cook-Greuter, 1999, pp. 159-160).
Figure 28

Distribution by TPR: Percentage of Stems Scored at Successive Ego Development Stages

Note. This graph is based on data reported in “Postautonomous Ego Development: A Study of Its Nature and Measurement” by S. R. Cook-Greuter, 1999, p. 160. Copyright 1999 by Integral Publishers. The blue line represents protocols given a TPR of Expert. The red line represents protocols given a TPR of Achiever. The yellow line represents protocols given a TPR of Individualist. The green line represents protocols given a TPR of Strategist. The orange line represents protocols given a TPR of Construct-Aware. The teal line represents protocols given a TPR of Unitive.

Cook-Greuter’s (1999) argument depends upon a visual analysis of her data and graph. As a result, her original dataset was reanalysed to determine whether her conclusions are supported. While Cook-Greuter did employ mathematical procedures to examine other aspects of the scoring system e.g., cut-scores, these procedures were not aimed at examining hard stage requirements.

The term ‘criteria’ was used in Studies 1A and 1B to acknowledge that not all criteria could be evaluated via formal statistical tests and to recognise the family approach typically used in Rasch analysis. In the context of Study 2, the term hypothesis is used because formal statistical tests are
applied. Rasch analyses could not be undertaken in Study 2A because data were only available at an aggregated level, but not for individual persons or stems. For several reasons, statistical analyses undertaken in Study 2A were not considered appropriate for Studies 1A and 1B. In the context of Studies 1A and 1B, the Rasch Model enabled more stringent analyses to be performed, particularly on hard stage requirements.

**Hypotheses for Study 2A**

The behaviour of scores that do and do not satisfy hard stage requirements is outlined in Figures 6 and 7. Given this, the following hypotheses were tested:

- **Hypothesis 1 (invariant sequence):** There is a statistically significant difference in mean scores between successive TPR distributions.

- **Hypothesis 2 (qualitatively distinct):** There is a statistically significant difference between the location of successive TPR distributions.

- **Hypothesis 3 (structure of the whole):**
  - **Hypothesis 3a)** Protocols have a range of no more than three stages, i.e., modal stage plus two adjacent stages.
  - **Hypothesis 3b)** Protocols have a mode which accounts for >80% of the response distribution.

The data structure did not allow statistical analyses to be performed on unidimensionality and hierarchical integrations because Cook-Greuter (1999) only reported frequency data but not scores for individual persons or stems.
Methods

Sample

Cook-Greuter’s (1999, p. 160) original dataset is presented in Table 27. The dataset is in the public domain and was originally presented as a table of frequency distributions. As a result, the data in Table 27 is not reconstructed from another set of reported parameters.

Table 27

Frequency Distribution by TPR: Percentage of Stems Scored at Successive Ego Development Stages

<table>
<thead>
<tr>
<th>TPR</th>
<th>N protocols</th>
<th>N stems</th>
<th>Percentage of Stems Scored at Successive Ego Development Stages</th>
<th>95% C.I. around Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Opportun&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Diplomat</td>
</tr>
<tr>
<td>Unitive</td>
<td>8</td>
<td>288</td>
<td>-</td>
<td>1.0</td>
</tr>
<tr>
<td>Construct-Aware</td>
<td>15</td>
<td>540</td>
<td>0.2</td>
<td>1.7</td>
</tr>
<tr>
<td>Strategist</td>
<td>12</td>
<td>432</td>
<td>0.7</td>
<td>4.6</td>
</tr>
<tr>
<td>Individualist</td>
<td>13</td>
<td>468</td>
<td>0.6</td>
<td>3.4</td>
</tr>
<tr>
<td>Achiever</td>
<td>11</td>
<td>396</td>
<td>1.5</td>
<td>8.8</td>
</tr>
<tr>
<td>Expert</td>
<td>1</td>
<td>36</td>
<td>-</td>
<td>27.8</td>
</tr>
<tr>
<td>Totals</td>
<td>60</td>
<td>2160</td>
<td>0.6%</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

Notes. A dash (-) means no data were collected for this cell. *Refers to the TPR given to protocols. **The total number of stems for protocols given each TPR. For example, 15 protocols were given a TPR of Construct-Aware and given that SCTs consist of 36 stems, there were 15 x 36 = 540 stems. *The percentage of stems scored at successive stages for protocols organised by TPR. For example, of the 15 protocols scored at Construct-Aware, 1.7% of
the 540 stems were scored at the Diplomat stage. The 95% confidence interval around the mode was calculated by the author of this thesis and was not included in Cook-Greuter’s (1999) original data table.

The dataset consisted of 60 protocols which were given a TPR from Expert to Unitive, and 2160 stems (i.e., 60 protocols x 36 stems=2160). Detailed sampling procedures are described in Cook-Greuter (1999, pp. 142-144). The sample consisted of 31 women and 29 men aged between 16 and 79. Most were well educated and had earned at least a Masters degree. Various occupations were represented, including students, artists, and counsellors.

**Materials**

Graphs and statistical analyses were undertaken in Jamovi (2021).

**Procedure**

Stages were coded along an ordinal scale from 1 to 9 such that: Impulsive=1, Opportunist=2, Diplomat=3, Expert=4, Achiever=5, Individualist=6, Strategist=7, Construct-Aware=8 and Unitive=9. Given the simple nature of the dataset, only basic statistical analyses were undertaken. SCT stem scores and TPRs awarded to protocols formed the basis for analysis.

**Results**

To facilitate visual analysis, a boxplot of TPR distributions is presented in Figure 29.
Notes. As seen in Figure 29, it is possible for the median score, which is identical to the second quartile, to coincide with the first or third quartile. This occurs when a single score is repeated multiple times in a skewed distribution, and the same score occurs at the first and second quartile, and/or the second and third quartile simultaneously. This effect can be observed for the Achiever and Individualist stages.

Visual analysis does not clearly support Cook-Greuter’s conclusions. The boxplot suggests that the distributions had similar ranges, which began in Opportunist and ended in either Construct-Aware or Unitive. Most distributions had a similar shape, and it was ambiguous whether the median scores between distributions were distinctly different. For example, Achiever and Individualist distributions share the same median score, as did Construct-Aware and Unitive. Finally, the interquartile ranges appeared to overlap which did not clearly suggest that the distributions were distinctly different. Statistical analyses were undertaken to verify these observations, and descriptive statistics are provided in Table 28.
Table 28

Descriptive statistics for TPR Distributions (Total N of Stems = 2160)

<table>
<thead>
<tr>
<th>TPR</th>
<th>Ordinal Code</th>
<th>N stems</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>SD</th>
<th>SE</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>95% C.I. Min</th>
<th>95% C.I. Max</th>
<th>C. I. Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>4</td>
<td>36</td>
<td>4.06</td>
<td>4</td>
<td>4</td>
<td>0.89</td>
<td>0.15</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>2.71 to 3.29</td>
<td>5.71 to 6.29</td>
<td>2.42</td>
</tr>
<tr>
<td>Achiever</td>
<td>5</td>
<td>396</td>
<td>4.65</td>
<td>5</td>
<td>5</td>
<td>0.95</td>
<td>0.05</td>
<td>2</td>
<td>8</td>
<td>7</td>
<td>1.91 to 2.09</td>
<td>7.91 to 8.09</td>
<td>5.81</td>
</tr>
<tr>
<td>Individualist</td>
<td>6</td>
<td>468</td>
<td>5.22</td>
<td>5</td>
<td>5</td>
<td>1.01</td>
<td>0.05</td>
<td>2</td>
<td>8</td>
<td>7</td>
<td>1.19 to 2.09</td>
<td>7.91 to 8.09</td>
<td>5.82</td>
</tr>
<tr>
<td>Strategist</td>
<td>7</td>
<td>432</td>
<td>5.71</td>
<td>6</td>
<td>6</td>
<td>1.33</td>
<td>0.06</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>1.87 to 2.13</td>
<td>8.87 to 9.13</td>
<td>6.75</td>
</tr>
<tr>
<td>Construct-Aware</td>
<td>8</td>
<td>540</td>
<td>6.60</td>
<td>7</td>
<td>7</td>
<td>1.32</td>
<td>0.06</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>1.89 to 2.11</td>
<td>8.89 to 9.11</td>
<td>6.78</td>
</tr>
<tr>
<td>Unitive</td>
<td>9</td>
<td>288</td>
<td>7.19</td>
<td>7</td>
<td>8</td>
<td>1.37</td>
<td>0.08</td>
<td>3</td>
<td>9</td>
<td>7</td>
<td>2.84 to 3.16</td>
<td>8.84 to 9.16</td>
<td>5.68</td>
</tr>
</tbody>
</table>

Note. *This column presents the distance between the upper boundary for the minimum C.I. and the lower boundary for the maximum C.I. e.g., For Expert, 5.71 – 3.29 = 2.42.

Invariant Sequence

A one-way ANOVA was conducted to test for significant differences in mean scores between successive TPR distributions. The results of the ANOVA were significant, F(5, 330.6) = 251.0, p < .001. Levene’s test for equality of variances was significant which indicated a significant difference between group variances, F(5, 2154) = 22.82, p < .001. As a result, the Games-Howell Test was used to conduct post hoc pairwise comparisons (refer to Table 29).
Table 29

Means, Standard Deviations, and One-Way Analysis of Variance (ANOVA) on Mean Scores Between Successive TPRs

<table>
<thead>
<tr>
<th>TPR</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Expert</th>
<th>Achiever</th>
<th>Individualist</th>
<th>Strategist</th>
<th>Construct-Aware</th>
<th>Unitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>36</td>
<td>4.06</td>
<td>0.89</td>
<td>-</td>
<td>-0.59**</td>
<td>-1.16***</td>
<td>-1.66***</td>
<td>-2.54***</td>
<td>-3.13***</td>
</tr>
<tr>
<td>Achiever</td>
<td>396</td>
<td>4.65</td>
<td>0.95</td>
<td>-</td>
<td>-0.57***</td>
<td>-1.06***</td>
<td>-1.95***</td>
<td>-2.54***</td>
<td></td>
</tr>
<tr>
<td>Individualist</td>
<td>468</td>
<td>5.22</td>
<td>1.01</td>
<td>-</td>
<td>-</td>
<td>-0.49***</td>
<td>-1.38***</td>
<td>-1.97***</td>
<td></td>
</tr>
<tr>
<td>Strategist</td>
<td>432</td>
<td>5.71</td>
<td>1.33</td>
<td>-</td>
<td>-0.89***</td>
<td>-1.48***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construct-Aware</td>
<td>540</td>
<td>6.60</td>
<td>1.32</td>
<td>-</td>
<td>-</td>
<td>-0.59***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unitive</td>
<td>288</td>
<td>7.19</td>
<td>1.37</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. * p<.05, ** p<.01, *** p <.001

Significant differences were identified for all pairwise comparisons. The means for Expert (M=4.06) and Achiever (M=4.65) were found to be significantly different by 0.59 of a stage, but both means were located within the Expert stage i.e., 4.00 to 4.99. Similarly, the means for Individualist (M=5.22) and Strategist (M=5.71) were significantly different by 0.49 of a stage, but both means were located within the Achiever stage i.e., 5.00 to 5.99. While significant differences between means were found, the results must be interpreted with caution from a theoretical perspective as discussed below.

Qualitatively Distinct

A series of two-sample Kolmogorov-Smirnov Tests were conducted to identify significant differences between the location of successive TPR distributions. That is, whether the distribution of protocols with a TPR of Stage S was significantly different to the distribution of protocols with a TPR of Stage S+1. A two-sample Kolmogorov-Smirnov Test is a nonparametric test which determines
whether two samples are likely to come from the same distribution and was therefore used as a statistical means to operationalise the requirement of qualitative distinctness. A Bonferroni correction yielded a critical alpha of .01 (i.e., .05/5 = .01). Pairwise comparisons demonstrated that D-statistics consistently exceeded the D-critical values, which suggested that there were significant differences between successive TPR distributions. Results suggested that there was a significant difference between Expert versus Achiever (D-stat=.34, D-critical=.28, p=.01), Achiever versus Individualist (D-stat=.22, D-critical=.11, p=.01), Individualist versus Strategist (D-stat=.20, D-critical=.11, p=.01), Strategist versus Construct-Aware (D-stat=.28, D-critical=.10, p=.01), and also Construct-Aware versus Unitive TPRs (D-stat=.20, D-critical=.12, p=.01).

Despite these differences, visual analysis of Figures 28 and 29 suggests a considerable amount of overlap between TPR distributions. The Coefficient of Overlap (OVL) was calculated to identify the amount of overlap. OVL estimates the area shared by two distributions and can be used to make inferences about their relatedness. OVL ranges from 0 to 1, where zero indicates no overlap and one indicates an identity relation between two distributions. OVL is given by

\[ OVL = 2 \Phi(-|\delta|/2) \]

where \( \Phi \) is the cumulative distribution function of the standard normal distribution and \( \delta \) is the population Cohen’s \( d \) (Magnusson, 2021; Reiser & Faraggi, 1999). Results are presented in Table 30.
Table 30

OVL Between Successive TPR Distributions

<table>
<thead>
<tr>
<th>TPR</th>
<th>M</th>
<th>SD</th>
<th>Expert</th>
<th>Achiever</th>
<th>Individualist</th>
<th>Strategist</th>
<th>Construct-Aware</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>4.06</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achiever</td>
<td>4.65</td>
<td>0.95</td>
<td></td>
<td>74.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individualist</td>
<td>5.22</td>
<td>1.01</td>
<td></td>
<td></td>
<td>77.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategist</td>
<td>5.71</td>
<td>1.33</td>
<td></td>
<td></td>
<td></td>
<td>83.6%</td>
<td></td>
</tr>
<tr>
<td>Construct-Aware</td>
<td>6.60</td>
<td>1.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>73.6%</td>
</tr>
<tr>
<td>Unitive</td>
<td>7.19</td>
<td>1.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82.6%</td>
</tr>
</tbody>
</table>

All TPR distributions overlapped by 74% to 84%. The empirical literature does not provide a criterion against which to compare these percentages. However, visual inspection of the Probability Characteristic Curves presented in Study 1A suggest an uncharacteristically high amount of overlap between distributions, relative to scores that have been shown to satisfy hard stage requirements.

Structure of the Whole

The descriptive statistics provided in Table 28 show that protocols had a range of seven or eight stages except for Expert, which had a range of four. All ranges exceed the critical value of 3 i.e., modal stage plus two adjacent stages. Ninety-five percent confidence intervals were constructed around the min and max values, based on 1.96 x SE. As can be seen from Table 28, the distance between the upper boundary for the min value and the lower boundary for the max value exceeded three in all cases, except for Expert.

The frequency distributions presented in Table 27 demonstrate that the modal stages accounted for 27.4% to 47.2% of the response distribution. For example, for protocols with a TPR of Achiever, only 45% of stems were scored at Achiever and 55% of responses were scored at other
stages. Ninety-five percent confidence intervals were constructed around the modal percentages, based on 1.96 x SE. As can be seen from Table 27, none of the confidence intervals included 80%.

Hierarchical Integrations

While the dataset did not allow statistical tests to be conducted on the requirement of hierarchical integrations, the logic of Cook-Greuter’s (1999) arguments was analysed. Cook-Greuter makes two key claims about the hierarchically integrated nature of her stages. Claim 1: “For a stage theory to be hierarchical, the stages have to be distinct and sequential...” (Cook-Greuter, 1999, p. 160). Arguably, qualitative distinctness and sequentiality are the products of hierarchical integrations. As a result, stages must be distinct and sequential if they are hierarchically integrated. This is a truism and a statement of logical necessity. However, this does not establish whether the requirement is satisfied.

Cook-Greuter (1999, p. 160) goes on to make Claim 2: “[Strategist], [Construct-Aware] and [Unitive] follow each other in a sequential order and are discontinuous according to the percentage frequency distributions for the different groups of TPRs. Therefore, the Construct-Aware and Unitive stage seem to fulfil the stage requirements for being higher-order hierarchical integrations” Cook-Greuter (1999, p. 160). The syllogism underpinning her argument is reflected in the following form:

- Premise 1: If stages are sequential and discontinuous, then they are hierarchically integrated.
- Premise 2: Stages are sequential and discontinuous.
- Conclusion: Therefore, stages are hierarchically integrated.

Her argument is valid but not sound. Her argument is valid in that the conclusion logically follows from the premises via a standard application of modus ponens i.e., given if A therefore B, and A, we may deduce B. However, her argument is not sound because the consequent does not follow from the antecedent in Premise 1. This renders Premise 1 a non sequitur, which is a logical fallacy. Sequentiality and discontinuity are definitional properties of all ordinal scales and reveal nothing
about hierarchical integrations. Similar patterns have been demonstrated in a multitude of cumulative ordinal scales (Bond & Fox, 2015; Hagquist et al., 2009; Klein et al., 2022; Pallant & Tennant, 2007; Steele & Day, 2020).

Hypothesis Tests for Study 2A

- Hypothesis 1 (invariant sequence): This was supported because there was a significant difference in mean scores between successive TPR distributions.
- Hypothesis 2 (qualitatively distinct): This was partially supported. Two-sample Kolmogorov-Smirnov Tests demonstrated there was a significant difference between the location of successive TPR distributions. However, OVLs occurred between 74-84% which may reflect an uncharacteristically high amount of overlap between distributions, relative to scores that have been shown to satisfy hard stage requirements.
- Hypothesis 3 (structure of the whole):
  - Hypothesis 3a) This was not supported because protocols had ranges of seven or eight stages which exceeded the critical value of three stages. This suggested that stem completions do not consolidate at a single stage.
  - Hypothesis H3b) This was not supported because protocols have a mode which accounts for considerably less than 80% of the response distribution.

Discussion

Few hypotheses were supported. SCT scores were shown to satisfy the requirement of invariant sequence. However, pairwise comparisons demonstrated that the means for Expert and Achiever occurred in the Expert stage, and the means for Individualist and Strategist occurred in the Achiever stage. This finding might be used to collapse pairs of stages in the context of many psychometric analyses. Scores were shown to partially satisfy the requirements of qualitative distinctness but not satisfy the requirement of structured wholeness. Requirements of
unidimensionality and hierarchical integrations could not be explicitly tested. However, the argument for hierarchical integrations appeared to rest on a logical fallacy.

Cook-Greuter’s interpretation of her dataset is strikingly different from that presented here. Three explanations may account for these findings. First, her interpretation may be a function of her visual analysis, which did not consider statistical evidence. Second, these findings corroborate those presented in Study 1B, which tentatively suggested that SCT scores may operationalise a soft stage, rather than a hard stage, model. This is an important finding because the two studies used different datasets and analytical procedures, thereby reducing the likelihood that the findings were an artefact of research methods. Third, the findings may be an artefact of the broad response distributions. As discussed in Study 1B, this makes it almost impossible for hard stage requirements to be satisfied in how they have been operationalised hitherto. To address this objection directly, Studies 2B and 2C test for hard stage requirements in a manner that circumvents this broad response distribution.
Study 2B: Scoring Ego Development Scoring Exemplars with a Cognitive-Developmental Scoring System

To address objections to Studies 1B and 2A, Studies 2B and 2C tested hard stage requirements in a manner that circumvented the broad response distribution typically found in SCT protocols. Given the provisional evidence in support of Lectical Phase Scores presented in Study 1A, they were used as a criterion variable against which to evaluate the ego development scoring system.

Correspondence Between Ego Development Stages and Cognitive-Developmental Stages

As discussed in Chapter 2, ego developmentalists argue that their stages satisfy hard stage requirements partially because they share a structural isomorphism with cognitive-developmental stages. Ego development and cognitive-developmental stages are contended to be different with respect to their breadth and content, but to be similar with respect to their hierarchically integrated structure. For example, Loevinger argued for a correspondence between her stages, Piaget’s stages of moral reasoning (Loevinger, 1976, pp. 79-85) and Kohlberg’s stages of moral reasoning (Loevinger, 1976, pp. 118-122). Torbert argued for a correspondence between his stages, Kohlberg’s stages of moral reasoning (Fisher et al., 1987, p. 260; Lichtenstein et al., 1995, p. 98), the Model of Hierarchical Complexity (Torbert, 1994, pp. 185-189), Kegan’s constructive-developmental stages (Fisher et al., 1987, p. 260; Lichtenstein et al., 1995, p. 98; Torbert, 1994, p. 184) and Selman’s stages of socio-moral perspective taking (Fisher et al., 1987, p. 260). Cook-Greuter argued for a correspondence between her stages and the Model of Hierarchical Complexity (Cook-Greuter, 1990, p. 104; 2002, pp. 7-28; 2013, pp. 22-76). While some developmentalists do not explicitly distinguish between claims at the scale of theory versus measurement, most claims about correspondence between stages seem to be made at the scale of theory. However, the research methods used to establish these correspondences are unclear. Many arguments depend on comparing superficial
features of between stages and assuming, rather than demonstrating, correspondences. Several of the cognitive-developmental stages to which ego development stages have been mapped, have been shown to satisfy hard stage requirements at the scale of measurement (Commons et al., 2008; Dawson, 2000a, 2002b; Dawson et al., 2010).

Torbert (1995) and Cook-Greuter (1990, 2002, 2013) posit correspondences, but map them somewhat differently, as presented in Table 31. The stages of moral reasoning to which Torbert maps his stages (Lichtenstein et al., 1995, p. 98), can be directly mapped to Lectica’s complexity levels based on Dawson’s earlier empirical findings (2002a, 2003b; 2005; 2003; 2003). Cook-Greuter’s mapping of her stages to the Model of Hierarchical Complexity (Cook-Greuter, 1990, p. 104; 2002, pp. 7-28; 2013, pp. 22-76) can be directly mapped to Lectica’s complexity levels, because these levels are partially informed by the Model of Hierarchical Complexity.

Table 31

<table>
<thead>
<tr>
<th>Stage No.</th>
<th>Ego Development Stage</th>
<th>Cook-Greuter’s mapping</th>
<th>Torbert’s mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Model of Hierarchical Complexity</td>
<td>Lectical Level</td>
</tr>
<tr>
<td>1</td>
<td>Impulsive</td>
<td>5^d</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Opportunist</td>
<td>6 &amp; 7</td>
<td>6 &amp; 7</td>
</tr>
<tr>
<td>3</td>
<td>Diplomat</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>Expert</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>Achiever</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>Individualist</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>Strategist</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>Construct-Aware</td>
<td>14^e</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>Unitive</td>
<td>15^f</td>
<td>15</td>
</tr>
</tbody>
</table>
Notes. A dash (-) indicates a cell that is deliberately missing content.  

*a* Cook-Greuter’s mapping of stages to the Model of Hierarchical Complexity is based on information provided by Cook-Greuter (1990, 2002, 2013).  

*b* Torbert’s mapping of stages to Kohlberg’s stages of moral judgement is based on information provided in Lichtenstein et al. (1995).  

*c* Kohlberg’s stages of moral judgement are mapped to complexity levels based on Dawson’s empirical findings (2002a, 2003b; 2005; 2003; 2003).  

*d* Cook-Greuter calls this stage Primary Actions (3a) based on one of the earliest versions of the Model of Hierarchical Complexity (Commons & Richards, 1984), but in the current version of the model (Commons & Pekker, 2008) this stage may map to Stage 5, Sentential.  

*e* Cook-Greuter only refers to one of the earliest versions of the Model of Hierarchical Complexity in which there is a direct leap from stage 12 to stage 14 (Commons & Richards, 1984), but considering the changes made to this model, she may mean stage 13 instead, but this is unclear.  

*f* The Unitive stage is claimed to be a hierarchical integration of the previous stage and also favours direct phenomenological experience of reality which is claimed to transcend the need for symbolically mediated ways of knowing via language, so it is unclear whether the Unitive stage can be mapped to a cognitive-developmental stage at all.

The Ego Development Scoring System

The ego development scoring system aims to identify the primary stage from which test-takers respond to the SCT. The TPR is intended to reflect this primary stage and is used to make inferences about test-takers’ centre of gravity. The ego development scoring system contains a chapter for each stem, and each chapter contains examples and exemplars extracted from actual protocols (Hy & Loevinger, 1996). While there are subtle distinctions between examples and exemplars, the term ‘exemplar’ will be used throughout this thesis for terminological consistency. A SCT stem score is awarded based on the match between test-takers’ responses and exemplars in the manual. For example, test-takers’ responses to the stem “When a child will not join in group activities ...” would be scored in the manner described in Table 32 (Hy & Loevinger, 1996, pp. 88-92).
**Table 32**

*Illustrative Example of Scoring the Stem “When a child will not join in group activities ... “*

<table>
<thead>
<tr>
<th>SCT Stem Completion</th>
<th>Ego Development</th>
<th>Stage #</th>
</tr>
</thead>
<tbody>
<tr>
<td>“the game is boring”</td>
<td>Impulsive</td>
<td>1</td>
</tr>
<tr>
<td>“he is a loser”</td>
<td>Opportunist</td>
<td>2</td>
</tr>
<tr>
<td>“check for possible signs of illness”</td>
<td>Diplomat</td>
<td>3</td>
</tr>
<tr>
<td>“he may be a shy, timid individual”</td>
<td>Expert</td>
<td>4</td>
</tr>
<tr>
<td>“maybe it is due to lack of self-confidence”</td>
<td>Achiever</td>
<td>5</td>
</tr>
<tr>
<td>“he’s probably marching to a different drummer”</td>
<td>Individualist</td>
<td>6</td>
</tr>
<tr>
<td>“it may mean that he has an inner strength and sees a different world”</td>
<td>Strategist</td>
<td>7</td>
</tr>
<tr>
<td>“it may be because he feels no identity with the group, is too self-conscious, or has no real interest in the activity”</td>
<td>Integrated/Construct-Aware</td>
<td>8</td>
</tr>
</tbody>
</table>

**Logic Underpinning Study 2B**

Study 2B used Lectical Phase Scores as a criterion variable to score exemplars from the ego development scoring system. The LAS was used to score exemplars to determine whether a structural isomorphism between ego development stages and complexity levels would be obtained. This approach is informed by Dawson’s (2003b; 2003) earlier work in which LAS/HCSS scores were given to exemplars in domain-specific scoring manuals. This was undertaken to determine if conceptual content unfolded in a similar sequence between different scoring systems. If stages are structurally isomorphic, then the law of transitivity should yield the following relations:

- Exemplar E is definitional of ego development stage S; and
- ego development stage S corresponds to complexity level L; so that
- when exemplar E is scored with the LAS, E should also be located at complexity level L.
An example of this hypothesised correspondence is described in Table 33.

**Table 33** Correspondence Between Ego Development Stages and Lectical Levels for the Stem “When a child will not join in group activities ...”

<table>
<thead>
<tr>
<th>Examples of responses</th>
<th>Ego Development Stage</th>
<th>Cook-Greuter’s correspondence to Lectical Levels a</th>
<th>Torbert’s Correspondence to Lectical Levels b</th>
</tr>
</thead>
<tbody>
<tr>
<td>“the game is boring”</td>
<td>Impulsive</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>“he is a loser”</td>
<td>Opportunist</td>
<td>6–7</td>
<td>9</td>
</tr>
<tr>
<td>“check for possible signs of illness”</td>
<td>Diplomat</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>“he may be a shy, timid individual”</td>
<td>Expert</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>“maybe it is due to lack of self-confidence”</td>
<td>Achiever</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>“he’s probably marching to a different drummer”</td>
<td>Individualist</td>
<td>11</td>
<td>-</td>
</tr>
<tr>
<td>“it may mean that he has an inner strength and sees a different world”</td>
<td>Strategist</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>“it may be because he feels no identity with the group, is too self-conscious, or has no real interest in the activity”</td>
<td>Integrated</td>
<td>14</td>
<td>13</td>
</tr>
</tbody>
</table>

**Notes.** a Cook-Greuter’s mapping of stages to the Model of Hierarchical Complexity is based on information provided by Cook-Greuter (1990, 2002, 2013). b Torbert’s mapping of stages to Kohlberg’s stages of moral judgement is based on information provided by Lichtenstein et al. (1995).

**Using Lectical Phases as a Criterion Variable**

Cook-Greuter (1999) drew upon various developmental models—including those with cognitive-developmental origins—to test the definitional attributes of her two highest stages i.e., Construct-Aware and Unitive. These models included The Model of Hierarchical Complexity.
(Commons & Richards, 1984) and Kegan’s (1982) constructive-developmental stages. Cook-Greuter employed three criteria to identify models against which her stages could be verified. They should (1) posit the existence of a self as an organising principle across domains; (2) investigate conceptions beyond the general systems stage; (3) include a focus on meaning-making and not reasoning in isolation; and (4) be domain-general to examine the multifaceted nature of ego development.

The LAS satisfies most of these criteria and therefore Lectical Phase Scores are well suited to be used as a criterion variable. While cognitive-developmental theory does not explicitly posit the existence of an ego, it is compatible with an integrative psychological structure such as an ego or a general executive function. The Lectical Scale investigates conceptions beyond the general systems stage because it comprises complexity levels beyond the abstract systems level, i.e., Lectical Level 11. As described in the literature review for Study 1A, meaning-making is explicitly considered when scoring with the LAS. Finally, the LAS is domain-general and can score hierarchical complexity in any domain in which it occurs.

**Hypotheses for Study 2B**

- **Visual Analysis 1**: There is a structural isomorphism between ego development stages and complexity levels. Lectical Phase Scores awarded to exemplars for each ego development stage occur within a single complexity level. This demonstrates that each ego development stage reflects a unique magnitude of hierarchical complexity. This is phrased as a visual analysis because it cannot be tested via statistical procedures.

- **Hypothesis 1**: There is a statistically significant increase in mean Lectical Phases Scores awarded to exemplars between successive ego development stages.
Method

Sample

Two samples of exemplars were scored with the LAS: Loevinger’s exemplars from the Impulsive through to Integrated stages (Hy & Loevinger, 1996, pp. 88-92), and Cook-Greuter’s (1999, pp. 73-83) exemplars for the Construct-Aware and Unitive stages. Cook-Greuter contends that Loevinger’s Integrated stage approximately corresponds to her Construct-Aware stage (1999, pp. 230-231). This approach ensured that exemplars for all nine stages could be taken into consideration. All exemplars analysed are found in the public domain.

Loevinger’s Exemplars (Impulsive Through to Integrated). Analyses were conducted on 171 exemplars for the stem “When a child will not join in group activities …”. Exemplars were taken from the second edition of the scoring system (Hy & Loevinger, 1996). Exemplars were distributed over eight stages from Impulsive through to Integrated and the number of exemplars varied by stage: Impulsive: n=3, Opportunist: n=13, Diplomat: n=17, Expert: n=64, Achiever: n=47, Individualist: n=19, Strategist=6, Integrated=2. The stem “When a child will not join in group activities …” was selected for several reasons. It is utilised in most versions of the SCT and is typically perceived to be “… easier to answer than most. It does not make the person answering feel self-conscious, and it is more structured” (Hy & Loevinger, 1996, p. 88). This stem is characterised by clear exemplars which could readily be scored with the LAS with an appropriate level of confidence for research purposes.

Cook-Greuter’s Exemplars (Construct Aware and Unitive).

In an earlier book chapter, Cook-Greuter (1994) provides a detailed description of the inductive methods that were used to generate the definitional attributes of the later stages of ego development. This earlier work also offers examples of stem completions to illustrate the forms of self-understanding that start to become available at the later stages. Her later work provides somewhat revised conceptions of the highest two stages (Cook-Greuter, 1999, pp. 194-195) and a
larger sample of 46 stem completions that exemplify these two stages (Cook-Greuter, 1999, pp. 73-83). Only the stem completions from her later work were included in Study 2B due to the size of the sample and its recency. Considered collectively, these two publications appear to be amongst the few instances of late stage exemplars in the public domain.

Other Materials

The author of this thesis had some direct influence over scores analysed in the context of Study 2B. The author of this thesis had not undertaken formal training in the ego development scoring system, but after completing various self-training exercises (Hy & Loevinger, 1996; Loevinger et al., 1970), achieved an inter-rater agreement of greater than 90% in a sample of several hundred SCTs. Ego development scores were taken directly from the respective ego development scoring systems (Cook-Greuter, 1999; Hy & Loevinger, 1996) so the author of this thesis had no influence over the ego development scores that were analysed.

The author of this thesis used to be a Certified Senior Lectical Analyst™ with over eight years of training in the application of the LAS and CLAS. As a result, the author of this thesis may be regarded as having demonstrated proficiency in both scoring systems considered in the context of this study. All analyses were conducted under the supervision of a Certified Master Lectical Analyst™ with over 20 years of relevant scoring experience in order to verify the accuracy of scores. Neither the Senior Analysts nor the Master Analyst were blind to the purpose of this study. To mitigate against potential bias, when the ego development exemplars were scored with the LAS, key terms were analysed in a manner that is consistent with how they are analysed in other samples. This ensured that the analysis was consistent with the LAS scoring rules and was not subject to special treatment because the exemplars were taken from the ego development scoring system. The LAS was used to award a Lectical Phase Score to each exemplar. Jamovi (2021) was used for all visual and statistical analyses.
**Procedure**

Ego development scorers are asked to “Rate each completion according to its own item manual without regard to context, and record the rating” (Hy & Loevinger, 1996, p. 39). As a result, the complexity of the stem itself is not scored. However, in the context of the LAS, the SCT stem is scored because responses are only scored in the context in which they occur (Dawson, 2010; Dawson-Tunik, 2006). The stem “When a child will not join in group activities ...” was scored at Lectical Phase 09c because the term *group activities* is a single abstraction at Lectical Level 9 and reasoning at this level of hierarchical complexity would need to be understood to meaningfully respond to the stem. Because of this difference in scoring rules, each exemplar was scored in context and out-of-context of the stem. A detailed description of the LAS is provided in Chapter 2, Study 1A. Similar patterns were obtained in both cases, so only results from scoring in context are reported in the Results section below. Findings from scoring out-of-context are provided in Appendix L.

**Results**

A boxplot depicting the relation between ego development stages and the Lectical Phase Scores awarded to exemplars is presented in Figure 30.
Except for the Impulsive stage, for which only three exemplars were available, Lectical Phase Scores did not occur within a single complexity level. There appeared to be a general trend for the median Lectical Phase Score to increase between successive stages. Descriptive statistics are presented in Table 34.
### Table 34

*Descriptive statistics for Lectical Phase Scores awarded to Exemplars for Each Ego Development*

<table>
<thead>
<tr>
<th>Stage</th>
<th>No. exemplars</th>
<th>Min Lectical Phase</th>
<th>Max Lectical Phase</th>
<th>Range of Complexity Levels</th>
<th>Range of Lectical Phases</th>
<th>Mean Lectical Phase</th>
<th>Median Lectical Phase</th>
<th>Modal Lectical Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impulsive</td>
<td>3</td>
<td>09c</td>
<td>09c</td>
<td>1</td>
<td>1</td>
<td>09c</td>
<td>09c</td>
<td>09c</td>
</tr>
<tr>
<td>Opportunist</td>
<td>13</td>
<td>09c</td>
<td>10a</td>
<td>2</td>
<td>3</td>
<td>09c</td>
<td>09c</td>
<td>09c</td>
</tr>
<tr>
<td>Diplomat</td>
<td>17</td>
<td>09c</td>
<td>10c</td>
<td>2</td>
<td>5</td>
<td>09d</td>
<td>09c</td>
<td>09c</td>
</tr>
<tr>
<td>Expert</td>
<td>64</td>
<td>09c</td>
<td>11a</td>
<td>3</td>
<td>7</td>
<td>10a</td>
<td>09d</td>
<td>09c</td>
</tr>
<tr>
<td>Achiever</td>
<td>47</td>
<td>09c</td>
<td>11a</td>
<td>3</td>
<td>7</td>
<td>10b</td>
<td>10b</td>
<td>10b</td>
</tr>
<tr>
<td>Individualist</td>
<td>19</td>
<td>09d</td>
<td>11a</td>
<td>3</td>
<td>6</td>
<td>10c</td>
<td>10c</td>
<td>10c</td>
</tr>
<tr>
<td>Strategist</td>
<td>6</td>
<td>10c</td>
<td>11a</td>
<td>2</td>
<td>3</td>
<td>10d</td>
<td>10c/10d</td>
<td>10c</td>
</tr>
<tr>
<td>Integrated</td>
<td>2</td>
<td>10d</td>
<td>11a</td>
<td>2</td>
<td>2</td>
<td>10d/11a</td>
<td>10d/11a</td>
<td>NA</td>
</tr>
<tr>
<td>Construct-Aware</td>
<td>20</td>
<td>10a</td>
<td>11c</td>
<td>2</td>
<td>7</td>
<td>10d</td>
<td>10d</td>
<td>10d</td>
</tr>
<tr>
<td>Unitive</td>
<td>26</td>
<td>10c</td>
<td>11b</td>
<td>2</td>
<td>4</td>
<td>10d/11a</td>
<td>11a</td>
<td>11a</td>
</tr>
</tbody>
</table>

As can be seen from the descriptive statistics, the Impulsive stage mapped to Lectical Level 9. For all other stages, exemplars spanned two or three complexity levels or three to seven Lectical Phases. This pattern does not resemble the correspondences predicted by Torbert or Cook-Greuter described in Table 33 above. When the predicted relations were compared to the actual relations obtained, a loose correspondence was identified for Opportunist, Expert, and Achiever but not for other stages.

The full range and the interquartile ranges between successive stages overlap. For example, the Expert and Achiever stages ranged from Lectical Phase 09c to 11a. Because of these overlapping ranges, it is possible for leaders to demonstrate an increase in their SCT score while simultaneously demonstrating an increase (Scenario 1), statis (Scenario 2), or regression (Scenario 3) in their Lectical Phase Score. These three scenarios are demonstrated visually in Figure 31.

**Figure 31**

*Three Scenarios: Increase, Statis, and Regression in Lectical Phase Scores*
Note. Scenario 1: eight stages of growth from Impulsive to Construct-Aware and two Lectical Levels of growth from 09c to 11c. Scenario 2: six stages of growth from Expert to Unitive and no Lectical growth from 10d. Scenario 3: five stages of growth from Expert to Construct-Aware and regression by one Lectical Level from 11a to 10a.

A one-way Analysis of Variance (ANOVA) was conducted to test for significant differences in mean Lectical Phase Scores awarded to exemplars between successive stages. The Impulsive (n=3) and Integrated (n=2) stages were removed from the analysis because of the small number of exemplars available. The ANOVA was significant (F(7, 48.6)=101.50, p<.001). Levene’s test of equality of error variances was also significant (F(7,204)=8.31, p<.001), so post hoc comparisons were conducted using the Games-Howell Test, which does not assume equality between group variances. For all pairwise comparisons except one, there was a nonsignificant difference. Pairwise comparisons are presented in Table 35.
Table 35

Means, Standard Deviations, Mean Differences, and One-Way Analysis of Variance (ANOVA) on
Lectical Phase Scores Between Successive Ego Development Stages

<table>
<thead>
<tr>
<th>Stage</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Opportun</td>
<td>13</td>
<td>09c</td>
<td>0.60</td>
<td>-0.48</td>
<td>-1.79***</td>
<td>-2.34***</td>
<td>-3.35***</td>
<td>-4.60***</td>
<td>-5.22***</td>
<td>-5.27***</td>
<td></td>
</tr>
<tr>
<td>3. Diplomat</td>
<td>17</td>
<td>09d</td>
<td>1.10</td>
<td>-1.31*</td>
<td>-1.87***</td>
<td>-2.87***</td>
<td>-4.13***</td>
<td>-4.74***</td>
<td>-4.79***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Expert</td>
<td>64</td>
<td>10a</td>
<td>1.85</td>
<td>-0.56</td>
<td>-1.56*</td>
<td>-2.82**</td>
<td>-3.43***</td>
<td>-3.48***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Achiever</td>
<td>47</td>
<td>10b</td>
<td>1.75</td>
<td>-1.00</td>
<td>-2.26*</td>
<td>-2.88***</td>
<td>-2.93***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Individual</td>
<td>19</td>
<td>10c</td>
<td>1.47</td>
<td>-1.25</td>
<td>-1.87**</td>
<td>-1.92***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Strategist</td>
<td>6</td>
<td>10d</td>
<td>0.98</td>
<td></td>
<td>-0.62</td>
<td>-0.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Construct Aware</td>
<td>20</td>
<td>10d</td>
<td>1.57</td>
<td></td>
<td>-</td>
<td>-0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Unitive</td>
<td>26</td>
<td>11a</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. * p<.05, ** p<.01, *** p<.001

As can be seen from Table 35, there was a significant difference between Diplomat and Expert, but not for other pairwise comparisons. The analysis was repeated using non-parametric procedures (i.e., Kruskal-Wallis and Mann-Whitney U tests) but yielded identical findings, so they are not reported.

An exploratory analysis found that most Lectical Phase Scores spanned 1.5 complexity levels from Lectical Phase 09c to 11a. Most Impulsive, Opportunistic, and Diplomat exemplars occurred within Lectical Level 9. Most exemplars for Expert through to Unitive occurred between Lectical Phases 10a to 11a. A few exemplars for Construct-Aware and Unitive were scored at 11b and 11c. This is not consistent with the predicted correspondences.
Hypothesis Tests for Study 2B

- Visual Analysis 1 was not supported. There did not appear to be a structural isomorphism between ego development stages and complexity levels. Lectical Phase Scores awarded to exemplars for each ego development stage did not occur within a single complexity level. Instead, they spanned two or three complexity levels or three to seven Lectical Phases. This did not demonstrate that ego development stages reflect a unique magnitude of hierarchical complexity.

- Hypothesis 1 was not supported because there was no significant increase in mean Lectical Phase Scores awarded to exemplars between successive ego development stages. The only exception was the difference between Diplomat and Expert.

Discussion

Findings did not support the hypotheses. Findings suggest that when ego development exemplars are scored with the LAS, ego development stages do not reflect a unique magnitude of hierarchical complexity. Furthermore, there is no significant increase in hierarchical complexity
between most ego development stages, as operationalised by Lectical Phase Scores. As a result, Study 2B failed to demonstrate a structural isomorphism between the two sets of stages.

It was striking that most exemplars occurred within 1.5 complexity levels, including the highest stages of ego development. Some of Cook-Greuter’s writing reflects a critical disposition towards formal operations, which is the theoretical equivalent of Lectical Level 10. For example, she contends that stages beyond formal operations bestow advantages such as “more flexible self-views, long-term perspectives” (Cook-Greuter, 1999, p. 187). In other places, she claims that the “… Construct-aware and Unitive stage seem to fulfil the stage requirements for being higher-order hierarchical integrations” (Cook-Greuter, 1999, p. 160). Ironically, findings suggest that Construct-aware and Unitive exemplars may be poetic expressions of formal operations which fail to fulfil the requirements of higher-order hierarchical integrations.

Cook-Greuter (1999, p. 96) admitted that “[Her] notions about postautonomous development can be readily proven false if there is evidence that individuals at earlier ego stages are capable of making the kinds of distinctions at the required level of complexity that, [she is] suggesting, are reserved for the most advanced meaning makers”. Findings suggest that not only have her notions about postautonomous development been ‘proven false’, but the general claim that stages satisfy the requirement of hierarchical integrations has also been challenged. This conclusion is informed by an analysis that does not depend on the broad distribution of stem scores.

These findings may explain the results reported in Studies 1B and 2A. In both cases, the requirement of invariant sequence was satisfied, but other requirements were not. If these findings generalised to other stems, it may be suggested that the ego development scoring system is not awarding scores in a manner that yields information about hierarchical complexity. These findings may account for why a correlation of .35 to .52 is generally found between ego development and cognitive-developmental scores (Cohn & Westenberg, 2004), but a correlation of .86 to .92 is generally found between scores based on hard stage models (Dawson, 2001, 2002a, 2003b; Dawson et al., 2003).
Findings may also challenge the contention that the SCT can be used to measure development through the lifespan. Given that most exemplars occurred between 09c to 11a, the SCT can only account for development in adolescents and the general adult population. However, it may not be appropriate for leader populations shown to reason between 10c and 12b in Study 1A.

Study 2B also surfaces another possibility that was not apparent from Studies 1B and 2A. Namely, that ego development stages may reflect a sequence of worldviews that are distinct with respect to their content—and perhaps underlying epistemic assumptions—but nevertheless can be expressed at a range of different complexity levels. For example, the Expert and Achiever stages are generally contended to reflect a conventional worldview as a result of their emphasis on empirical positivism, intellectual mastery, performance improvement, outcomes and assessments, etc. (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 2013). However, Study 2B suggests that both stages can be expressed at a range of complexity levels, from Lectical Level 9 to 11. Analogously, the Individualist and Strategist stages are generally contended to reflect a postconventional worldview as a result of their emphasis on relativistic reasoning, individuality, interpretation, pluralistic perspectives, etc. (Cook-Greuter, 1999; Loevinger, 1976; Torbert, 2013). However, Study 2B suggests that both stages can be expressed at a range of complexity levels, from Lectical Level 9 to 11. This is consistent with the finding that relativistic reasoning—which is often considered to be indicative of the Individualist stage—may be found at a range of complexity levels (Commons & Richards, 2002; Dawson & Stein, 2004a). Similar observations can be made about the preconventional worldviews of the Impulsive and Opportunist stages and also the stages that follow the post-conventional worldviews i.e., Construct-Aware and Unitive stages. These findings may surface the possibility that growth in hierarchical complexity may occur within, rather than between, distinct worldviews. This is not to positively suggest that ego development stages may form a nominal scale of worldviews that do not reflect an increase in complexity, however Study 2B does not rule out this possibility.

Notwithstanding the convergence between findings on ego development reported thus far, a significant limitation of Study 2B is that it depends upon an analysis of a single stem. As a result, it
remains an empirical question whether similar patterns would be obtained in a representative sample of stems.
**Study 2C: Regression Analyses on the Cognitive-Developmental Scores and Ego Development Scores**

The purpose of Study 2C was to explicitly test whether cognitive-developmental and ego development scores reflect the logic underpinning cultural age, soft stage, or hard stage models. Similarly to Study 2B, Study 2C afforded a method of testing hard stage requirements in a manner that circumvented the broad response distributions typically found in SCT protocols.

**Relation Between Ego Development and Cognitive-Developmental Scores**

As discussed in Studies 1B and 2B, ego developmentalists argue for a structural isomorphism between ego development and cognitive-developmental stages. While this claim seems to be made at the scale of theory, it can be operationalised at the scale of measurement. However, correlations between SCT scores and cognitive-developmental scores are typically lower than that found between scores based on hard stage models, as discussed in Chapter 2. A closer examination suggests these findings may be an artefact of the research methodologies employed in earlier research.

Earlier research administered separate assessments to a common sample of test-takers, scored each type of assessment with its respective scoring system, and then analysed correlations between scores (refer to Figure 33). This is a relatively standard methodology used in the social sciences but introduces conflating factors which may account for the correlations obtained. The SCT and cognitive-developmental assessments are typically administered under different conditions. The SCT is administered with open-ended instructions, the response format is focused on ego projections, and responses typically focus on broad aspects of test-takers’ worldviews. In contrast, cognitive-developmental assessments are administered with direct instructions, the response format is focused on reasoning in specific domains, and responses typically focus on judgements (‘what do you think?’) and justifications (‘why do you think what you think?’). Study 2C mitigates these
confounding factors by scoring a common sample of SCTs with CLAS and the ego development scoring system, so that clearer attributions can be made about the relations between scores.

**Figure 33**

*Mitigation of Confounding Factors Which May Have Impacted Earlier Research*

Proxy Variables to Operationalise the Logic Underpinning Cultural Ages, Soft Stages, and Hard Stages

As discussed in Chapter 1, Kohlberg’s three types of adult development theories may exhaustively account for those considered in the field of developmental psychology. Cultural ages are operationalised by nominal categories that are sequenced temporally. Soft stages are operationalised by ordinal stages that are sequenced cumulatively. Hard stages are operationalised
by ordinal stages that are sequenced hierarchically. The logics underpinning these three models can be operationalised via meaningful proxy variables: increasing age for cultural ages, a cumulative increase in a theoretically relevant construct for soft stages, and hierarchical complexity for hard stages.

The number of perspectives taken, or word count may reflect cumulative increases in theoretically relevant constructs in the context of adult development theory. The acquisition of different perspectives has also been proposed to account for the logic underpinning some adult development stage theories (Cook-Greuter, 1999; Selman, 1980; Torbert, 1994). Word count is often contended to increase between successive stages because a greater number of words is often required to express increasingly complex ideas (Cohn & Westenberg, 2004).

Operationalising the Logic Underpinning the Lectical Scale

As discussed in Study 1A, the Lectical Scale is a unidimensional ordinal scale which reflects a single underlying construct, i.e., hierarchical complexity. Unidimensionality has been supported by Dawson’s earlier findings that fit statistics are consistent with the Rasch Model, and LAS/HCSS scores are highly correlated with scores based on other hard stage cognitive-developmental models (Dawson, 2002a). In Study 1A, a negligible amount of multidimensionality was identified in Lectical Phase Scores, which further supports the unidimensionality of the scale (c = .01). Previous research has demonstrated that LAS scores (Dawson, 2010, 2022b; Dawson-Tunik, 2006) and HCSS scores (Dawson, 2003a) are only weakly related to age (r = .16) (Dawson et al., 2005), and the number of perspectives taken in adulthood (r = .22) (Fuhs, 2016). As a result, age, word count, and the number of perspectives taken are not expected to account for a substantial amount of variability in CLAS scores.
Operationalising the Logic Underpinning the Ego Development Scale

As discussed in Studies 1B, 2A, and 2B, there is limited evidence to support the contention that SCT scores reflect the underlying construct of hierarchical complexity. Previous studies have demonstrated that TPRs are also weakly related to age in adulthood ($r=.04$) (Cohn, 1998). However, there are reasons to believe that SCT scores may be related to the number of perspectives taken and word count. The acquisition and deepening of perspectives have been proposed to account for the transformation between successive ego development stages, as discussed in Chapter 2 (Cook-Greuter, 1999; Torbert, 1994). Furthermore, Lanning’s (2018) analysis of responses to nearly 44,000 stems indicated a significant relation between SCT stem scores and pronoun use.

Numerous studies have examined the relation between SCT scores and word count and/or verbal ability (Browning, 1987; Cohn & Westenberg, 2004; Drewes & Westenberg, 2001; Einstein & Lanning, 1998; Ginsburg & Orlofsky, 1981; Hansell et al., 1984; Labouvie-Vief et al., 1989; Lanning et al., 2018; Loevinger & Wessler, 1970; McCrae & Costa, 1980; Morros et al., 1998; Skoe & von der Lippe, 2002). Many of these studies suggest that SCT scores are significantly related to, but not completely accounted for, by word count and/or verbal ability. For example, Cohn and Westenberg (2004) undertook a meta-analysis of 23 studies consisting of over 2,700 participants and found a correlation between SCT scores and word count that ranged from approximately .20 to .80 with a weighted average of .54. In an analysis of responses to nearly 44,000 stem completions, Lanning et al. (2018) found a significant correlation of .45 between SCT stem scores and words per sentence, and .40 correlation between SCT stem scores and word count. As a result, there are reasons to hypothesise that number of perspectives and word count may account for a significant amount of variability in SCT scores.

Hypotheses for Study 2C

Study 2C aimed to identify which proxy variable—age, number of perspectives, word count, or hierarchical complexity—best predicts CLAS scores and SCT scores awarded to a common set of
protocols. From the perspective of theory, complexity levels and ego development stages are contended to satisfy hard stage requirements. From the perspective of measurement, cognitive-developmentalists have demonstrated—and ego developmentalists have asserted—that their respective scores satisfy hard stage requirements. If CLAS and SCT scores are a sound reflection of hierarchical complexity, then age, number of perspectives, and word count should not account for additional variability in CLAS scores or SCT scores beyond the variance shared between them. This is demonstrated conceptually in Figures 34 and 35.

Figure 34

Conceputal Relation between Proxies and CLAS Scores
Hypothesis 1: The correlation between CLAS scores and SCT scores awarded to a common sample of SCTs occurs between .86 and .92, which is typically found between scores based on hard stage models.

Hypothesis 2: Beyond the variance shared between CLAS scores and SCT scores, age, number of perspectives, and word count do not account for a statistically significant increase in the predictive power of a regression model performed on CLAS scores (refer to Figure 34 above).

Hypothesis 3: Beyond the variance shared between SCT scores and CLAS scores, age, number of perspectives and word count do not account for a statistically significant increase in the predictive power of a regression model performed on SCT scores (refer to Figure 35 above).
Method

Sample

This study drew upon the same sample that was used for Study 1B. As a result, the nature of the sample and data permissions were described in the context of Study 1B. The original sample consisted of 996 SCTs, however TPRs and completed protocols were only available for 937 of the original sample. Filters were applied to the original sample to yield an analysable sample and thus, the sampling process was not entirely random. The analysable sample consisted of protocols at test-time 1 (n=321, 81.7%), test-time 2 (n=69, 17.6%), and test-time 3 (n=3, 0.8%). To mitigate violations to the assumption of independent observations, only assessments administered at test-time 1 were included in the final sample. When data were collected, test-takers were provided with a binary option to select for sex. Demographic details for these leaders are provided in Table 36. There is a large percentage of cases for this sample for which demographic details are unidentified. The reason(s) are unknown/unknowable. The demographics fields at the top of the SCT are not mandatory and so test-takers can choose not to answer these fields. However, the author of this thesis is not aware of anything unusual about this sample that would make test-takers sensitive to being identified.
Table 36

Demographic Details of Leaders for Study 2C

<table>
<thead>
<tr>
<th>Demographics</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (at test time)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 35 years</td>
<td>23</td>
<td>7.2%</td>
</tr>
<tr>
<td>35 to 44 years</td>
<td>41</td>
<td>12.8%</td>
</tr>
<tr>
<td>45 to 54 years</td>
<td>42</td>
<td>13.1%</td>
</tr>
<tr>
<td>&gt; 54 years</td>
<td>21</td>
<td>6.5%</td>
</tr>
<tr>
<td>Unidentified</td>
<td>194</td>
<td>60.4%</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>67</td>
<td>20.8%</td>
</tr>
<tr>
<td>Female</td>
<td>63</td>
<td>19.6%</td>
</tr>
<tr>
<td>Unidentified</td>
<td>191</td>
<td>59.5%</td>
</tr>
<tr>
<td>First language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>228</td>
<td>71.0%</td>
</tr>
<tr>
<td>Not English</td>
<td>68</td>
<td>21.1%</td>
</tr>
<tr>
<td>Unidentified</td>
<td>25</td>
<td>7.8%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>70</td>
<td>21.8%</td>
</tr>
<tr>
<td>Non-Caucasian</td>
<td>18</td>
<td>5.6%</td>
</tr>
<tr>
<td>Unidentified</td>
<td>233</td>
<td>72.6%</td>
</tr>
</tbody>
</table>

Materials

As described in Study 1B, five versions of the SCT were administered. Invariance between versions of the SCT, and SCT scorers, was established in Study 1B. Rasch person estimates were produced by RUMM software in Study 1B and were used to inform these analyses because of their scaling precision (Andrich et al., 2015).

As discussed in Chapter 2, test-takers’ responses can be scored via Lectica’s two scoring systems—the LAS and the CLAS. The purpose of both systems is to locate the magnitude of hierarchical complexity exhibited in samples of reasoning along the Lectical Scale. CLAS was used to score leaders’ responses to SCT protocols for Study 2C. As discussed in Chapter 2, LAS and CLAS have a high inter-rater agreement and can therefore be treated as interchangeable for research purposes within certain constraints. Lectical decimal scores have been used in some empirical
research (Fuhs, 2016; Van Rossum, 2013). Statistical models were constructed in Mplus (Muthén & Muthén, 1998-2015).

**Procedure**

**Scoring the Protocols with the Ego Development Scoring System.** Protocols had already been scored via the ego development scoring system. Each protocol was awarded a TPR, a mean SCT score, and an SCT Rasch person estimate based on the analyses undertaken in Study 1B.

**Training the Computerised Lectical Assessment System (CLAS) to Score Protocols.** At the later stages of ego development, stems elicit responses with a more poetic or esoteric nature e.g., “seeing the world in a grain of sand” (Cook-Greuter, 1999, p. 195). To ensure that CLAS could confidently score protocols containing these poetic terms, 112 of the original 937 protocols (12% of the original sample) were randomly selected and scored by a Certified Senior Lectical Analyst™ or a Certified Master Lectical Analyst with a minimum of 4+ years of experience. Through this scoring process, CLAS learnt how a human scorer would score a protocol containing these uncommon terms. As a result, the CLAS scoring algorithm was trained to score protocols at the higher stages of ego development.

The author of this thesis had some indirect influence over the scores analysed in the context of Study 2C by playing a role in determining the Lectical Phase at which these uncommon terms should be located along the Lectical Scale. These terms were analysed in a manner that is consistent with how they are analysed in other samples. This mitigated against potential bias by ensuring that the analysis was consistent with Lectica’s approach to scoring and was not subject to special treatment because the terms were derived from SCT protocols. CLAS then applied a discrimination function to each protocol based on the density of concepts at successive Lectical Phases to yield a CLAS Score as it would for any other sample of reasoning. Because CLAS is a computerised scoring system, neither the author of this thesis nor the Certified Master Lectical Analyst had any additional influence over the CLAS score awarded to each protocol.
Selection of Cases. In addition to awarding a CLAS score, CLAS also awards other scores that can be used as filters. Filters were applied to identify assessments from the original sample that had been scored with sufficient confidence to include in the analysable sample. Two filters were used in the context of Study 2C. The first filter was CLAS’s confidence in the CLAS score based on the density of concepts at successive Lectical Phases. The confidence score ranged from 1 to 10, and protocols with a confidence score equal to or greater than 6 were included. This threshold was considered suitable for research purposes and is consistent with the general procedures applied to previous research (Fuhs, 2016; Van Rossum, 2013). The second filter was the unique word count. Protocols with a unique word count equal to or greater 120 words were included to ensure sufficient text for CLAS to score.

Modelling Approach. CLAS awards a single score to each assessment rather than a unique score to each item. For this reason, regression models on manifest variables were used, rather than a combination of measurement and structural models. This approach imposes some limitations on the ability to account for error variance. However, this is justified based on the measurement properties displayed by the LDMA in Study 1A.

Construction of Additional Variables. Two additional variables were constructed to operationalise the cumulative logic underpinning soft stage models: unique perspectives and unique word count. The unique perspectives variable was based on the number of grammatical persons used in each protocol. There are eight grammatical persons in the English language. They are the first person (e.g., “I”), the second person (e.g., “you”), and the third person (e.g., “it”) in the masculine gender (e.g., “he”) and feminine gender (e.g., “she”) in the singular (e.g., “I”) and plural (e.g., “we”). There are five pronouns and adjectives for each of these eight grammatical persons. They are subject pronouns (e.g., “I”), object pronouns (e.g., “me”), possessive adjectives (e.g., “my”), possessive pronouns (e.g., “mine”) and reflexive pronouns (e.g., “myself). After removing duplicated terms, 31 unique grammatical persons remained (i.e., I, me, my, mine, myself, you, your, yours,
yourself, he, him, his, himself, she, her, hers, herself, it, its itself, we, us, our, ours, ourselves, yourselves, they, them, their, theirs, themselves). For details, refer to Table 37.

**Table 37**

*Grammatic Persons Used to Develop the ‘Unique Number of Perspectives’ Variable*

<table>
<thead>
<tr>
<th>Grammatical person</th>
<th>Gender</th>
<th>Subject Pronoun</th>
<th>Object Pronoun</th>
<th>Possessive Adjectives</th>
<th>Possessive Pronoun</th>
<th>Reflexive Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>First person singular</td>
<td>Neutral</td>
<td>I</td>
<td>me</td>
<td>my</td>
<td>mine</td>
<td>myself</td>
</tr>
<tr>
<td>Second person singular</td>
<td>Neutral</td>
<td>you</td>
<td>you</td>
<td>your</td>
<td>yours</td>
<td>yourself</td>
</tr>
<tr>
<td>Third person singular</td>
<td>Masculine</td>
<td>he</td>
<td>him</td>
<td>his</td>
<td>his</td>
<td>himself</td>
</tr>
<tr>
<td>Third person singular</td>
<td>Feminine</td>
<td>she</td>
<td>her</td>
<td>her</td>
<td>hers</td>
<td>herself</td>
</tr>
<tr>
<td>Third person singular</td>
<td>Neutral</td>
<td>it</td>
<td>it</td>
<td>its</td>
<td>its</td>
<td>itself</td>
</tr>
<tr>
<td>First person plural</td>
<td>Neutral</td>
<td>we</td>
<td>us</td>
<td>our</td>
<td>ours</td>
<td>ourselves</td>
</tr>
<tr>
<td>Second person plural</td>
<td>Neutral</td>
<td>you</td>
<td>you</td>
<td>your</td>
<td>yours</td>
<td>yourselves</td>
</tr>
<tr>
<td>Third person plural</td>
<td>Neutral</td>
<td>they</td>
<td>them</td>
<td>their</td>
<td>theirs</td>
<td>themselves</td>
</tr>
</tbody>
</table>

This approach was more aligned with Cook-Greuter’s (1999) perspectives on the self, than Torbert’s (1994) phenomenologically oriented perspectives. Counting the number of unique perspectives mitigated the risk of inflated scores for leaders who may have repeatedly used the same grammatical person. It also afforded an ordinal scale with a sufficient range to conduct meaningful analyses.

CLAS automatically generates unique word count. This was used instead of total word count to mitigate the risk of inflated scores because of the repetitious use of language. Some covariance between unique perspectives and unique word count is expected because a greater number of perspectives would be reflected in longer responses.
Results

Hierarchical Clustering in the Data Structure

The data were nested at multiple layers, so the amount of variability attributable to hierarchical clustering needed to be determined. *Program* was used as the clustering variable because the leader development program was arguably the most immediate context that could have influenced leaders’ scores. Hierarchical clustering was analysed via two methods—intraclass correlations (ICCs) and the design effect (Deff). ICCs partition variance at the within and between levels and then determine the proportion of within variance relative to the total variance (i.e., within and between). Deff indicates the magnitude of bias in the standard errors (SEs) when the data are not independent such that $\text{Deff} = 1 + (c - 1)p$, where $c$ is the average cluster size at 45.86, and $p$ is the ICC. ICCs greater than .05 and Deffs greater than 2 are typically considered to require multi-level modelling. Results are presented in Table 38.

Table 38

Variability Attributable to Hierarchical Clustering in the Data Structure for Study 2C

<table>
<thead>
<tr>
<th>Variable</th>
<th>ICC (p)</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCT Ogive</td>
<td>.17</td>
<td>8.76</td>
</tr>
<tr>
<td>SCT Mean</td>
<td>.27</td>
<td>12.93</td>
</tr>
<tr>
<td>SCT Rasch</td>
<td>.27</td>
<td>12.98</td>
</tr>
<tr>
<td>CLAS Score</td>
<td>.20</td>
<td>9.97</td>
</tr>
<tr>
<td>Unique perspectives</td>
<td>.19</td>
<td>9.57</td>
</tr>
<tr>
<td>Unique word count</td>
<td>.20</td>
<td>9.93</td>
</tr>
<tr>
<td>Age</td>
<td>.06</td>
<td>3.51</td>
</tr>
</tbody>
</table>

The ICCs ranged from .06 to .27 and therefore exceeded the critical value of .05. All Deffs exceeded the critical value of two. As a result, aggregated multi-level modelling was used to obtain
unbiased estimates and standard errors. This approach was favoured over disaggregated modelling because analyses were aggregated at the level of overall scores for each protocol rather than disaggregated by comparing sources of variation at various levels of the data hierarchy. An aggregated multi-level modelling approach also offered greater flexibility when constructing the model syntax (Preacher et al., 2010).

**Intercorrelations Between Main Variables**

A correlation matrix was constructed to conduct a preliminary exploration of the relations between variables (refer to Table 39).

**Table 39**

*Descriptive Statistics and Correlations Between Variables for Study 2C*

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. SCT ogive</td>
<td>321</td>
<td>6.09</td>
<td>.81</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. SCT mean</td>
<td>321</td>
<td>5.42</td>
<td>.63</td>
<td>.88*</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. SCT Rasch</td>
<td>321</td>
<td>0.27</td>
<td>.86</td>
<td>.88*</td>
<td>.99*</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. CLAS score</td>
<td>321</td>
<td>10.99</td>
<td>.31</td>
<td>.46*</td>
<td>.51*</td>
<td>.52*</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Unique perspectives</td>
<td>321</td>
<td>13.06</td>
<td>3.63</td>
<td>.51*</td>
<td>.53*</td>
<td>.55*</td>
<td>.37*</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Unique word count</td>
<td>321</td>
<td>294.76</td>
<td>160.6</td>
<td>.63*</td>
<td>.67*</td>
<td>.70*</td>
<td>.49*</td>
<td>.74*</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7. Age</td>
<td>127</td>
<td>44.35</td>
<td>10.10</td>
<td>-.05</td>
<td>-.01</td>
<td>-.02</td>
<td>-.006</td>
<td>-.11</td>
<td>-.09</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note.* * p<.05, ** p<.001.
As expected, there was a strong correlation between all three versions of the SCT score \((.88 \leq r \leq .99, .77 \leq r^2 \leq .98)\). Because of this strong correlation and measurement precision offered by Rasch scores, all subsequent analyses were based on SCT Rasch scores. SCT Rasch scores were moderately correlated with CLAS scores \((r=.52, r^2=.27, p<.001)\), moderately correlated with unique perspectives \((r=.55, r^2=.30, p<.001)\), strongly correlated with unique word count \((r=.70, r^2=.49, p<.001)\), but not significantly correlated with age \((r=-.02, r^2=.00, p=.86)\). CLAS scores were weakly correlated with unique perspectives \((r=.37, r^2=.14, p<.001)\), moderately correlated with unique word count \((r=.50, r^2=.24, p<.001)\), but not significantly correlated with age \((r=-.01, r^2=.00, p=.94)\).

Unique perspectives and unique word count were strongly correlated \((r=.74)\), but the Variance Inflation Factor (VIF) < 10 and Tolerance > .10 which did not indicate significant multicollinearity. As a result, they were treated as independent variables in the regression models. Age was removed from subsequent analyses because it was only available for 40% of cases and was statistically unrelated to all other variables.

**Sequential Regression on CLAS Scores**

A sequential regression analysis was undertaken on CLAS scores in which SCT Rasch score, unique perspectives, and unique word count were introduced to the model sequentially. As Model 1 only included the SCT Rasch score, the results are identical to the bivariate correlation between CLAS and SCT Rasch scores reported in the correlation matrix above \((r=.52, r^2=.27, p<.001)\). For Model 2, SCT Rasch scores and unique perspectives were included as predictors. Considered jointly, these predictors accounted for 28% of the variability in CLAS scores \((r^2=.28, p<.001)\). This represented a nonsignificant increase of 1% in predictive power relative to Model 1 \((F(1,318)=3.98, p=.05)\). The standardised beta-estimates indicated that SCT Rasch scores were a significant predictor of CLAS scores \((\beta=.46, p<.001)\) but unique perspectives were not \((\beta=.11, p=.19)\). For Model 3, SCT Rasch scores, unique perspectives, and unique word count were included as predictors. Considered jointly, this set of predictors accounted for 30% of the variability in CLAS scores \((r^2=.30, p<.001)\). This
represented a significant increase of 2% in predictive power relative to Model 2 ($F(1, 317) = 10.92, p = .001$). The standardised beta-estimates indicated that SCT Rasch scores were a stronger predictor of CLAS scores ($\beta_s = .35, p < .001$) than unique word count ($\beta_s = .27, p < .001$) and unique perspectives no longer significantly contributed to the model ($\beta_s = -.020, p = .60$).

**Sequential Regression on SCT Rasch Scores**

A sequential regression analysis was undertaken on SCT Rasch scores in which CLAS scores, unique perspectives, and unique word count were introduced to the model sequentially. As Model 1 only included the CLAS scores, the results are identical to the bivariate correlation between SCT Rasch scores and CLAS scores reported in the correlation matrix above ($r = .52, r^2 = .27, p < .001$). For Model 2, CLAS scores and unique perspectives were included as predictors. Considered jointly, this set of predictors accounted for 42% of the variability in SCT Rasch scores ($r^2 = .42, p < .001$). This represented a significant increase of 15% predictive power relative to Model 1 ($F(1, 318) = 82.3, p < .001$). The standardised beta-estimates indicated that unique perspectives were a stronger predictor of SCT Rasch scores ($\beta_s = .42, p < .001$) relative to CLAS scores ($\beta_s = .37, p < .001$). For Model 3, CLAS scores, unique perspectives, and unique word count were included as predictors. Considered jointly, this set of predictors accounted for 53% of the variability in SCT Rasch scores ($r^2 = .53, p = .01$). This represented a significant increase of 11% predictive power relative to Model 2 ($F(1, 317) = 74.4, p < .001$). The standardised beta-estimates indicated that unique word count displayed more than twice the predictive power ($\beta_s = .527, p = .002$) of CLAS scores ($\beta_s = .234, p = .006$) but unique perspectives no longer significantly contributed to the model ($\beta_s = .078, p = .28$).

**Comparison Between Sequential Multiple Regressions on CLAS Scores and SCT Rasch Scores**

For comparative purposes, Table 40 summarises the results of the sequential regressions on SCT Rasch scores and CLAS scores.
Table 40

Summary of Sequential Regression on CLAS Scores and SCT Rasch Scores

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression on CLAS Scores</th>
<th>Regression on SCT Rasch Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Variable</td>
<td>β</td>
</tr>
<tr>
<td>1</td>
<td>SCT Rasch score</td>
<td>.52***</td>
</tr>
<tr>
<td>2</td>
<td>SCT Rasch score</td>
<td>.46***</td>
</tr>
<tr>
<td></td>
<td>Unique perspectives</td>
<td>.11</td>
</tr>
<tr>
<td>3</td>
<td>SCT Rasch score</td>
<td>.35***</td>
</tr>
<tr>
<td></td>
<td>Unique perspectives</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>Unique word count</td>
<td></td>
</tr>
</tbody>
</table>

Note. *p<.05. **p<.01. ***p<.001.

From Table 40, the amount of variability accounted for in CLAS scores increased negligibly from 27% to 28% to 30% between successive models. Based on the standardised beta-coefficients, unique perspectives and unique word count never explained more variability in CLAS scores than SCT Rasch scores. For the SCT Rasch scores, the variability accounted for increased considerably from 27% to 42% to 53% between successive models. Based on the standardised beta-coefficients, unique perspectives and unique words explained more variability in SCT Rasch scores than CLAS scores.

Hypotheses Tests for Study 2C

- Hypothesis 1 was not supported. The correlation between CLAS scores and SCT Rasch scores was .52 which fell outside the range of correlations typically found between scores based on hard stage models i.e., .86 to .92, where z=-9.98, p<.001.
• Hypothesis 2 was partially supported. Beyond the variance shared between CLAS scores and SCT Rasch scores, number of perspectives and word count did account for an increase in the predictive power of a regression model on CLAS scores. This increase was statistically significant but practically negligible, i.e., 2%.

• Hypothesis 3 was not supported. Beyond the variance between SCT Rasch scores and CLAS scores, number of perspectives and word count did account for a significant increase in the predictive power of the regression model on SCT Rasch scores. The size of this increase was practically significant, i.e., up to 26%.

Discussion

Findings provide mixed support for the hypotheses. Correlations between CLAS scores and SCT Rasch scores ($r=.52$) occurred within the range found between ego development and cognitive-developmental scores (.35 to .52) (Cohn & Westenberg, 2004) but below the range typically found between scores based on hard stage models (.86 to .92). Considered collectively, the evidence suggests that hierarchical complexity only accounts for 12-27% of the variability in SCT scores, which means that 73-88% of the variability in SCT scores cannot be directly attributed to hierarchical complexity. This is insufficient to suggest that ego development stages are hierarchically integrated and that an increase in hierarchical complexity is a necessary but not sufficient condition for transformation between successive stages. Neither unique perspectives nor unique words were found to contribute substantially to predicting CLAS scores beyond their shared variance with SCT Rasch scores. However, unique perspectives and unique words contributed significantly to predicting SCT Rasch scores beyond their shared variance with CLAS scores.

The regression models on CLAS scores are consistent with earlier research on the unidimensionality of Lectical Scores and the negligible amount of multidimensionality found in Lectical Phase Scores in Study 1A. The models demonstrate that CLAS scores are weakly related to age, which is the logic underpinning cultural ages, and also weakly related to cumulative constructs,
which is the logic underpinning soft stages. Notwithstanding the need to rule out alternative explanations discussed in Study 1A, these findings strengthen the conclusion that Lectical Scores are a sound reflection of hierarchical complexity.

The regression models on SCT Rasch scores reveal a different story. Loevinger argued that “If all we want is the endgame, a Total Protocol Rating, we can just set up a regression equation with measures of age, intelligence, SES, and so on. No need to give an SCT. The purpose of the SCT measure of ego development was not just to arrive at some kind of score or prediction. It is a source of insight into personality development in general and of the individual subject” (Loevinger, 1998a, p. 348). Loevinger may be correct that SCT scores provide insight into general personality development. However, she may also be confessing that ego development cannot be clearly distinguished from other facets of the human condition because SCT scores can be derived from an amalgam of other variables.

Unique perspectives and unique words increased the predictive power of the regression models on the SCT Rasch scores. This may account for the finding that when SCT scores are factor analysed, the first principal component only accounts for approximately 20% of the variability in the data, as discussed in Chapter 2. It may also explain the moderate amount of multidimensionality identified in some subsamples in Study 1B. Not only this, but unique perspectives and unique words were found to be stronger predictors of SCT Rasch scores than CLAS scores. SCT Rasch scores were more strongly related to cumulative constructs, which are definitional of soft stage models. These findings corroborate those presented in Studies 1B, 2A, and 2B, in which SCT scores were tentatively suggested to be a stronger reflection of a cumulative ordinal scale.

Age could not be included in the models because of missingness, so temporal factors may have increased the predictive power of both models. Given the nonsignificant bivariate correlations between age, CLAS scores and SCT Rasch scores, this does not seem particularly tenable. Also, Step 3 of the regression model on SCT Rasch scores only accounted for 53% of the data’s variability, leaving 47% unexplained, which allows for the possibility that other variables may be more strongly
related to SCT Rasch scores than those included in this study. This is an important consideration because even with a moderately sized sample and the inclusion of three theoretically relevant predictors, the model only accounted for approximately half of the variability in the data.

It seems reasonable to conclude that findings corroborate those reported in previous studies. Scores on the Lectical Scale are weakly related to temporal change and cumulative constructs and seem to be a sound reflection of hierarchical complexity. In contrast, the hard stage claims about SCT scores do not seem defensible.
Synthesised Discussion on Ego Development

Summary of Findings

SCT scores are contended to operationalise nine hierarchically integrated stages that the ego may traverse over the lifespan as it constructs increasingly complex worldviews. Despite this contention, patterns observed in SCT scores appear distinctly different to those observed in scores that have been shown to operationalise hard stages. Findings from Studies 1B, 2A, 2B, and 2C drew upon different samples and research methods yet converged on the conclusion that scores operationalise approximately four or five cumulative stages that are moderately related to hierarchical complexity and more strongly related to the number of perspectives and/or words used by test-takers. Scores seem to be a stronger reflection of the cumulative breadth—but not the hierarchical complexity—of test-takers’ responses. Why have other scholars not arrived at similar conclusions in previous research? There may be several ways of accounting for this.

Potential Explanations

First, the research methods used to test hard stage requirements have been relatively limited. Based on a literature review of over 500 publications, Cook-Greuter’s (1999) research seemed to be one of the few attempts to explicitly test these requirements. As discussed in Study 2A, her conclusions depended on visual analysis of SCT distributions and what appeared to be a logical fallacy. Her conclusions were not supported when her data were reanalysed using quantitative procedures.

Second, ego developmentalists may have overgeneralised the applicability of hard stage requirements to broader personality structures i.e., the ego/worldview, to which they may not apply. As discussed in Chapter 2, the ego development scoring system enumerates an extensive list of manifestation of ego development in protocols. These include feelings, cognitive style, conscious
preoccupations, internal dimensions, problem-solving, decision-making styles, organisational types, defence mechanisms, counselling style, main anxiety, character of depression, language clues, impulse control, interpersonal style, conception of truth, etc. (Cook-Greuter, 2002). The breadth of these manifestations is consistent with the aspiration to characterise a construct as broad as a worldview, and the general trend for SCT scores to be weakly to moderately correlated with almost all other psychological variables against which they have been tested (Blasi, 1993; Broughton & Zahaykevich, 1988; Hauser, 1976, 1993; Loewinger, 1979, 1993b; Manners & Durkin, 2001).

However, integrating a multitude of psychological dimensions into a scoring system does not seem to result in sound measurement of a more holistic set of hard stages. Findings from this thesis suggest that it results in an increasingly integrative conception of the self that is decreasingly related to hierarchical complexity. Not only does this observation apply to the higher stages of ego development, but it also seems to apply across the entire ego development scale. Relatively few dimensions listed above have been shown to develop hierarchically or even covary. As a result, integrating all these dimensions into a scoring system is conceptually and empirically questionable and may explain why the unidimensionality of SCT scores is compromised. This situation may be analogous to a doctor who tries to express all biometric information about their clients (e.g., height, weight, heart rate, etc.) in millimetres of mercury (mmHg), even though mmHg is typically only used to express units of blood pressure. The many manifestations of ego development may be critical to developing an integrated conception of the human condition, but it does not follow that all these manifestations can be calibrated along a common scale. Rather, each manifestation may be deserving of its own unique scale as discussed by other developmentalists such as Noam (1993). As a result, there may be a fundamental misfit between some of the constructs considered by ego development theory and its scoring system (e.g., main anxiety) and the properties of a hierarchically integrated scale of development. While some scholars have explored the applicability of The Model of Hierarchical Complexity to diverse psychological constructs such as empathy (Commons & Wolfsont, 2002), it is unclear whether these stages of complexity characterise third-person
conceptions about empathy or first-person phenomenological experiences of it. Loevinger (1993a, p. 61) addressed the possibility that different manifestation of ego development needed to be considered separately by stating that “the question as to whether or to what extent ego development is composed of separable strands of development that can be assessed separately is open—empirically, theoretically, and metatheoretically.” This line of argumentation may support Noam’s (1993, p. 48) general position that “[it] is this conceptual problem—that complexity and maturity are built into one stage model—that is exposed when we study the relation between ego development and psychopathology.” This line of argumentation may also support aspects of the earlier developmental literature in which it was posited that some psychological processes may not be subject to hierarchical integrations (Commons, Richards, et al., 1982; Labouvie-Vief, 1980, 1982). In the long run, it may turn out that only symbolically mediated aspects of reality such as sensorimotor actions, arguments, and concepts, are subject to hierarchical integrations.

Third, the general acceptance of the hierarchically integrated nature of stages may be a function of what Ross (2008b) refers to as ‘the telephone game’ through which knowledge is distorted by relying on nth-hand information. She provides an example of how The Model of Hierarchical Complexity (Commons & Pekker, 2008) has been used to make false inferences and highlights the ethical implications of this phenomenon. The telephone game may be equally apparent in the ego development literature. For example, one of Loevinger’s (1976, pp. 68-135) main arguments for the hierarchically integrated nature of ego development stages was based upon relatively superficial similarities between her stages and other stage models. She argued that the “various accounts [i.e., different developmental models] are too different to be collapsed into one account, yet too much alike to refer to altogether different aspects of human nature” (Loevinger, 1976, p. 69). The ego development literature reflects a general acceptance of Loevinger’s assertion. This seems to have influenced practitioners who are promoting the use of ego development theory and measurement to characterise leaders’ ability to navigate organisational complexity, despite the limited evidence to support this contention e.g., Petrie (2014).
Finally, ego developmentalists may have conflated the distinction between sequences and hierarchical integrations. Whereas cultural ages, soft stages, and hard stages are all sequences of sorts, only hard stages are organised by the logic of hierarchical integrations. The literature review from Study 1B and the findings reported in this thesis demonstrate that SCT scores are correctly sequenced, but this is inadequate to infer that scores operationalise hard stages. This is not to suggest, however, that ego development theory and the SCT are unimportant to the field of leadership. SCT scores are related to various leadership phenomena, as discussed in detail in Study 1B. However, the relation between SCT scores and other leadership variables may not be attributable to the SCT’s ability to measure complexity.

Addressing Potential Rebuttals from Ego Developmentalists

The following section foreshadows and addresses anticipated rebuttals about the suggested conclusions drawn in this thesis. The first potential rebuttal is that ego development is more inclusive than cognitive development. Therefore, cognitive-developmental scores cannot be used as a criterion against which SCT scores can be tested. “In contrast to this, self-inquiring, full-bodied developmental theories, which focus on meaning rather than mechanics of cognition, and which include cognitive, affective and behavioural components, tend to assess their own methods and arguments for hidden assumptions and limitations” (Cook-Greuter, 2000, p. 10). This rebuttal reflects a fundamental misunderstanding of how cognitive-developmentalists conceive of cognition. Whereas cognition is defined more narrowly in other fields of enquiry (M. D. Mumford et al., 2015), cognitive-developmentalists in the Piagetian and neo-Piagetian traditions have conceived of cognition as an integrative construct which integrates thought, emotion, motivation, meaning, and action (Fischer & Bidell, 2006; Piaget, 1968; Snarey et al., 1983). Cognitive-developmentalists may conceive of ego development as cognitive theory precisely because it strives to integrate thought, affective, and behavioural components. The LAS awards scores based on two definitional properties—hierarchical orders of abstraction and logical structure of arguments. Orders of
abstraction reflect the implicit amount of hierarchical complexity associated with the meaning of concepts. LAS and CLAS scores cannot be seriously considered to focus on the mechanics of cognition at the expense of meaning-making. Lectica’s approach to attributing orders of abstraction to the meaning of concepts may have something in common with the considerations that are explicitly acknowledged in the ego development literature. For example, Cook-Greuter (1990) recognises that some phenomena only become available at certain ego development stages because of the magnitude of development they reflect e.g., shame emerging at the Opportunist / Diplomat stages and guilt emerging at the Expert / Achiever stages. If ego developmentalists pursue this rebuttal, then they would need to identify specific psychological processes that have been shown to be subject to hierarchical integrations and are included by ego developmental theory but are not accounted for by the integrative conception of cognition that is typically promoted by cognitive-developmentalists.

The second potential rebuttal is that cognitive-developmental approaches are limited by their focus on reasoning. However, the higher stages of ego development transcend reasoning skills in favour of other psychological faculties such as deeper insight and wisdom. As a result, cognitive-developmental scores cannot be used as a criterion against which SCT scores can be tested. “Because they take the representational, symbolically mediated reality as the only kind of reality that exists, they are limited to assessing the structural complexity of rational thought. They cannot detect its fundamental limitations. As a result, purely cognitive approaches tend to privilege intellectual prowess, logical sophistication, complexity and formal aspects of meaning making at the expense of deeper insight and wisdom” (Cook-Greuter, 1999, p. 178). As discussed above, this rebuttal fails to reflect a sound understanding of the cognitive-developmental approach, in which cognition is conceived to be an integrative process that interweaves thought, emotion, motivation, meaning, and action. Cognitive-developmentalists are likely to conceive of deeper insight and wisdom as manifestations of cognition. Whereas some leaders may value deeper insight and wisdom (Cook-Greuter, 2000), it does not follow that insight and wisdom are the products of
hierarchical integrations or that their emergence can be located along hard stage scales. If ego developmentalists wish to measure deep insight and wisdom, they may be better served by identifying a set of items that measure personal maturity and/or existential depth, rather than hierarchically integrated forms of complexity.

The third potential rebuttal is that cognitive-developmental assessments measure narrower constructs and will, therefore, consistently display stronger psychometric properties than those that measure broader constructs. As a result, it may be unfair to compare the psychometric properties of a cognitive-developmental assessment with those of an ego development assessment. “Cognitive theoreticians can rightfully claim that their measures are more robust in their narrowly defined domains than measures that try to assess broader areas of human being. This is so because, the narrower the focus in general, the easier it is to create tasks to measure the target features in question and, if well done, the better the statistical outcome” (Cook-Greuter, 2013, p. 12). However, this rebuttal does not reflect an accurate understanding of metrology or psychometric procedures used to evaluate developmental assessments. Assessments of broad constructs such as literacy and numeracy have been shown to display strong psychometric properties (Cavanagh & Waugh, 2011; Irwin & Irwin, 2005; Weller et al., 2013), and assessments of narrower constructs such as the self-efficacy of nurses have been shown to display weaker psychometric properties (Hagquist et al., 2009). It does not stand to reason that assessments that measure broader constructs display weaker measurement properties than those that measure narrower constructs. Psychometric properties are primarily determined by how successfully a set of items function together to measure a unidimensional construct which can be used to make distinctions between persons (Andrich, 1988; Rasch, 1980). Assessments that measure broader constructs, such as ego development, require a greater number of items to adequately reflect the breadth of the construct in question. However, assessments with a greater number of items, such as the SCT, are likely to exhibit higher reliabilities and therefore display stronger measurement properties than assessments with fewer items (Traub & Rowley, 1991). This thesis does not depend on the psychometric properties of LAS and SCT scores.
Studies 2A, 2B, and 2C do not employ psychometric procedures but still converged on the conclusion that SCT scores failed to satisfy most hard stage requirements.

The fourth potential rebuttal is that while approaches to measuring cognitive development may satisfy hard stage requirements, this may present a barrier to their application to leadership studies. “We propose that the work of vertical leadership development (VLD) is sometimes off-balance in certain ways. This shows up as confusion, complaints, critiques, resistance, or outright failure. Our colleagues may point out, for example, that VLD is ‘too complex’, ‘hierarchical and judgmental’, ‘too Western’, or that ‘stage change takes too long’. The list goes on. The points have merit.” (Palus et al., 2020, p. 60). This rebuttal may be the function of an appeal to emotion fallacy. Even if cognitive-developmental approaches are ‘hierarchical and judgmental’, this has no bearing on the definitional attributes of theory or the psychometric properties of assessments. As discussed in Chapter 1, approaches to measurement that successfully operationalise hard stage requirements may account for a unique proportion of variability in leadership phenomena that cannot be readily accounted for via other approaches. This rebuttal might be readily addressed by outlining the affordances of hard stage models and associated measurement approaches.

The fifth potential rebuttal is that because ego development stages include cognitive, emotional, and behavioural growth (Cook-Greuter, 1999), we cannot expect all aspects of ego development stages to correspond to cognitive-developmental stages. If correspondences were obtained, they would only occur between the cognitive facet of ego development and cognitive-developmental stages. However, this rebuttal is not available to ego developmentalists because a distinct cognitive factor has not been identified in the literature. As discussed in Chapter 2, factor analyses suggest that stems tap a single construct, i.e., ego development (Loevinger & Wessler, 1970). Structural equation models have not supported the claim that ego development subsumes cognitive development (Novy et al., 1994), and factor analyses that have revealed more than one factor have not been interpreted to suggest a distinct cognitive factor (Torbert & Livne-Tarandach, 2009). If the ego is conceived to be a structural unity, then cognition cannot be regarded as a
separable aspect of growth because egos weave cognitive, emotional, and behavioural processes together in a unified structure. This point is also related to the potential counter-argument that ego development stages do not ‘just reflect’ an increase in hierarchical complexity, but they also reflect an increase in a multitude of other constructs e.g., emotional maturity, defence mechanisms, etc. (Cook-Greuter, 1990, 2013). However, accepting that ego development stages do not ‘just’ increase in hierarchical complexity is to accept that stages do increase in hierarchical complexity amongst other constructs, and this necessitates that Stage S+1 must be more hierarchically complex than Stage S. Based on the literature presented in Chapter 2 and the findings reported in this thesis, this is potentially a challenging position to defend.

The sixth potential rebuttal that may be levelled against the suggested conclusions draw in this thesis is related to the general argument that people cannot grasp concepts that transcend their level of ego maturity (Blumentritt et al., 1996; Jurich & Holt, 1987; Redmore, 1976). However, this rebuttal would reflect a presuppositional fallacy by assuming the validity of the developmental stages that the findings reported in this thesis call into question. This rebuttal would also constitute an ad hominem logical fallacy that overlooks the scholarly arguments that have been laid out in this thesis. Broader implications, limitations, and future research directions are discussed in Chapter 6.
Chapter 4: The Complexity Gap Between Leader Reasoning Skills and the Task Demands of Their Roles

Given the properties displayed by cognitive-developmental scores in Studies 1 and 2, they were used as the basis for analysis in the context of Studies 3 and 4. In study 3, cognitive-developmental scores were used to compare the hierarchical complexity of leaders’ reasoning skills to the task demands of their roles. In study 4, cognitive-developmental scores were used to evaluate longitudinal growth over time.

Introduction to the Complexity Gap

The complexity gap occurs when the hierarchical complexity of leaders’ reasoning skills is significantly lower than the hierarchical complexity of the task demands that characterise their roles. In this context, tasks are downward assimilated as the requisite complexity of tasks is reduced to a lower level, and the inherent complexity of tasks is not solved (Dawson & Stein, 2008). Piaget’s famous experiment on the conservation of volume which is acquired at concrete operations vividly illustrates downward assimilation and the complexity gap (Piaget, 1954; Piaget & Szeminska, 1941/1952). A child is presented with two equally sized glasses containing equal amounts of liquid. The liquid from the second glass is poured into a third glass which is taller and narrower than the first and second glasses respectively. Children with less developed reasoning skills at Lectical Level 7 assert that the third glass (which is taller and narrower) now contains more water than the first glass
(which is shorter and wider). They arrive at this conclusion because they reason in a linear way which directly maps the height of the water in the narrower glass to a false conclusion about its volume. These children have downward assimilated the inherent complexity of the task and have unwittingly demonstrating the complexity gap. Children with more developed reasoning skills at Lectical Level 8 assert that the third glass (which is taller and narrower) contains the same amount of water as the first glass (which is shorter and wider). They arrive at this conclusion by constructing a system of conceptual relations about the height and width of the third glass to reach a true conclusion about its volume. These children do not exhibit the complexity gap because their reasoning satisfies the inherent complexity of the task.

This example pertains to child development, but it is analogous to the complexity gap that is likely to be experienced by leaders who display reasoning skills that fail to satisfy the task demands of their roles. Leaders who are aware of this inability may experience a sense of being ‘in over their heads’ (Kegan, 1994), while leaders who are unaware of this inability may be oblivious to their underperformance (Jaques, 1976, 2006). Arguments about the complexity gap are found in the philosophical, management and leadership, and developmental psychological literature. Each body of literature will be reviewed in turn.

**Philosophical Arguments**

Arguments pertaining to the complexity gap date back to early Greek Philosophy. For example, in The Republic, Plato (pp. 427d-445e) recounts a conversation between Socrates and a fellow Athenian, Adeimantus. Socrates argued for the existence of the complexity gap by providing examples in which citizens lacked the requisite knowledge and reasoning skills to participate in a democratic society.

In the second half of the 20th century, several philosophers contributed to the intellectual discourse about the complexity gap. American scholar Daniel Bell (1973) predicted that post-industrial societies would exceed the complexity of systems available to manage them. He argued
that post-industrial societies would emphasise service-based industries, science-based sectors, and the rise of technological elites. He predicted that these characteristics would result in more complex and differentiated workplaces. However, he argued that “the major intellectual and sociological problems of the post-industrial society are ... those of ‘organized complexity’ — the management of large-scale systems with a large number of interacting variables, which have to be coordinated to achieve specific goals. It is the hubris of the modern systems theorist that the techniques for managing these systems are now available” (Bell, 1973, p. 38).

In a similar vein, German philosopher Jurgen Habermas (1975) predicted that organisations and their leaders would suffer from a ubiquitous challenge in post-industrial societies, which he referred to as the legitimation crisis. Because of the rate at which economic, political, and socio-cultural systems have evolved, he contended there would be a lack of capability to establish systemic structures that would be required to achieve organisational outcomes. He anticipated conditions in which institutions would retain their legal authority while citizens’ confidence in their legitimacy declined.

**Arguments Informed by the Management and Leadership Literature**

The management and leadership literature has made a modest contribution to our understanding of the complexity gap, perhaps because of its limited integration with the field of complexity science. This literature is limited to consideration of perceived, inferred, and conceptual capability gaps, rather bodies of empirical literature that directly establishes its existence. Perceived capability gaps are typically identified by survey-based research which summarises followers’ perceptions of leaders’ gaps. Inferred capability gaps are typically described by research which infers that leaders are likely to exhibit gaps because they possess low levels of a construct that has been linked to favourable leadership outcomes such as intelligence. The possibility of a capability gap has been surfaced by relatively nascent theorising about constructs such as requisite complexity (Lord et al., 2011). These three types of research are reviewed in turn.
Perceived Capability Gaps

Perceived capability gaps are identified by survey-based research which summarises followers’ perceptions of leaders’ gaps. For example, the Karpin report was a landmark study on the inadequacy of Australia's leadership at the turn of the 20th century (Karpin, 1995). The report argued that while effective leaders were necessary for a more competitive economy and higher-performing organisations, the leadership landscape was being transformed because of globalisation, technology, and the need for innovative products and services. This landscape required a new paradigm of leadership that was characterised by more adaptive and inclusive ways of working. The report presented evidence that most Australian leaders lacked the requisite education and skills to foster this new paradigm and surfaced perceived gaps in entrepreneurship, global orientation, soft skills, strategic skills, and management development. Recommendations included reforming management education, achieving enterprise best practice management development, and capitalising on diversity. It indicated that Australia may risk limited competitiveness in the global marketplace without a fundamental reform to leader development. Subsequent reports acknowledged the robustness of Karpin’s recommendations but argued that insufficient attention has been placed on implementing the recommendations (ISBA, 2011).

The Study of Australian Leadership was the first major review of Australian leadership since the Karpin report (Gahan et al., 2016). It surveyed almost 8000 people across 2,703 organisations, from senior executives to front-line employees. The report found that over 40% of Australian organisations failed to meet their performance targets, and key gaps in leadership capabilities needed to be addressed. Leadership and management practices were positively correlated with several organisational outcomes, including financial performance, innovation, culture, and engagement. Leaders rated their workplaces in terms of various leadership and management practices, including performance monitoring, the quality of performance targets, and the use of incentives on a scale that ranged from .0 (least capable) to 1.0 (most capable). Results suggested that different industry sectors ranged from .42 to .66, with a median score of .51. This was
interpreted to indicate a moderate degree of perceived leadership and management capability. Perceived capability gaps were identified in various leadership fundamentals, including effectively addressing workplace issues, using key performance indicators (KPIs), reward and recognition, and consequence management. The report argued that significantly more resources were needed to address these capability gaps, particularly at the front-line leadership level.

An extensive literature review suggested that negligible research has been undertaken on perceived capability gaps over the recent past. The few studies identified present evidence that is generally consistent with earlier findings. For example, gaps are perceived to occur for general leadership attributes (Bunker et al., 2021), specific industries (Weese, 2005), roles (Graham, 2020), and for topics such as gender balance and diversity (Kim et al., 2021; Tarbutton, 2019).

**Inferred Capability Gaps**

Inferred capability gaps are typically described by research which infers that leaders are likely to exhibit gaps because they possess low levels of a construct that has been linked to favourable leadership outcomes. For example, general intelligence has been shown to predict leader effectiveness, particularly when leaders occupy complex roles (Salgado et al., 2003; Schmidt & Hunter, 1998; Schmidt et al., 2008). Meta-analyses generally reveal a modest correlation of approximately .17 between intelligence and perceived leader effectiveness, but a somewhat stronger correlation of .33 between intelligence and objective measures of performance (Hoffman et al., 2011; Judge et al., 2004). Despite these mixed findings, intelligence appears to among the strongest predictors of leaders’ performance. For this reason, some scholars advocate for using intelligence as a selection criterion in recruitment processes because lower levels of intelligence may imply a cognitive gap which is likely to impact performance (Antonakis et al., 2020).
**Conceptual Capability Gaps**

The possibility of a capability gap has been surfaced by nascent theorising about leader complexity. Many of these conceptions are based on the premise that leaders must integrate various psychological, behavioural, and social factors to successfully navigate organisational complexity.

Argyris and Schon (1977) were among the first management theorists to distinguish between two mental *theories of action*. Model 1 focused on achieving behavioural outcomes by maximising winning, being rational, and reducing emotionality. Model 2 focused on achieving psychological and behavioural outcomes through informed choice and environmental monitoring. Argyris contended that Model 2 was qualitatively different and more inclusive than Model 1. Furthermore, these theorists argued that leaders who displayed Model 1 were likely to exhibit gaps relative to the requirements of the contemporary workplace.

In a similar vein, American systems theorist Peter Senge (1990) delineated between linear thinking and systems thinking. Linear thinking consists of causal chains that are close in space and time and often leads to simplistic inferences about the cause of workplace dilemmas. Senge argued that this would be inadequate to address the inherent complexity of contemporary workplaces because linear thinkers typically reach for low-leverage interventions. On the other hand, systems thinking focuses on emergent wholes and dynamic patterns, rather than static snapshots. Systems thinking enables complex workplace dilemmas to be perceived in a multivariate manner, thereby increasing the opportunity for strengthened leader effectiveness.

The *Leaderplex Model* is a more advanced iteration of these earlier conceptions (Hooijberg et al., 1997). According to this model, leaders must integrate cognitive, social, and behavioural complexity to demonstrate effective leadership in the face of organisational complexity. Cognitive complexity was defined in terms of differentiation and integration of information. Social complexity was defined in terms of differentiation and integration of social roles. Behavioural complexity was defined in terms of the breadth of leaders’ behavioural repertoire. Cognitive and social complexity
were conceived to be precursors of behavioural complexity, and behavioural complexity was predicted to have a direct causal impact on leader effectiveness. The Leaderplex Model provides a sound conception of the attributes which may contribute to the emergence of leader effectiveness, but it has motivated little empirical research to date. Despite this, Day and Lance (2004) argue that the three main elements of the Leaderplex Model may play an important role in a broader conceptualisation of leadership, in which the processes of differentiation and integration underpin the process of leader development. This is consistent with the position advanced by other scholars who have argued that differentiation and integration are fundamental to some forms of human development (Kjellström & Sjölander, 2014), and also play a critical role in accounting for a variety of developmental phenomena in various philosophical and scientific disciplines (Akrivou, 2008; Johnson, 2000; Kolb, 1984).

Building on the Leaderplex Model, Hannah et al. (2009) presented a conception of leader complexity. This was hypothesised to predict leaders’ ability to navigate organisational challenges and influence followers. They posited that leader complexity emerges when leaders’ self-concept is complex and positive. This requires integrating various psychological processes, including cognition, affect, goals, values, expectancies, and self-regulatory plans. Effective leadership may emerge when leader complexity interacts with the social context and the task demands of leaders’ roles. Like the Leaderplex Model above, leader complexity provides a useful conceptualisation, but it has not been used to empirically test for the presence of the complexity gap.

Most recently, requisite complexity has been hypothesised to define leaders’ ability to exhibit responses that are congruent with complex organisational challenges (Hannah et al., 2011; Lord et al., 2011). Requisite complexity consists of two elements: static and dynamic. Static requisite complexity consists of four stable attributes which leaders recruit when responding to organisational challenges. These four attributes are general cognitive complexity to process abstract information; affective complexity to be sensitive to emotions; social complexity to integrate multiple social roles; and self-complexity to draw upon one’s self-concept to mediate between intra and interpersonal
processes. Dynamic requisite complexity consists of leaders’ ability to construct in-the-moment responses to organisational challenges by integrating diverse information. Diverse information includes the nature of the task, social context, and other information provided by an internal set of mental structures. Dynamic requisite complexity leads to the emergence of adaptive structures in the brain, which are recruited to comprehend the social context and act. Static and dynamic complexity are hypothesised to influence whether the leaders’ responses are congruent with the complex challenges they are presented with.

**Arguments Informed by Developmental Psychology**

A few developmental psychologists have begun to quantify the complexity gap. Elliott Jaques (1976, 2006) was one of the first scholars to conceive of *requisite fit*, which he defined as the fit between the complexity of work and the capability of the role holder. According to Jaques, requisite fit was the strongest predictor of workplace performance. He claimed that lack of fit lead to limited effectiveness and unintended business consequences. Jaques predicted that if work complexity exceeds the level of capability of a role holder, then consequences might include burnout, low retention, low job engagement, simplified decision-making, and the experience of information overload. On the other hand, if the level of capability of a role holder exceeds work complexity, he predicted overly complicated decision-making, disengagement with detail, and boredom would ensue. Although requisite fit is a useful conception, empirical evidence on the percentage of leaders who fall below, within, and above work complexity ranges could not be identified. Some scholars have applied the Hierarchical Complexity Scoring System and the ego development scoring system to analyse Jaques’ conception of work complexity (Törnblom et al., 2018). Analyses were undertaken on the requisite level of cognitive complexity and meaning making at multiple organisational layers. Preliminary analyses suggest that the demands of cognitive complexity occur between stages 9 and 12 (which may loosely correspond to Lectical Levels 9 to 12), and the demands of meaning making occur between the Diplomat and Strategist stages on the ego
development scale. Notwithstanding the possibility that requisite fit may contribute to our understanding of the complexity gap, Jaques’ conception of work complexity has been criticised by developmentalists for having limited applicability to organisations that lack hierarchical structures and also because of the limited validation of his conception of work complexity e.g., number of organisational layers that are accounted for, the cognitive requirements at successive organisational layers, etc. (Törnblom, 2018).

Developmental psychologist Robert Kegan was one of the first scholars to apply developmental stages—which he refers to as *orders of consciousness*—to characterise the gap between the demands of leaders’ roles and their actual mental ability (Kegan, 1994). He argued that the complexity gap would lead to the experience of being ‘in over one’s head’. Kegan’s theory consists of stages through which the mind develops (Kegan, 1982; Kegan & Lahey, 2009). The first stage is called the impulsive mind and is driven by basic impulses and perceptions. The second stage is called the imperial mind and is driven by needs, interests, and wishes. The third stage is called the socialised mind, it is driven by interpersonal mutuality, and is shaped by the internalisation of social expectations. The fourth stage is called the self-authoring mind, it is driven by the need for self-authorship, identity, and ideology, and is shaped by an internal sense of personal authority. The fifth stage is called the self-transforming mind, it is driven by the need for interindividuality, and is shaped by the ability to perceive one’s own ideological limitations, and coordinate multiple systems of ideas simultaneously. Kegan (1994) contends that approximately one-third of adults are in transition between the third and fourth orders, and another third have been located at the fourth order. In his earlier writing, Kegan argued that “At any given moment, around one-half to two-thirds of the adult population appear not to have fully reached the fourth order of consciousness” (Kegan, 1994, pp. 188-191). However, later in his writing he argued that team members who are predominantly at the third stage need to develop to the fourth order, and leaders who are predominantly at the fourth order need to develop to the fifth order. By doing so, not only would leaders display self-authorship but they would also perceive the limitations of their ideologies and
continuously re-author them accordingly (Kegan & Lahey, 2009). Thus, Kegan contends that the size of the complexity gap between requisite and actual mental abilities is one full stage. Although the direction of Kegan’s research is promising, he offers relatively few details of how the requisite complexity of roles was established, and also the reliability and validity of his approach to measurement (Kegan, 1994; Lahey et al., 2001).

Finally, Dawson and Stein (2004b) conducted a study on the relation between the hierarchical complexity of leaders’ decision-making skills and the task demands of their roles at various management layers. A total of 512 leaders across four management layers in the federal government were assessed via the LDMA between 2002 and 2006 as part of a series of cognitive-developmental assessments. Figure 36 depicts the relation between the task demands of leaders’ roles (in teal) and the hierarchical complexity of their reasoning skills (in ochre). A widening gap between can be observed at successive management layers. This research is promising, but only visual analyses were undertaken. It is unknown whether similar patterns would be observed in reasoning skills applied to domains other than decision-making. Study 3 attempts to resolve these issues in order to determine how the hierarchical complexity of leaders’ reasoning skills is related to the task demands of their roles.
Figure 36

**Average Lectical Score and Role Complexity by Management Layer**


**Hypotheses for Study 3**

Hypotheses are based on the previous literature review, paying particular attention to the research undertaken by developmental psychologists.

- **Hypothesis 1**: There is a statistically significant difference in the hierarchical complexity of reasoning skills between leaders at increasingly senior management layers i.e., from mid-leaders to upper leaders to senior leaders to executive leaders. Management layers are defined as layers within an organisational structure at which leadership roles can be located. In the context of this thesis, leadership roles at four management layers are considered: mid-leader, upper leader, senior leader, and executive leader. These are defined in detail below.

- **Hypothesis 2**: The hierarchical complexity of leaders’ reasoning skills is significantly lower than the task demands of their roles—but only for upper leaders, senior leaders, and executive leaders.
Method

Participants

Lectica had previously collected the data between the years 2006 and 2016. The original sample was extracted from Lectica’s database of over 40,000 cognitive-developmental assessments and consisted of 3661 Lectical Assessments completed by 1172 leaders. It included all persons who occupied a leadership role, had participated in a leader development initiative, had completed at least one Lectical Assessment, and had provided sufficient information about their organisational context for Lectica to estimate role complexity. Leaders consented to their data being shared with third parties for legitimate research purposes via Lectica’s registration process which includes privacy and legal statements (refer to Appendix C). Filters were applied to the original sample to yield an analysable sample and thus, the sampling process was not entirely random. The analysable sample consisted of 692 leaders who had completed 1709 Lectical Assessments at test-time 1. When data were collected, test-takers were provided with a binary option to select for sex. Demographic details for these leaders are provided in Table 41.
Table 41

Demographic Details of Leaders for Study 3

<table>
<thead>
<tr>
<th>Demographics</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leadership provider (geography, sector)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (Australia, public)</td>
<td>4</td>
<td>0.58</td>
</tr>
<tr>
<td>3 (North America, private)</td>
<td>20</td>
<td>2.89</td>
</tr>
<tr>
<td>4 (Multinational, private)</td>
<td>18</td>
<td>2.60</td>
</tr>
<tr>
<td>6 (North America, public)</td>
<td>7</td>
<td>1.01</td>
</tr>
<tr>
<td>7 (Australia, private)</td>
<td>7</td>
<td>1.01</td>
</tr>
<tr>
<td>8 (North America, public)</td>
<td>539</td>
<td>77.89</td>
</tr>
<tr>
<td>9 (Australia, private)</td>
<td>8</td>
<td>1.16</td>
</tr>
<tr>
<td>10 (North America, private)</td>
<td>2</td>
<td>0.29</td>
</tr>
<tr>
<td>12 (Australia, private)</td>
<td>21</td>
<td>3.03</td>
</tr>
<tr>
<td>13 (Australia, private)</td>
<td>14</td>
<td>2.02</td>
</tr>
<tr>
<td>14 (Australia, private)</td>
<td>9</td>
<td>1.30</td>
</tr>
<tr>
<td>15 (North America, private)</td>
<td>14</td>
<td>2.02</td>
</tr>
<tr>
<td>16 (North America, private)</td>
<td>28</td>
<td>4.05</td>
</tr>
<tr>
<td>17 (Multinational, private)</td>
<td>1</td>
<td>0.14</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>445</td>
<td>64.31</td>
</tr>
<tr>
<td>Female</td>
<td>216</td>
<td>31.21</td>
</tr>
<tr>
<td>Unidentified</td>
<td>31</td>
<td>4.48</td>
</tr>
<tr>
<td>First language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>585</td>
<td>84.54</td>
</tr>
<tr>
<td>Not English</td>
<td>58</td>
<td>8.38</td>
</tr>
<tr>
<td>Unidentified</td>
<td>49</td>
<td>7.08</td>
</tr>
<tr>
<td>Management layer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Layer 2 mid-leader</td>
<td>245</td>
<td>35.40</td>
</tr>
<tr>
<td>Layer 3 upper leader</td>
<td>295</td>
<td>42.63</td>
</tr>
<tr>
<td>Layer 4 senior leader</td>
<td>79</td>
<td>11.42</td>
</tr>
<tr>
<td>Layer 5 executive leader</td>
<td>73</td>
<td>10.55</td>
</tr>
</tbody>
</table>

Materials

As discussed in Chapter 2, Lectica offers a suite of cognitive-developmental assessments called Lectical Assessments that assess the hierarchical complexity of reasoning in different domains. A variety of Lectical Assessments were administered to leaders, including the LDMA, the Lectical Leadership Reasoning Assessment (LLRA), the Lectical Self-Understanding Assessment (LSUA), the Lectical Ethical Reasoning Assessment (LERA), and the Lectical Reflective Judgement Assessment.
A detailed description of Lectical Assessments, including a sample dilemma and a list of items, is provided in Chapter 2, Study 1A. The number of assessments administered by management layer at test-time 1 is presented in Table 42.

**Table 42**

*Number of Lectical Assessments Administered by Management Layer*

<table>
<thead>
<tr>
<th>Management layer</th>
<th>LDMA</th>
<th>LLRA</th>
<th>LSUA</th>
<th>LERA</th>
<th>LRJA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>154</td>
<td>98</td>
<td>171</td>
<td>61</td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>264</td>
<td>186</td>
<td>252</td>
<td>111</td>
<td>110</td>
</tr>
<tr>
<td>4</td>
<td>77</td>
<td>22</td>
<td>30</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>73</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>568</td>
<td>306</td>
<td>453</td>
<td>189</td>
<td>193</td>
</tr>
</tbody>
</table>

As discussed in Chapter 2, test-takers’ responses can be scored via Lectica’s two scoring systems—the LAS and the CLAS. The purpose of both systems is to locate the magnitude of hierarchical complexity exhibited in samples of reasoning along the Lectical Scale. CLAS was used to score leaders’ responses to Lectical Assessments for Study 3. As discussed in Chapter 2, LAS and CLAS have a high inter-rater agreement and can therefore be treated as interchangeable for research purposes within certain constraints. Lectical decimal scores have been used in some empirical research (Fuhs, 2016; Van Rossum, 2013). Statistical models were constructed in Mplus (Muthén & Muthén, 1998-2015), and the syntax for all analyses presented in Chapter 4 is documented in Appendix M. Jamovi (2021) was used to produce boxplots.

**Procedure**

**Administration.** Lectical Assessments had been administered through various accredited providers whose identities cannot be disclosed due to contractual agreements with Lectica.

**Selection of Cases.** In addition to awarding a CLAS score, CLAS also awards other scores that can be used as filters. Filters were applied to identify assessments from the original sample that had
been scored with sufficient confidence to include in the analysable sample. Three filters were used in the context of Study 3. The first filter was the difference between each assessment’s CLAS score and LAS score. Assessments with a difference of less than or equal to $\pm 0.3$ were included to ensure sufficient similarity between scores and consistency with the general procedures applied to previous research (Fuhs, 2016; Van Rossum, 2013). The second filter was CLAS’s confidence in the CLAS score based on the density of concepts at successive Lectical Phases. The confidence score ranged from 1 to 10, and assessments with a confidence score equal to or greater than 6 were included. This threshold was considered suitable for research purposes and is consistent with the general procedures applied to previous research (Fuhs, 2016; Van Rossum, 2013). The third filter was the unique word count. A minimum threshold was set for each year of educational attainment to ensure that leaders had elaborated on their responses sufficiently for CLAS to award a score confidently. Assessments with a unique word count equal to or greater than the minimum threshold were included (refer to Table 43).

**Table 43**

*Thresholds Applied to Unique Word Count by Years of Education and Approximate Grade*

<table>
<thead>
<tr>
<th>Years of education</th>
<th>Low threshold</th>
<th>High threshold</th>
<th>Approximate grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>110</td>
<td>800</td>
<td>End of high school</td>
</tr>
<tr>
<td>13</td>
<td>120</td>
<td>1000</td>
<td>Start of university</td>
</tr>
<tr>
<td>14</td>
<td>150</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>150</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>175</td>
<td>1000</td>
<td>Bachelor’s degree</td>
</tr>
<tr>
<td>17</td>
<td>175</td>
<td>1000</td>
<td>Start of Masters</td>
</tr>
<tr>
<td>18</td>
<td>175</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>175</td>
<td>1000</td>
<td>Start of PhD</td>
</tr>
<tr>
<td>&gt; 19</td>
<td>200</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>
Role Complexity Estimation. Role complexity was estimated via a consultative process between Lectica, accredited Lectical consultants, and leaders. Lectica gathers information about leaders’ organisational context, such as the number of layers in the organisational structure, the total headcount, the average number of direct reports, where the leaders’ roles fit into the organisational hierarchy, etc. This information enabled Lectica to match leaders’ roles to a set of general management layers. Lectica used the LAS to identify a role complexity range for each management layer. The role complexity range reflected the hierarchical complexity of task demands that characterised each management layer. Summarised descriptors and role complexity ranges for management layers 2 through 5 are presented in Table 4.
### Table 44

**General Lectical Management Layers 2 Through 5 Including Descriptors and Role Complexity Ranges**

<table>
<thead>
<tr>
<th>Layer No.</th>
<th>Management Layer</th>
<th>Skills required</th>
<th>Lower boundary of role complexity</th>
<th>Mid-point of role complexity</th>
<th>Upper boundary of role complexity</th>
</tr>
</thead>
</table>
| 5         | Executive leader | • Take into account the numerous nested interconnections among variables that characterise highly complex situations  
• Expertly employ iterative, distributed decision-making processes  
• Deftly determine the level of collaboration or perspective seeking that is optimal for a given context | 11.60                             | 11.83                         | 12.05                            |
| 4         | Senior leader    | • Identify multiple relations between nested variables  
• Take change over time into account  
• Consider the level of collaboration and perspective seeking that is optimal for a given context  
• Maintain awareness of common impediments to good decision-making in organizations | 11.40                             | 11.63                         | 11.85                            |
| 3         | Upper leader     | • Work with both individual and group perspectives  
• Identify and attempt to balance competing factors, such as human needs versus organisational needs  
• Identify implicit and contextual causes  
• Identify common impediments to good decision-making in organisations  
• Shift frame of reference to explore alternate ways of seeing a problem | 11.20                             | 11.43                         | 11.65                            |
<table>
<thead>
<tr>
<th>Layer No.</th>
<th>Management Layer</th>
<th>Skills required</th>
<th>Lower boundary of role complexity</th>
<th>Mid-point of role complexity</th>
<th>Upper boundary of role complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Mid-leader</td>
<td>• Identify and seek the perspectives of relevant stakeholders</td>
<td>11.00</td>
<td>11.23</td>
<td>11.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Take a sceptical approach (because there is always bias or faulty information)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Identify multiple relations between variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Identify own frame of reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Attempt to remain objective and impartial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Maintain self-awareness</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Construction of Additional Variables.** Several transformations on management layer were performed in order to construct the statistical models. The first step was to zero all management layers so that management layer 2 became management layer 0, layer 3 became 1, layer 4 became 2, and layer 5 became 3. This meant the intercept was interpretable as the score attained by leaders at management layer 2. The quadratic function was calculated by squaring the value of each adjusted management layer (i.e., management layer^2). The piecewise function was constructed via two pieces with a knot at adjusted management layer 2 (i.e., original management layer 4). The freed function was constructed through four dummy variables.

**Data Structure to Satisfy Statistical Assumptions.** The original sample included assessments administered at test-times 1, 2, 3 and 4. Only assessments administered at test-time 1 were included in the analysable sample to avoid violations to the assumption of independent observations. Some providers (e.g., Provider 8) administered multiple Lexical Assessments at test-time 1 (e.g., LDMA, LLRA, LSUA, LERA and LRJA). A correlation matrix between CLAS scores for different assessments suggested a moderately-high correlation between reasoning skills in different domains, where r ranged from .63 to .75, and r^2 ranged from .40 to .56 (refer to Table 45).
However, the Variance Inflation Factor (VIF) < 10 and Tolerance >.10 which did not indicate significant multicollinearity. Because only management layer 5 took the LDMA, each Lectical Assessment was analysed separately to avoid conflating the domain of reasoning with management layer.

Table 45

*Descriptive Statistics and Correlations Between Lectical Assessments Administered at Test-Time 1 for Provider 8*

<table>
<thead>
<tr>
<th>Lectical Assessment</th>
<th>n</th>
<th>M</th>
<th>s²</th>
<th>LDMA</th>
<th>LLRA</th>
<th>LSUA</th>
<th>LERA</th>
<th>LRJA</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDMA</td>
<td>410</td>
<td>11.31</td>
<td>0.036</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLRA</td>
<td>306</td>
<td>11.34</td>
<td>0.029</td>
<td>.68**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSUA</td>
<td>453</td>
<td>11.23</td>
<td>0.036</td>
<td>.63**</td>
<td>.69**</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LERA</td>
<td>189</td>
<td>11.35</td>
<td>0.030</td>
<td>.68**</td>
<td>.71**</td>
<td>.70**</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>LRJA</td>
<td>193</td>
<td>11.22</td>
<td>0.043</td>
<td>.69**</td>
<td>.71**</td>
<td>.70**</td>
<td>.75**</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note.* *p*<.05, **p**<.001

**Modelling Approach.** CLAS awards a single score to each assessment rather than a unique score to each item. For this reason, regression models on manifest variables were used, rather than a combination of measurement and structural models. This approach imposes some limitations on the ability to account for error variance. However, this is justified based on the measurement properties displayed by the LDMA in Study 1A.
Results

Separate analyses were conducted on the LDMA (n=568), LLRA (n=306), LSUA (n=453), LERA (n=189), and LRJA (n=193). Analyses reported below were conducted on LDMA at test-time 1 because this sample had the largest number of cases, and it was the only assessment completed by leaders at management layer five. Similar analyses were conducted on the other assessments and are presented in a more abbreviated format in Appendix N.

Hierarchical Clustering in the Data Structure

The data were nested at multiple layers, so the amount of variability attributable to hierarchical clustering needed to be determined. Leadership providers were considered to be the primary sampling unit (PSU), however hierarchical clustering was examined for the three clusters that were most likely to have impacted leaders’ scores i.e., leadership providers, leader development program, and specific training cohort. Hierarchical clustering was analysed via two methods—intraclass correlations (ICCs) and the design effect (Deff). ICCs partition variance at the within and between levels and then determine the proportion of within variance relative to the total variance (i.e., within and between). Deff indicates the magnitude of bias in the standard errors (SEs) when the data are not independent such that $\text{Deff} = 1 + (c-1)\rho$, where $c$ is the average cluster size, and $\rho$ is the ICC. ICCs greater than .05 and Deffs greater than 2 are typically considered to require multi-level modelling. Results are presented in Table 46.
Table 46

Hierarchical Clustering in the Data Structure for Study 3

<table>
<thead>
<tr>
<th>Clustering variable</th>
<th>ICC (p)</th>
<th>Average cluster size (c)</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider</td>
<td>.14</td>
<td>40.57</td>
<td>6.54</td>
</tr>
<tr>
<td>Program</td>
<td>.16</td>
<td>37.87</td>
<td>7.05</td>
</tr>
<tr>
<td>Cohort</td>
<td>.23</td>
<td>17.75</td>
<td>4.87</td>
</tr>
</tbody>
</table>

The ICCs ranged from .14 to .23 and therefore exceeded the critical value of .05. All Deffs exceeded the critical value of 2. As a result, aggregated multi-level modelling was used to obtain unbiased estimates and standard errors. This approach was favoured over disaggregated modelling because analyses were aggregated at the level of CLAS scores, rather than disaggregated by comparing sources of variation at various levels of the data hierarchy. Aggregated multi-level modelling also offered greater flexibility when constructing the model syntax (Preacher et al., 2010). 

*Cohort* was chosen as the clustering variable because it accounted for the greatest variability in CLAS scores. Furthermore, it was the most immediate context in which leaders were nested, so it may have had the greatest impact on their performance.

*Relation between CLAS Scores and Management Layer*

A boxplot of CLAS scores at the four management layers is presented in Figure 37.
As can be seen from the boxplot, CLAS scores were broadly distributed for all four management layers. Despite this, there appeared to be an increase in median scores from mid-leaders to upper leaders to senior leaders, but not from senior leaders to executive leaders. Based on the visual analysis, the relation between CLAS scores and management layer appeared to be non-linear or curvilinear.

**Model Fit**

The relationship between CLAS scores and management layer was analysed by comparing the relative fit of a series of models that successfully deviated more significantly from linearity. These were a linear model, a quadratic model, a piecewise model, and a freed model. The relation between CLAS score and management layer was freed in the freed model. Additional models were explored (e.g., Log10, LogE) but did not display high levels of fit, so they were not explored further. A Maximum Likelihood Robust (MLR) estimator was employed in all cases because of its robustness to non-normality and non-independence.
Given that the models had a different number of equality constraints placed on their parameters, relative fit was examined statistically via the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), the sample-size adjusted BIC, and shared variance ($r^2$) because these fit statistics can estimate the incremental improvement in fit between successive models (Muthén & Muthén, 1998-2015; Vrieze, 2012). The models were saturated, so the chi-square test of model fit could not be used (Muthén & Muthén, 1998-2015). Given that this fit statistic is sensitive to deviations from normality and sample size, it was not considered appropriate to use this statistic to make inferences about model fit (Muthén & Muthén, 1998-2015). For the AIC, BIC, and sample size adjusted BIC, lower scores reflected a stronger fit. Other relative fit indices such as the Tucker Lewis Index (TLI) and Comparative Fit Index (CFI) were less relevant to the current analyses, so they were not used to inform decision-making (Bentler, 1990; Hu & Bentler, 1998). Fit statistics are presented in Table 47.

**Table 47**

*Fit Statistics for the Four Models Applied to Study 3*

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>AIC</th>
<th>BIC</th>
<th>Sample adjusted BIC</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>568</td>
<td>-295.30</td>
<td>-282.28</td>
<td>-291.80</td>
<td>.12</td>
</tr>
<tr>
<td>Quadratic</td>
<td>568</td>
<td>-304.80</td>
<td>-287.43</td>
<td>-300.13</td>
<td>.13</td>
</tr>
<tr>
<td>Piecewise</td>
<td>568</td>
<td>-309.61</td>
<td>-292.24</td>
<td>-304.94</td>
<td>.14</td>
</tr>
<tr>
<td>Freed</td>
<td>568</td>
<td>-307.76</td>
<td>-286.05</td>
<td>-304.94</td>
<td>.14</td>
</tr>
</tbody>
</table>

*Note.* AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

For the linear model (n=568), mid-leaders had a mean CLAS score of 11.27 (SE=0.02, $p<.001$). There was an increase of 0.07 in CLAS scores between successive management layers (SE=0.02, $p<.001$). The residual variance in this model was significant ($s^2_r=0.03$, SE=0.01,
However, the linear model was not a strong fit because the graphical evidence did not appear to be consistent with a linear relation between variables, several fit statistics were higher relative to the other models, and the model accounted for the least amount of variance ($r^2 = .12$).

For the quadratic model ($n=568$), mid-leaders had a mean CLAS score of 11.24 (SE=0.02, $p<.001$). There was an increase of 0.15 in CLAS scores between successive management layers (SE=0.02, $p<.001$). The quadratic function indicated that the increase in CLAS scores was reduced by 0.03 between successive management layers (SE=0.01, $p=.001$). The residual variance in this model was significant ($s^2=0.03$, SE=0.01, $p<.001$). However, the quadratic model was not a strong fit because even though the graphical evidence may be consistent with a quadratic relation between variables, several fit statistics were higher relative to the other models, and the model accounted for the second least amount of variance ($r^2 = .13$).

For the piecewise model ($n=568$), with a knot at the senior leader layer, mid-leaders had a mean CLAS score of 11.24 (SE=0.021, $p<.001$). For piece 1, which occurred between mid-leaders and senior leaders, there was an increase of 0.11 in CLAS scores between successive management layers (SE=0.02, $p<.001$). For piece 2, which occurred between senior and executive leaders, there was a decrease of 0.03 in CLAS scores between successive management layers, but this was not significant (SE=0.03, $p=.20$). The residual variance in this model was significant ($s^2=0.03$, SE=0.01, $p<.001$). The piecewise model was a stronger fit because the graphical evidence was consistent with a piecewise relation between variables, several fit statistics were lower than the other models, and the model accounted for the highest amount of variance ($r^2 = .14$).

For the freed model ($n=568$), mid-leaders had a mean CLAS score of 11.24 (SE=0.02, $p<.001$). Upper leaders had a mean CLAS score that was 0.10 higher than mid-leaders (SE=0.018, $p<.001$). Senior leaders had a mean CLAS score that was 0.22 higher than mid-leaders (SE=0.04, $p<.001$). Executive leaders had a mean CLAS score that was 0.18 higher than mid-leaders (SE=0.04, $p<.001$).
The residual variance in this model was significant ($s^2_r=0.03$, SE=0.01, $p<.001$). The freed model was not a strong fit because the graphical evidence did not appear consistent with a freed relation between variables, and the AIC and BIC were higher than for the piecewise model.

Overall, the piecewise model seemed to be the strongest fit. However, the differences between fit statistics for the four models were relatively small. As a result, pairwise comparisons between successive management layers were undertaken via z-tests. These analyses were based on the Full Information Maximum Likelihood (FILM) estimator and the corrected standard errors (SEs). A Bonferroni correction for three comparisons was applied ($0.05/3 = 0.02$). Analyses indicated that there was a significant increase in CLAS scores of 0.10 between mid-leaders ($n=154$, $M=11.24$, $s^2=0.04$) and upper leaders ($n=264$, $M=11.35$, $s^2=0.03$) based on the z-test (SE=0.02, $p<.001$). There was a significant increase in CLAS scores of 0.12 between upper leaders ($n=264$, $M=11.35$, $s^2=0.03$) and senior leaders ($n=77$, $M=11.46$, $s^2=0.034$) based on the z-test (SE=0.04, $p=.001$). There was a decrease in CLAS scores of 0.04 between senior leaders ($n=77$, $M=11.46$, $s^2=0.034$) and executive leaders ($n=73$, $M=11.42$, $s^2=0.03$), but this was not significant according to the z-test (SE=0.03, $p=.19$). This confirmatory analysis suggested the piecewise model was the most tenable fit, and the model is presented in Figure 38.
Note. This regression model predicts CLAS scores from Piece 1 and Piece 2 from the Piecewise Model (N=568). Statistics are standardised regression coefficients and standard errors. The estimated covariance between Piece 1 and Piece 2 is .129, SE=.038, p=.001.

Comparison Between Skill Complexity and Role Complexity Ranges

A boxplot of CLAS scores relative to role complexity ranges for the four management layers is presented in Figure 39.
Notes. Role complexity ranges are overlaid on the boxplots. The green lines illustrate the upper boundary, midpoint, and lower boundary for each management layer.

As can be seen from Figure 39, for mid-leaders and upper leaders, the interquartile ranges occur within the role complexity range. For senior leaders, approximately half of the distribution occurs within the role complexity range. For executive leaders, most of the distribution occurs below the role complexity range. Based on the visual analysis, there is no complexity gap for mid-leaders and upper leaders. However, a complexity gap emerges for senior leaders and then widens for executive leaders.

To compare the observed versus expected distributions, a chi-square goodness-of-fit test was undertaken for each management layer. Then a confirmatory analysis was undertaken in which the mean CLAS score was compared to the role complexity range using 95% confidence intervals.
No empirical literature could be identified to inform predictions about the manner in which leaders’ actual reasoning skills may be related to role complexity ranges. Notwithstanding that expected distributions may vary depending on the organisational context, deductive reasoning was used to establish expected distributions. As a result, the findings should be interpreted with caution. In general terms, organisations expect that leaders’ skills are commensurate with the requirements of their roles. As a result, it seems reasonable to deduce the following: approximately 5% of leaders will likely fall below the lower boundary to acknowledge the minority of leaders who are appointed to roles based on other desirable leadership characteristics and are actively seeking stretch opportunities. Approximately 65% of leaders will likely fall between the lower boundary and the midpoint to acknowledge the majority of leaders whose skills match the requirements of their roles and need sufficient time for growth with a view towards longer-term promotion opportunities. Approximately 25% of leaders will likely fall between the midpoint and the upper boundary to acknowledge the minority of leaders whose skills match the requirements of their roles and need sufficient time for growth with a view toward medium-term promotion opportunities. Approximately 5% of leaders will likely fall above the upper boundary to acknowledge the minority of leaders whose skills exceed the requirements of their roles and who are ready for immediate promotion opportunities.

The observed distributions of CLAS scores and the results from the chi-square goodness-of-fit tests are presented in Table 48.
Table 48

Frequencies, Percentages, and Chi-Square Goodness-of-Fit Tests on CLAS Scores

<table>
<thead>
<tr>
<th>Management layer</th>
<th>n</th>
<th>CLAS score ≤ lower boundary</th>
<th>Lower boundary &lt; CLAS score ≤ midpoint</th>
<th>Midpoint &lt; CLAS score ≤ upper boundary</th>
<th>CLAS score &gt; upper boundary</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>Mid-leaders</td>
<td>154</td>
<td>18</td>
<td>52</td>
<td>33.8%</td>
<td>65</td>
<td>42.2%</td>
</tr>
<tr>
<td>Upper leaders</td>
<td>264</td>
<td>35</td>
<td>130</td>
<td>49.2%</td>
<td>93</td>
<td>35.2%</td>
</tr>
<tr>
<td>Senior leaders</td>
<td>77</td>
<td>32</td>
<td>34</td>
<td>44.2%</td>
<td>8</td>
<td>10.4%</td>
</tr>
<tr>
<td>Executive leaders</td>
<td>73</td>
<td>57</td>
<td>16</td>
<td>21.9%</td>
<td>0</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

* p<.001.

For mid-leaders, the chi-square test was significant \( (\chi^2, n=154)=71.71, p<.001) \), which meant there was a significant difference between the observed and expected distributions. This probably occurred because there was a lower percentage of leaders between the lower boundary and midpoint (33.8%) relative to expectations (65%), which increased the percentage of leaders in the three other categories. A 95% confidence interval around the mean score of 11.24 (11.17 to 11.31) suggested that the mean score was not statistically different from the midpoint (11.23).

For upper leaders, the chi-square test was significant \( (\chi^2, n=264)=61.06, p<.001) \), which meant there was a significant difference between the observed and expected distributions. There was a higher percentage of leaders below the lower boundary (13.3%) than expected (5%), and also a higher percentage of leaders between the midpoint and the upper boundary (35.2%) than...
expected (25%). There was a lower percentage of leaders between the lower boundary and midpoint (49.2%) than expected (65%), and also a lower percentage of leaders above the upper boundary (2.3%) than expected (5%). A 95% confidence interval around the mean score of 11.35 (11.29 to 11.41) suggested that the mean score fell between the lower boundary (11.20) and midpoint (11.43).

For senior leaders, the chi-square test was significant ($\chi^2(3, n=77)=217.7, p<.001$), which meant there was a significant difference between the observed and expected distributions. This probably occurred because there was a higher percentage of leaders who fell below the lower boundary (41.6%) than expected (5%), and this reduced the percentage of leaders in the three other categories. A 95% confidence interval around the mean score of 11.46 (11.39 to 11.53) suggested that the mean score was not statistically different from the lower boundary (11.40).

The chi-square test could not be calculated for executive leaders because leaders were only distributed between two categories. This probably occurred because there was a higher percentage of leaders who fell below the lower boundary (78.1%) than expected (5%), and this reduced the percentage of leaders in the three other categories. A 95% confidence interval around the mean score of 11.42 (11.36 to 11.48) suggested that the mean score was statistically lower than the lower boundary (11.60).

**Hypothesis Tests for Study 3**

- Hypothesis 1 was partially supported. Findings suggest that there was a statistically significant difference in the hierarchical complexity of reasoning skills between mid-leaders and upper leaders, and between upper leaders and senior leaders, but not between senior leaders and executive leaders.
• Hypothesis 2 was partially supported. Findings demonstrated that the hierarchical complexity of leaders’ reasoning skills is significantly lower than the task demands of their roles, but only for senior and executive leaders. However, a gap was not found for mid-leaders and senior leaders. Additional on the distribution of CLAS scores suggested that there was no complexity gap for mid-leaders, a shift towards a complexity gap for upper leaders, a beginning of the complexity gap for senior leaders, and a widening of the complexity gap for executive leaders.

Discussion

Findings provide mixed evidence for the hypotheses. Findings suggest a significant difference in the hierarchical complexity of reasoning skills between mid-leaders and upper leaders, and between upper and senior leaders, but not between senior and executive leaders. Findings also suggest a relatively sound fit between leaders’ reasoning skills and task demands for mid-leaders and upper leaders but a considerably weaker fit for senior and executive leaders. There is insufficient evidence to suggest the presence of a complexity gap for mid-leaders. However, there appears to be a shift towards the complexity gap for upper leaders, a beginning of the complexity gap for senior leaders, and a widening of the complexity gap for executive leaders. This pattern seemed relatively consistent across decision-making, reasoning about leadership, self-understanding, ethical reasoning, and reflective judgement. However, there is some variation as to the management layer at which the complexity gap begins for other reasoning skills (refer to Appendix N).

Each management layer exhibited a unique relation between reasoning and role complexity. Most mid-leaders displayed reasoning skills within the role complexity range (76%). However, a considerable percentage displayed reasoning skills below the lower boundary (11.7%) and above the upper boundary (12.3%). As a result, almost a quarter of mid-leaders displayed reasoning skills that are less or more complex than their roles. For upper leaders, this pattern was somewhat different.
At this management layer, an even greater majority (84.4%) displayed reasoning skills within the role complexity range. However, relatively few upper leaders were over-skilled (2.3%), while a considerable proportion (13.3%) were under-skilled. This reflects a slight increase in the percentage of under-skilled upper leaders relative to mid-leaders. For senior leaders, this pattern changed considerably. At this management layer, just over half of senior leaders (54.6%) displayed reasoning skills within the role complexity range, relatively few were over-skilled (3.9%), but a striking percentage displayed reasoning skills below the lower boundary (41.6%). This pattern became even more pronounced for executive roles because at this management layer, only 21.9% of leaders displayed reasoning skills within the role complexity range, whereas the majority (78.1%) appeared to be under-skilled relative to their role requirements.

These findings might be explained in several ways. The possibility that findings are an artefact of the sample cannot be ruled out. However, the plausibility of this explanation is mitigated by the size of the sample, the relative consistency across multiple reasoning skills, and the consistency of these findings with the general predictions discussed in the literature review. Although direct causal attributions cannot be made, these findings may be attributable to various leader and/or organisational factors.

In terms of leader-related factors, it is generally recognised that the level of hierarchical complexity at which organisms can perform is tightly coupled with—and in some contexts it is also constrained by—a variety of biological factors. These biological factors may include the number of neurons in brain (Harrigan & Commons, 2014), functional patterns of activity that occur in the brain as demonstrated through electroencephalograms and other methods (Collins et al., 2010; Fischer, 2008; Fischer et al., 1984; Leite et al., 2016), the functioning of the connectome that may impact the extent to which cognitive structures can be hierarchically integrated in the brain (Smith et al., 2019), and also as a logarithmic function of age (Commons, Miller, et al., 2014). As a result, it seems reasonable to suggest that some leaders may be subject to neurological constraints that impose upper limits on the level of hierarchical complexity at which they can reason, or the pace at which
their reasoning skills may develop. Incumbents of management roles are a function of the talent pool from which they are drawn, and the recruitment processes used in their selection. The complexity gap at more senior management layers may reflect the reasoning skills of the general leadership population or perhaps the lack of emphasis placed on complexity-related criteria in the context of recruitment (Clark, 1992; Fitzsimmons & Callan, 2016).

In terms of organisational factors, information on the amount time each leader was incumbent in their role and the number of leader development programs they had attended was unavailable. As a result, it is not possible to infer that on-the-job learning or leader development processes contributed to the complexity gap. Several meta-analyses have provided mixed results on the effectiveness of leader and management development initiatives (Avolio et al., 2010; Avolio et al., 2009; Burke & Day, 1986; Collins & Holton, 2004; Day & Sin, 2011; Lacerenza et al., 2017; Reyes et al., 2019; Rosch et al., 2017). Perhaps development initiatives leaders attended in the past had limited impact on their reasoning skills unless they were deliberately designed to do so. Alternatively, leaders may have attended development initiatives informed by developmental models that were not well suited to inform initiatives that focused on complexity. Notwithstanding the possibility of providing leaders with practices that may catalyse higher levels of mental complexity, some scholars have questioned the effectiveness and ethics of their use on some populations (Kjellström, 2009, 2010). For example, the application of such practices would need to consider whether participants have provided informed content, the role and competence of facilitators, the task demands of the practices, the suitability to leaders’ learning needs, and whether there are sufficient bodies of scientific evidence to justify their use to foster growth in complexity-related constructs.

Finally, earlier research suggested that contingencies nested in organisational culture may impose constraints on the hierarchical complexity of reasoning that leaders may attain (Commons et al., 1993). Perhaps the organisational cultures in which these leaders were operating rewarded
more simplistic forms of reasoning and extinguished more complex forms of reasoning, through processes related to operant conditioning.

On average, mid-leaders (M=11.24), upper leaders (M=11.35), senior leaders (M=11.46), and executive leaders (M=11.42) scored between 11.24 to 11.42, which falls in the early zone of Lectical Level 11, i.e., Abstract Systems Thinking. These scores suggest that leaders are reasoning in terms of relatively simple systems when making leadership decisions. These simple systems likely consist of implicit conceptions of systemic phenomena or more explicit descriptions of systemic dynamics. However, this level of reasoning cannot account for relations between systems, nested systems, or higher-order principles that govern the relations between multiple systems. Despite this, many leadership roles require a more sophisticated understanding of relations between multiple systems, including financial, stakeholders, human capital, ethical, environmental, political, community, etc. However, these advanced reasoning skills typically unfold in late Lectical Level 11 and early Lectical Level 12. The impetus for developing the cognitive skills required to navigate a complex stakeholder environment seems conceptually and empirically clear. For example, a review of over 200 peer-reviewed articles demonstrated that stakeholder-focused approaches to leadership outperformed other leadership approaches across a range of shareholder and stakeholder criteria (Lemoine et al., 2021).

Given that the leaders in this sample had already been appointed to their roles, solutions to the complexity gap may focus on redeployment and/or development. Leaders experiencing a large complexity gap may need to be redeployed to more junior roles with less complex task demands. Some leaders may benefit from leader development initiatives which strengthen their ability to reason in more complex ways. Finally, leadership development process may be required to implement forms of social scaffolding that may reduce the impact of the complexity gap in individual leaders such as collaborative decision-making. Collaborative decision-making is an instance of collective leadership that typically involves practices that allow groups to arrive at decisions for which joint accountability is taken (Buck & Villines, 2007; Daniel, 2015; Hiller et al., 2006 ).
Collaborative practices may involve seeking diverse arguments, surfacing assumptions, building aggregated understanding, and fostering collective ownership.

This study has several limitations. The sample size was relatively limited at the senior (n=77) and executive (n=73) layers. As Lectica collects more data from senior leaders, it will be possible to validate whether similar patterns are obtained in larger samples. Analyses were limited to general complexity ranges. As Lectica refines its methodology, it may be possible to undertake similar analyses on precise role complexity estimates for specific organisations and/or job families. Broader implications, limitations, and future research directions are discussed in Chapter 6.

Given the evidence reported in Study 3, it seems reasonable to suggest that leaders in the 21st century are experiencing a complexity gap. Study 4 explores the possibility of whether, and to what extent, leader development processes may play a role in partially closing this gap.
Chapter 5: Longitudinal Growth in the Hierarchical Complexity of Leaders’ Reasoning Skills

Given the plausible existence of a complexity gap—and the potential implications for leader and organisational performance—it is reasonable to enquire into the development processes which may contribute to closing this gap. “There are only two logical ways to mend this [gap]—reduce the world’s complexity or increase our own. The first isn’t going to happen. The second has long seemed an impossibility in adulthood” (Kegan & Lahey, 2009, p. 12). As a result, the purpose of Study 4 was to address Research Question 3: To what extent do leaders develop the hierarchical complexity of their reasoning skills during participation in various leader development programs? This question was addressed by conducting a longitudinal analysis of a large sample of leaders who participated in various developmental initiatives. The goal was to better understand the underlying patterns of change in leader development.

Approach to Researching the Hierarchical Complexity of Leaders’ Reasoning Skills

An approach to closing and/or mitigating the complexity gap must consider a) the development of leaders, b) as a function of leader development processes, c) via longitudinal methods, and d) as it relates to change in hierarchical complexity, amongst other dependent variables. The development of leaders must consider the general developmental trajectory of reasoning skills over time. Leader development processes must consider the conditions in which
individual attributes of leaders may grow. Longitudinal methods must include at least three test intervals, administered with a relevant frequency, to track change over time (Day, 2011). Longitudinal methods are critical given that cognitive development has been shown to unfold slowly, if at all, in adulthood (Armon & Dawson, 1997). Finally, tracking change in hierarchical complexity is required to detect forms of development that may contribute to closing and/or mitigating the complexity gap. This approach is consistent with the argument put forward by Day et al. (2009) that individual leader development needs to be considered in the context of ongoing adult development processes over long timescales.

The challenge of pursuing such an approach is underscored by the facts that longitudinal research is notoriously challenging to undertake, and little has been done to integrate the fields of leadership, adult development, and cognitive development. The leader development literature has mostly applied quasi-longitudinal methods to detecting change in psychological and behavioural variables which are unrelated to hierarchical complexity e.g., Riggio and Mumford (2011). Quasi-longitudinal methods are limited to detecting change as a linear relation between two time points (Day, 2011). The general cognitive literature has demonstrated that cognitive skills are related to favourable leadership outcomes e.g., M. D. Mumford et al. (2015). However, relatively few studies have used longitudinal methods to examine the processes which can strengthen these skills. The longitudinal cognitive-developmental literature has tracked change in hierarchical complexity over long timescales e.g., Armon and Dawson (2002). However, most of this literature is conducted on the general adult population rather than leadership populations, and it evaluates naturalistic trajectories of growth rather than how it might occur given exposure to specific developmental catalysts. A second body of cognitive-developmental literature examines change in hierarchical complexity associated with various development processes e.g., Upadhyaya et al. (2015). Although this body of literature is most relevant to Study 4, it typically employs quasi-longitudinal methods on the general adult population and in relatively small samples. These bodies of literature provide
limited insight on the extent to which we might expect to close the complexity gap in organisational leaders. Each body of literature is reviewed in turn.

The Leader Development Literature

The broader field of leadership has a history of more than a century (Antonakis & Day, 2018; Avolio et al., 2009), but scientific enquiry into the nature of leadership development is still relatively nascent (Day et al., 2014). In this context, leader development must be distinguished from leadership development (Day, 2000). Leader development is analogous to the development of human capital and involves the growth of mindsets, behaviours, skills, and other intrapersonal attributes of individual leaders. Leadership development is analogous to the development of social capital and involves the growth of relationships, systems, and broader social networks in which the focus is on interpersonal connections among individuals. Leader development may be further subdivided into leader training and leader development (Day et al., 2021). Leader training is typically focused on more concrete skills over shorter timescales and is appropriate for broader audiences. Leader development is more holistic, occurs over longer timescales, and is more appropriate for leaders who occupy more senior roles. Given the nature of hierarchical complexity as a dependent variable in the context of Study 4, leader development processes are of greatest relevance.

Perhaps a more integrative conception of leader and leadership development was introduced by Kjellström et al. (2020). A phenomenographic analysis based on semi-structured interviews with workplace professionals suggested that there may be at least six increasingly complex ways of making sense of development processes. The first way focused on one’s own development with little generalisation to leadership. The second way focused on filling a leader role and emphasised coordination of one’s own development with one’s role requirements. The third way focused on personal development and was considered to transcend but include development within the role. The fourth way focused on integrating leader and organisational development and emphasised alignment between leader, personal, and organisational development. The fifth way
focused on collective leadership development and emphasised shared forms of leadership. The sixth way focused on human development and conceived of development as an evolutionary process that transcends organisational boundaries. In the context of this heuristic, the second way of making sense of leader development seems most relevant to Study 4.

As noted, the leader development literature has typically examined change in various psychological and behavioural variables, but there has not been a strong emphasis on complexity-related phenomena. For example, a special issue of The Leadership Quarterly in 2011 (Riggio & Mumford, 2011) was dedicated to longitudinal research, but it focused on variables such as personality traits and intelligence (Reichard et al., 2011), and leader identity (Day & Sin, 2011). Another special issue of The Leadership Quarterly in 2021 presented several longitudinal studies, but they focused on topics such as critical experiences at successive life phases (Liu et al., 2021) and the presentation of self-as-leader through vocal cues (Truninger et al., 2021). These studies have contributed to our general knowledge about longitudinal growth, but they do not directly address approaches which may close the complexity gap.

The General Cognitive Literature

A considerable proportion of the leadership literature has been informed by behavioural approaches (Yukl, 1971), even though cognition has been shown to predict a variety of leadership outcomes such as leader emergence and performance (M. D. Mumford et al., 2015). For example, cognitive skills, such as generating alternative solutions and the application of wisdom, are moderately predictive of leader performance in some contexts (M. Mumford et al., 2015; Mumford et al., 2012; Zaccaro et al., 2013). Despite emerging insights into the predictive validity of cognitive skills, relatively few studies have attempted to examine the growth of cognitive skills in leadership populations over time. The limited number of studies on this topic have addressed issues such as beliefs during an organisation crisis (Carrington et al., 2019) rather than the cognitive implications of navigating organisational complexity. Limited research has shown that developmental processes can
strengthen cognitive skills as evaluated by expert raters (Santos et al., 2015), but there is much to learn about the underpinning mechanisms that foster these types of growth.

**Cognitive-Developmental Literature (with Longitudinal Designs)**

The cognitive-developmental literature has typically focused on naturalistic growth trajectories in the general adolescent and adult populations, rather than in leader populations. Leader populations consist of individuals at any layer in an organisational hierarchy who practice the process of leadership. Three of the most relevant studies are presented below.

Kegan (1994) conducted a longitudinal study on 22 adults. Participants’ stage of development was assessed via the subject-object interview annually for four years and then intermittently after that. Preliminary findings from the first four years suggested that 15 adults grew by a maximum of two-fifths of a stage per year, two adults demonstrated no growth, and four adults regressed between at least two successive years (Kegan, 1994). Kegan concluded that there was a general trend for increasingly complex stages of development to unfold in adults.

Armon and Dawson conducted two longitudinal studies on approximately 30 adults (Armon & Dawson, 2002). Adults participated in two cognitive-developmental interviews focused on their conceptions of the good life and moral reasoning. They participated in four interviews which were spaced approximately four years apart. For reasoning about the good life, results demonstrated that one participant developed by a full stage, 22 participants developed by less than a full stage, and eight participants demonstrated no change (Armon & Dawson, 1997, 2002). For moral reasoning, results demonstrated a significant increase by approximately 0.7 of a stage across the whole group. However, there was a curvilinear relation between the rate of growth and age. For younger adults (23-50 years) there was significant growth at a rate of one stage per year. For older adults (50-72) there was no significant increase with age (Armon & Dawson, 1997).

Finally, Kitchener and King conducted a longitudinal study on development through stages of reflective judgement over 10+ years (King & Kitchener, 1994; Kitchener & King, 1981). Reflective
judgement stages account for increasingly complex conceptions of truth, evidence, reasoning, and knowledge. High school, college, and advanced doctoral students participated in cognitive-developmental interviews that were spaced two, four, and ten years after an initial test. Overall, 92% of participants demonstrated significant growth in their stage of reflective judgement. However, the trajectories of development were different between cohorts. High school students developed the most and demonstrated increasingly complex reasoning at successive test administrations. College students experienced a more modest increase. Advanced doctoral students demonstrated no/minimal change over time. This general pattern is consistent with six other longitudinal studies reported by Kitchener et al. (2006).

These studies suggest the possibility of growth in the general adult population. Nonetheless, there is mixed evidence about the rate of growth because findings suggest that growth may occur slowly or arrest completely in adulthood. Whether these patterns generalise to leaders who participate in leader development processes remains to be determined.

Cognitive-Developmental Literature (with Interventions)

The final body of literature examines interventions associated with increased in hierarchical complexity over time. This body of literature is particularly pertinent because it provides insight into the characteristics of leader development processes likely to contribute to closing the complexity gap and is therefore reviewed in greater detail.

Kohlberg’s moral education is a suite of developmental interventions explored in the early cognitive-developmental literature. Moral education includes the study of moral exemplars (e.g., Martin Luther King), the establishment of democratic decision-making systems called ‘just communities’, and the discussion of moral dilemmas. In this context, the discussion of moral dilemmas has probably attracted the most significant body of evidence. This approach aims to facilitate a Socratic discussion about moral dilemmas to create learning conditions in which participants can reason at higher stages of moral judgement (Kohlberg, 1975). Kohlberg (1975)
argued that discussion needed to satisfy the following three conditions for growth to occur. First, it must provide exposure to higher stages of reasoning. Second, it must expose contradictions and limitations of participants’ current stage of reasoning. Third, it must foster an atmosphere in which conflicting views can be openly exchanged. This was typically implemented by a facilitator who introduced an ambiguous moral dilemma to a group, identified which stages of reasoning were being applied by participants, elicited participants’ insight into the limitations of their current reasoning, and then introduced considerations at successively higher stages of moral judgement. This approach has been shown to facilitate the complexity of reasoning in various adult populations (Arbuthnot, 1984; MacPhail, 1989). Kohlberg’s research on moral reasoning has also been used to develop theoretical conceptions and measures related to the ethics of care (Juujärvi et al., 2020; Juujärvi et al., 2010; Skoe, 1995; Skoe & Diessner, 1994; Skoe & Marcia, 1991). As to whether Kohlberg’s suite of interventions may be used to consistently foster development in the ethics of care remains an open empirical question.

Deliberate Psychological Education is an approach to fostering advanced forms of reasoning (Mosher & Sprinthall, 1971). The approach is intended to create optimal learning conditions by providing experiences that increase perspective-taking, reflection, and psychological awareness. This typically involves taking diverse perspectives, guided journaling, and practising interpersonal skills such as listening and showing empathy. Deliberate psychological education is associated with—and possibly causally related to—significant increases in hierarchical complexity in diverse samples of adults such as young women (Erickson, 1977), teachers (Reiman & Parramore, 1993) and counsellors (Brendel et al., 2002).

Wolfsont (2002) conducted an experimental study to determine whether a single coaching session could foster more advanced forms of reasoning. A small sample of participants ($N = 4$) completed one coaching session based on the principle of whole brain understanding in which reflective, emotional, and physical exercises were employed to help participants think about their personal goals in new ways. Participants were interviewed pre- and post-intervention on their
conceptions of a personal goal and how to attain it. The interviews were scored with the Hierarchical Complexity Scoring System (HCSS) (Commons et al., 2005), which is based on The Model of Hierarchical Complexity (Commons & Pekker, 2008). Although the sample was too small to conduct statistical analyses, there was a general trend for the hierarchical complexity of reasoning to increase by at least one stage after the coaching session. The researchers suggested that whole brain approaches may help adults access their innate cognitive resources and reason at higher stages over short intervals.

Ross (2006a, 2006b) conducted an experimental study to determine whether a facilitated discourse process would increase the complexity of participants’ reasoning about community-related issues. A small sample of community members (N = 8) took part in a six-week discourse process called The Integral Process for Working on Complex Issues. Through the process, community members engaged in weekly sessions during which they identified complicated community issues, selected an issue to focus on, identified systemic actions, identified issues and risks, and deliberated on the options generated. An interview was administered pre- and post-participation, and responses were scored with the Hierarchical Complexity Scoring System (HCSS) (Commons et al., 2005), which is based on The Model of Hierarchical Complexity (Commons & Pekker, 2008). Quantitative analyses demonstrated a significant increase in reasoning which would not typically be expected after six weeks. The increase in hierarchical complexity had a large effect size of .79, as calculated by Cohen’s d. Additional qualitative analyses also demonstrated changes in interactions between community members which become more positive, coherent, and deliberative.

Lewis et al. (2005) conducted a longitudinal study on military cadets as they progressed through Kegan’s constructive-developmental stages. Cadets (N=38) attending college took part in three subject-object interviews in their freshmen, sophomore, and senior years respectively. Findings indicated that at test-time 1 which occurred in their freshmen year, 21% of cadets were at Stage 2, 63.2% were transitioning between Stages 2 and 3, 25.8% were at Stage 3, and 0% were at Stage 4. By test-time 3, which occurred in their senior year, significant growth occurred because
6.3% of cadets were at Stage 2, 31.3% were transitioning between Stages 2 and 3, 43.7% were at Stage 3, and 18.7% were transitioning between Stages 3 and 4. For senior year cadets, there was also a statistically significant relation between their performance on the subject-object interview and a composite score which reflected their performance in fulfilling military training requirements.

Some issues with the research design are noted because of attrition and also adding in new cadets to the study at test-time 2. Nevertheless, the researchers concluded that finding support the occurrence of growth over time.

Bartone et al. (2007) conducted a longitudinal study on college students as they progressed through Kegan’s constructive-developmental stages. College students (N = 52) attending the U.S. Military Academy at West Point took part in three subject-object interviews in years 1, 2 and 4 of college. Analyses demonstrated a significant increase for 47% of participants. Growth in stage was moderately related to perceived ratings of leadership from peers (ρ=.38) and subordinates (ρ =.30) but not to ratings of leadership from supervisors (ρ =.16). There was nonsignificant growth between years 1 and 2 but significant growth between years 1 and 4, and between years 2 and 4. The researchers concluded that while only correlational inferences could be made, growth occurred at a faster rate than what would be expected in some other settings.

Building upon earlier studies on interventions that caused stage development in children (Commons & Davidson, 2015) and college students (Commons, Miller, et al., 1982), other researchers demonstrated that environmental contingencies led to the development of hierarchical complexity in adults (Adhikari, 2016; Upadhyaya et al., 2015). The first study aimed to determine whether reinforcement of correct responses would increase stage performance in Nepalese adults (Upadhyaya et al., 2015). Participants (N = 39) undertook an assessment based on The Model of Hierarchical Complexity (Commons & Pekker, 2008), which consisted of questions at successive developmental stages. The assessment required participants to answer a series of questions at Stage S correctly before moving on to questions at Stage S+1. If participants did not answer the series correctly, they were repeatedly presented with the same questions. If participants still could
not respond correctly, they were asked questions from the previous stage, and correct responses were reinforced with a monetary reward. Results demonstrated that there was a significant increase by one stage post-reinforcement. The researchers concluded that reinforcement of higher stage responses may be sufficient to increase stage of adult performance.

A second quasi-experimental study closely mirrored the study described above (Adhikari, 2016). A sample of adults \( (N = 33) \) completed two assessments based on The Model of Hierarchical Complexity (Commons & Pekker, 2008) before and after a training intervention. Procedures like those described above were used, except the assessment was administered in groups. As a result, not only were participants rewarded with money and points for correct answers, but also individual points were summed to form a group total which provided an additional incentive for higher team performance. Overall, participants’ stage of performance increased by more than one stage after the intervention, and the researchers concluded that behavioural reinforcement was the most plausible explanation for this change over such a short time interval.

Most recently, Cavallaro and French (2021) explored adults’ growth through Kegan’s constructive-developmental stages during a graduate program. Students self-selected into two groups. The first was a comparison group \( (N=9) \) which completed core curriculum courses in military operations, decision-making, strategy, and policy. The second was an experimental group \( (N=7) \) which completed the core curriculum in addition to leadership subjects in ethics and self-awareness, and individual coaching sessions. Students undertook a subject-object interview at the beginning of the program and a post-test nine months later. Across the combined sample, 69% of students demonstrated some growth. However, 88% of students who began the course at the third stage demonstrated growth in contrast to only 46% who began the course at the fourth stage and demonstrated growth. There was no significant difference in growth between the experimental and the comparison group, suggesting that the additional leadership modules and coaching did not accelerate growth beyond that which occurred during the core curriculum.
Finally, coaching methodologies based on assumption-testing, and leader development processes based on the need for personal vulnerability, have been proposed to foster developmental complexity in leaders (Kegan & Lahey, 2009, 2016). At this point in time, it seems that only anecdotal evidence and case studies rather than empirical evidence, have been used to support these claims e.g., Helsing et al. (2008). This limits the causal claims that can be made about the effectiveness of these interventions.

Although these studies contribute to our knowledge about processes associated with growth in hierarchical complexity, they also reflect several methodological limitations. Most studies did not include control conditions, many used small samples, several applied interventions to non-leader populations, and all of them used quasi-experimental designs at best. Thus, the level of causal understanding from these studies is highly limited. Study 4 reflects an attempt to resolve a number of these issues.

The Process of Cognitive Development

Most cognitive-developmentalists agree that growth does not occur at the same rate between different knowledge domains (Fischer & Bidell, 1998). However, there has been considerable debate about whether a single underlying process accounts for growth in different domains. Early cognitive developmentalists argued that different domains were characterised by distinct stages and underlying processes (Kohlberg et al., 1983). Contemporary cognitive developmentalists argue that different domains consist of different conceptual content but reflect the same hierarchically integrated structures and growth through these structures is explained by a single underlying process (Dawson et al., 2005).

In the Piagetian (1985) tradition, growth occurs when new information cannot be assimilated into existing cognitive structures. Growth occurs when new information reveals limitations of existing cognitive structures and causes an uncomfortable state of cognitive disequilibrium. In this state, cognitive structures must be restructured to accommodate new
information and regain cognitive equilibrium. Piaget (2000) referred to the process of transforming lower-order cognitive structures into qualitatively distinct higher-order cognitive structures as *reflective (or reflecting) abstraction*. Reflective abstraction leads to a state in which cognitive structures become *hierarchically integrated* in the mind-brain (Smith et al., 2019). This hierarchical ‘chunking’ reduces the number of concepts which must be coordinated simultaneously and frees up cognitive processing space to enable increasingly complex reasoning to emerge (Burtis, 1982; Halford, 1999). Hierarchical complexity is conceived to be the behavioural manifestation of hierarchical integration and is directly observable in sensorimotor actions, verbal behaviour, and social contexts (Dawson et al., 2005; Dawson-Tunik, 2006).

**Rationale for Study 4**

The literature review suggests that growth in hierarchical complexity is possible, and this may occur as a function of leaders’ engagement with certain types of developmental processes. Lectica applies the process described above to create conditions in which leaders can engage with learning material located one Lectical Phase (i.e., one-quarter of a complexity level) above their current ability. This relates somewhat to the scaffolding processes proposed by various developmentalists (Fischer & Bidell, 2006; Vygotsky, 1978). Given this material is of a higher order of hierarchical complexity, it is unlikely to be assimilated into leaders’ existing cognitive structures. As leaders engage with this material in low-stakes conditions, it may lead to disequilibrating experiences that stimulate the process of reflective abstraction. Repeated engagement with this material may result in more hierarchically integrated cognitive structures. In turn, these structures are manifested in higher amounts of hierarchical complexity exhibited in samples of reasoning over time (Dawson & Stein, 2011).

Increased engagement with more complex material may result in a higher rate of growth in hierarchical complexity, and this may be a function of a) the number of cognitive-developmental assessments completed by leaders, b) the average instruction time received per month, c) and
formative versus summative application of cognitive-developmental assessments. Formative versus summative applications of assessments are more likely to result in growth because they incorporate learning material aimed at a level of hierarchical complexity that is more likely to stimulate the process of reflective abstraction. Finally, different leader development programs may also be associated with varying rates of growth in hierarchical complexity as a function of their pedagogical characteristics. The rate at which the hierarchical complexity develops in individuals may also help to determine whether leader development processes suffice to close the complexity gap identified in Chapter 4. These considerations lead to the following hypotheses to be tested in Study 4.

**Hypotheses for Study 4**

- **Hypothesis 1**: There is an increase in the hierarchical complexity of leaders’ reasoning skills during participation in various leader development programs, as determined by a statistically significant slope, s.
- **Hypothesis 2**: The rate of growth in hierarchical complexity as determined by the slope, s, will be significantly moderated by a) the number of cognitive-developmental assessments completed over the course of a development program, b) the average instruction time received per month, c) and formative versus summative application of cognitive-developmental assessments.
- **Hypothesis 3**: There is a statistically significant difference in rates of growth between different development providers as determined by differences in slopes, s.

**Method**

**Participants**

Lectica had previously collected the data between the years 2006 and 2016. The original sample was extracted from Lectica’s database of over 40,000 cognitive-developmental assessments and consisted of 5060 Lectical Assessments completed by 1787 leaders. It included all persons who
occupied a leadership role, had participated in various leader development initiatives offered by 11 leadership providers, and had completed at least two Lectical Assessments. Leaders consented to their data being shared with third parties for legitimate research purposes via Lectica’s registration process which includes privacy and legal statements (refer to Appendix C). Filters were applied to the original sample to yield an analysable sample and thus, the sampling process was not entirely random. The analysable sample consisted of 1152 leaders who had completed 2380 Lectical Assessments at test-times 1, 2, 3, and 4. When data were collected, test-takers were provided with a binary option to select for sex. Demographic details for these leaders are provided in Table 49.
Table 49

*Demographic Details of Leaders for Study 4*

<table>
<thead>
<tr>
<th>Demographics</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Leadership Provider (geography, sector)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (North America, private)</td>
<td>43</td>
<td>3.7</td>
</tr>
<tr>
<td>3 (Australia, private)</td>
<td>24</td>
<td>2.1</td>
</tr>
<tr>
<td>4 (Australia, public)</td>
<td>8</td>
<td>0.7</td>
</tr>
<tr>
<td>5 (North America, public)</td>
<td>576</td>
<td>50.0</td>
</tr>
<tr>
<td>6 (North America, public)</td>
<td>26</td>
<td>2.3</td>
</tr>
<tr>
<td>7 (South America, private)</td>
<td>77</td>
<td>6.7</td>
</tr>
<tr>
<td>8 (Multinational, private)</td>
<td>59</td>
<td>5.1</td>
</tr>
<tr>
<td>9 (North America private)</td>
<td>12</td>
<td>1.0</td>
</tr>
<tr>
<td>10 (North America, private)</td>
<td>42</td>
<td>3.6</td>
</tr>
<tr>
<td>11 (North America, public)</td>
<td>89</td>
<td>7.7</td>
</tr>
<tr>
<td>12 (North America, public)</td>
<td>196</td>
<td>17.0</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>591</td>
<td>51.3</td>
</tr>
<tr>
<td>Female</td>
<td>379</td>
<td>32.9</td>
</tr>
<tr>
<td>Unidentified</td>
<td>182</td>
<td>15.8</td>
</tr>
<tr>
<td><strong>First language</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>808</td>
<td>70.1</td>
</tr>
<tr>
<td>Not English</td>
<td>114</td>
<td>9.9</td>
</tr>
<tr>
<td>Unidentified</td>
<td>230</td>
<td>20.0</td>
</tr>
</tbody>
</table>

*Materials*

The pedagogical design of the initiatives varied widely and included custom leader development and/or coaching programs, public leader development programs, and postgraduate programs offered by tertiary institutions. Leadership providers were responsible for the design, delivery, and measurement approaches applied to their respective initiatives.
As discussed in Chapter 2, Lectica offers a suite of cognitive-developmental assessments called Lectical Assessments that assess the hierarchical complexity of reasoning in different domains. A variety of Lectical Assessments were administered to leaders, including the LDMA, the Lectical Leadership Reasoning Assessment (LLRA), the Lectical Self-Understanding Assessments (LSUA), and the Lectical Developmental Pedagogy Assessment (LDPA). A detailed description of Lectical Assessments, including a sample dilemma and a list of items, is provided in Chapter 2, Study 1A. The number of assessments administered at test-times 1 through 4 aggregated across leadership providers is presented in Table 50.

Table 50

<table>
<thead>
<tr>
<th>Lectical Assessment</th>
<th>Test-time 1</th>
<th>Test-time 2</th>
<th>Test-time 3</th>
<th>Test-time 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDMA</td>
<td>963</td>
<td>687</td>
<td>281</td>
<td>242</td>
<td>2173</td>
</tr>
<tr>
<td>LLRA</td>
<td>27</td>
<td>8</td>
<td>0</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>LSUA</td>
<td>42</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>LDPA</td>
<td>52</td>
<td>43</td>
<td>2</td>
<td>0</td>
<td>97</td>
</tr>
<tr>
<td>Total</td>
<td>1084</td>
<td>766</td>
<td>283</td>
<td>247</td>
<td>2380</td>
</tr>
</tbody>
</table>

As discussed in Chapter 2, test-takers’ responses can be scored via Lectica’s two scoring systems—the LAS and the CLAS. The purpose of both systems is to locate the magnitude of hierarchical complexity exhibited in samples of reasoning along the Lectical Scale. CLAS was used to score leaders’ responses to Lectical Assessments for Study 4. As discussed in Chapter 2, LAS and CLAS have a high inter-rater agreement and can therefore be treated as interchangeable for research purposes within certain constraints. Lectical decimal scores have been used in some empirical research (Fuhs, 2016; Van Rossum, 2013). Statistical models were constructed in Mplus.
(Muthén & Muthén, 1998-2015), and the syntax for all analyses presented in Chapter 5 is documented in Appendix O. Microsoft Excel was used to produce graphs.

**Procedure**

**Administration.** Lectical Assessments had been administered through various accredited providers whose identities cannot be disclosed due to contractual agreements with Lectica. There were considerable differences in the amount of time that passed between successive assessment administrations for different leadership providers.

**Selection of Cases.** In addition to awarding a CLAS score, CLAS also awards other scores that can be used as filters. Filters were applied to identify assessments from the original sample that had been scored with sufficient confidence to include in the analysable sample. Three filters were used in the context of Study 4. The first filter was the difference between each assessment's CLAS score and LAS score. Assessments with a difference of less than or equal to $\pm 0.3$ were included to ensure sufficient similarity between scores and consistency with the general procedures applied to previous research (Fuhs, 2016; Van Rossum, 2013). The second filter was CLAS's confidence in the CLAS score based on the density of concepts at successive Lectical Phases. The confidence score ranged from 1 to 10, and assessments with a confidence score equal to or greater than 6 were included. This threshold was considered suitable for research purposes and is consistent with the general procedures applied to previous research (Fuhs, 2016; Van Rossum, 2013). The third filter was the unique word count. A minimum threshold was set for each year of educational attainment to ensure that leaders had elaborated on their responses sufficiently for CLAS to award a score confidently. Assessments with a unique word count equal to or greater than the minimum threshold were included (refer to Table 51).
Table 51

Thresholds Applied to Unique Word Count by Years of Education and Approximate Grade

<table>
<thead>
<tr>
<th>Years of education</th>
<th>Low threshold</th>
<th>High threshold</th>
<th>Approximate grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>110</td>
<td>800</td>
<td>End of high school</td>
</tr>
<tr>
<td>13</td>
<td>120</td>
<td>1000</td>
<td>Start of university</td>
</tr>
<tr>
<td>14</td>
<td>150</td>
<td>1000</td>
<td>Start of university</td>
</tr>
<tr>
<td>15</td>
<td>150</td>
<td>1000</td>
<td>Bachelor’s degree</td>
</tr>
<tr>
<td>16</td>
<td>175</td>
<td>1000</td>
<td>Start of Masters</td>
</tr>
<tr>
<td>17</td>
<td>175</td>
<td>1000</td>
<td>Start of PhD</td>
</tr>
<tr>
<td>18</td>
<td>175</td>
<td>1000</td>
<td>Start of PhD</td>
</tr>
<tr>
<td>&gt; 19</td>
<td>200</td>
<td>1000</td>
<td>Start of PhD</td>
</tr>
</tbody>
</table>

Modelling Approach and Missing Data. CLAS awards a single score to each assessment rather than a unique score to each item. For this reason, regression models on manifest variables were used, rather than a combination of measurement and structural models. This approach imposes some limitations on the ability to account for error variance. However, this is justified based on the measurement properties displayed by the LDMA in Study 1A. Although this analytical approach does not have all the benefits of latent growth modelling (LGM) and multiple-indicator latent growth modelling (MLGM), it is consistent with the variety of approaches which can evaluate change over time (Day & Lance, 2004).

Missing data could be attributed to planned missingness or participant attrition. Planned missingness occurred when providers only administered assessments at test-times 1 and 2, but not at test-times 3 and 4. Participant attrition occurred when participants did not complete an assessment that was assigned to them at any test-time. Imputation was considered to address missing data but was not employed because it has not been found to provide significant advantages.
over Full Information Maximum Likelihood (FIML) which is the default computational approach in Mplus. FIML also yields similar chi-square and fit statistics to other imputation approaches, particularly with large samples with sufficient statistical power (Acock, 2012; Arbuckle, 1996).

Results

Hierarchical Clustering in the Data Structure

The data were nested at multiple layers, so the amount of variability attributable to hierarchical clustering needed to be determined. Leadership providers were considered to be the primary sampling unit (PSU), however hierarchical clustering was examined for the three clusters that were most likely to have impacted leaders’ scores i.e., leadership providers, leader development program, and specific training cohort. Hierarchical clustering was analysed via two methods—intraclass correlations (ICCs) and the design effect (Deff). ICCs partition variance at the within and between levels and then determine the proportion of within variance relative to the total variance (i.e., within and between). Deff indicates the magnitude of bias in the standard errors (SEs) when the data are not independent such that Deff=1+(c-1)ρ, where c is the average cluster size, and ρ is the ICC. ICCs greater than .05 and Deffs greater than 2 are typically considered to require multi-level modelling. Results are presented in Table 52.

Table 52

<table>
<thead>
<tr>
<th>Clustering variable</th>
<th>ICCs for CLAS 1 (ρ₁)</th>
<th>ICCs for CLAS 2 (ρ₂)</th>
<th>ICCs for CLAS 3 (ρ₃)</th>
<th>ICCs for CLAS 4 (ρ₄)</th>
<th>Average cluster size</th>
<th>Deff a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider</td>
<td>.36</td>
<td>.42</td>
<td>.40</td>
<td>.50</td>
<td>101.0</td>
<td>37.2</td>
</tr>
<tr>
<td>Program</td>
<td>.30</td>
<td>.37</td>
<td>.35</td>
<td>.39</td>
<td>67.3</td>
<td>19.9</td>
</tr>
<tr>
<td>Cohort</td>
<td>.28</td>
<td>.32</td>
<td>.25</td>
<td>.24</td>
<td>31.077</td>
<td>8.1</td>
</tr>
</tbody>
</table>

*Note.* a Deff was calculated on the lowest ICC for each clustering variable.
The ICCs ranged from .24 to .49 and therefore exceeded the critical value of .05. All Deffs exceeded the critical value of 2. As a result, aggregated multi-level modelling was used to obtain unbiased estimates and standard errors. This approach was favoured over disaggregated modelling because analyses were aggregated at the level of CLAS scores, rather than disaggregated by comparing sources of variation at various levels of the data hierarchy. Aggregated multi-level modelling also offered greater flexibility when constructing the model syntax (Preacher et al., 2010). Provider was chosen as the clustering variable as it accounted for the greatest amount of variability in CLAS scores. Furthermore, it was the most meaningful way to group data for comparative analysis.

Descriptive Statistics

Descriptive statistics for CLAS scores at the four test-times is presented in Table 53.

Table 53

Descriptive Statistics for CLAS Scores at the Four Test-Times Included in Study 4

<table>
<thead>
<tr>
<th>CLAS Scores at Successive Test-Times</th>
<th>n</th>
<th>M</th>
<th>( s^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLAS 1</td>
<td>1084</td>
<td>11.30</td>
<td>0.04</td>
</tr>
<tr>
<td>CLAS 2</td>
<td>766</td>
<td>11.33</td>
<td>0.04</td>
</tr>
<tr>
<td>CLAS 3</td>
<td>283</td>
<td>11.36</td>
<td>0.04</td>
</tr>
<tr>
<td>CLAS 4</td>
<td>247</td>
<td>11.41</td>
<td>0.03</td>
</tr>
</tbody>
</table>

A graph of the whole sample was uninterpretable because of the sheer number of cases, so a random sample of 250 persons was extracted from the analysable sample and plotted by test-time (Figure 40) and by month of administration post-test-time 1 (Figure 41).
Figure 40

CLAS Score by Test-time 1 Through 4 for a Random Sample of 250 Persons

Figure 41

CLAS Score by Month of Administration Post Test-Time 1 for a Random Sample of 250 Persons
**Model Fit**

The trajectory of development was analysed by comparing the relative fit of several regression models which successively deviated more significantly from linearity. These were an intercept only model, a linear model, a quadratic model, a piecewise model, and a time-variant model. A Maximum Likelihood Robust (MLR) estimator was employed in all cases because of its robustness to non-normality and non-independence. For the intercept only model, the intercept reflected scores at test-time 1, no growth over time was reflected through a slope fixed at 0 between successive test-times. The linear model fixed an equal amount of time between successive test-times and tested the fit of a linear growth trajectory. The quadratic model fixed an equal amount of time between successive test-times and tested the fit of a quadratic growth trajectory. The piecewise model fixed an equal amount of time between test-times and tested the fit of growth occurring via two separate linear relations. The first linear relation occurred between test-times 1 and 2, and the second linear relation occurred between test-times 2 and 4. Finally, the time-variant model freed the time between test-times and tested the fit of a growth trajectory based on the actual number of months between successive administrations.

For reasons discussed in Chapter 4, the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and the sample-size adjusted BIC were used to estimate the incremental improvement in fit between successive models (Muthén & Muthén, 1998-2015; Vrieze, 2012). Some models were saturated, so the chi-square test of model fit could not be used (Muthén & Muthén, 1998-2015). Besides these statistical criteria, the design of the programs was also considered when evaluating model fit. Fit statistics are presented in Table 54.

For these analyses, the intercept is denoted by $i$, the slope by $s$, standard errors by SE, variance by $s^2$, probability values by $p$, and CLASx denotes CLAS scores at test-time x. In the context of the current data set, the intercept, $i$, reflects test-takers’ scores at test-time 1 and the slope, $s$, reflects the average growth rate over time.
Table 54

*Fit Indices for the Four Models Applied to Study 4*

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>AIC</th>
<th>BIC</th>
<th>Sample adjusted BIC</th>
<th>χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
<td>df</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>Intercept only</td>
<td>1152</td>
<td>-1903.67</td>
<td>-1873.37</td>
<td>-1892.43</td>
<td>360.51</td>
</tr>
<tr>
<td>Linear</td>
<td>1152</td>
<td>-2065.24</td>
<td>-2019.80</td>
<td>-2048.39</td>
<td>9.53</td>
</tr>
<tr>
<td>Quadratic</td>
<td>1152</td>
<td>-2060.55</td>
<td>-1994.91</td>
<td>-2036.202</td>
<td>*</td>
</tr>
<tr>
<td>Piecewise</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Time-variant</td>
<td>1084</td>
<td>-2050.00</td>
<td>-2005.10</td>
<td>-2033.67</td>
<td>*</td>
</tr>
</tbody>
</table>

Note. An asterisk, *, indicates that the fit statistic could not be computed. The chi-square test of model fit could not be computed for all models because some models were saturated. However, given that this fit statistic is sensitive to deviations from normality and sample size, it was not considered the most appropriate statistic to make inferences about model fit. The piecewise model could not be identified even after 1000 starts, so it was not considered further. As a result, interpretation of relative fit between models was primarily based on the AIC and BIC statistics and design considerations.

For the intercept only model (n=1152), test-takers received a mean CLAS score of 11.33 at test-time 1 (SE=0.02, p<.001), and there was no growth over time. There was a significant amount of variance in CLAS scores at test-time 1 (s²= 0.03, SE= 0.003, p<.001). However, this model was not a good fit for various reasons. As reported above, the lack of growth was not consistent with the general increase in CLAS scores between successive test-times. The fit statistics demonstrated a weaker fit relative to other models because the AIC, BIC, and sample-size adjusted BIC statistics were higher than other models. Finally, the chi-square test of model fit was significant, which suggests a difference between the model implied values and the data matrix (χ²(8)= 360.51, p<.001).

For the linear model (n=1152), test-takers received a mean CLAS score of 11.30 at test-time 1 (SE=0.02, p<.001) and grew an average of 0.03 of a complexity level between successive test-times.
(SE=0.004, p<.001). The significant variance in CLAS scores at the intercept \( (s^2_i=0.03, \ SE=0.003, p<.001) \) and slope \( (s^2_s=0.001, \ SE=0.00, \ p=.03) \) suggests variability in test-takers’ initial scores and average rates of growth per month. There was significant covariance between the intercept and slope, suggesting that for each additional complexity level attained at test-time 1, there was a slowing in the average rate of growth between test-times by 0.001 \( (COV_{i,s}=-.001, \ SE=0.00, p=.004) \). The average rate of growth was consistent with the general increase in CLAS scores between successive test-times (refer above). The fit statistics showed a sound fit relative to other models because this model was awarded the lowest AIC, BIC, and sample-size adjusted BIC statistics.

The chi-square test of model fit was not significant, which indicated no significant difference between the model implied values and the data matrix. While the statistical fit of the linear model was strong, it was not consistent with the research design because it assumed that the time between successive test administrations was equivalent.

For the quadratic model \( (n=1152) \), test-takers received a mean CLAS score of 11.30 at test-time 1 \( (SE=0.016, \ p<.001) \) and had an average rate of growth of 0.03 of a complexity level between successive test-times, however, this was not significantly different from 0 \( (SE= 0.02, \ p=.06) \). The quadratic function did not make a significant contribution to the model \( (q=0.001, \ SE= 0.01, p=.85) \). There was significant variance in CLAS scores test-time 1 \( (s^2_i=0.02, \ SE=0.01, p<.001) \), but nonsignificant variance in the slope \( (s^2_s=-0.01, \ SE=0.01, \ p=.18) \) and quadratic function \( (s^2_q=0.00, \ SE=0.00, \ p=.44) \). This suggests significant variability in test-takers’ initial scores, but no significant variability in their average rates of growth per month. There was significant covariance between the intercept and slope \( (COV_{i,s}=0.001, \ SE=0.003, \ p=.02) \) and between the intercept and the quadratic function \( (COV_{i,q}=-.002, \ SE=0.001, \ p=.001) \) but not between the slope and the quadratic function \( (COV_{s,q}=.001, \ SE=0.001, \ p=.20) \). This model was neither a sound statistical fit nor a good fit from the perspective of research design. The nonsignificant rate of growth was not consistent with the general increase in CLAS scores between successive test-times. The fit statistics showed a weaker fit than other models because of the higher values for AIC, BIC, and the sample-size adjusted BIC. The
chi-square test of model fit could not be computed. Finally, this model was inconsistent with the research design because it assumed that the time between successive test administrations was equivalent.

A piecewise model (n=1152) was constructed with a knot at test-time 2 such that the first linear model occurred between test-times 1 and 2 and the second linear model occurred between test-times 2 through 4. However, this model could not be identified even when the number of STARTS was increased to 1000, so it was not considered further. A piecewise model with a knot at test-time 3 demonstrated similar challenges.

For the time-variant model (n= 1084), test-takers received a mean CLAS score of 11.31 at test-time 1 (SE= 0.02, p<.001) and grew at an average rate of 0.006 of a complexity level per month (SE=0.001, p<.001) which suggests an average annualized rate of 0.07 of complexity level per year (12 x 0.006=0.07). There was significant variance in CLAS scores at test-time 1 ($s^2_i= 0.02$, SE= 0.003, p<.001) and in the average rate of growth per month ($s^2_s= 0.00002$, SE=0.0000001, p<.001). There was significant covariance between the intercept and slope, suggesting that for each additional complexity level attained at test-time 1, there was a slowing in the average rate of growth per month by -0.00016 (COV$_{is}=-.00016$, SE=0.00002, p<.001). This model was a sound statistical fit and a good fit from the perspective of research design. The significant rate of growth per month was consistent with the general increase in CLAS scores between successive test-times. The fit statistics showed a somewhat weaker fit relative to the linear model but a stronger fit relative to the intercept only and quadratic models because of the AIC, BIC, and sample-size adjusted BIC values. Finally, this model was consistent with the research design because it considered the actual time (in months) which passed between successive test administrations. While the linear model was the best statistical fit, the time-variant model was a sound statistical fit and more consistent with the research design. As a result, the time-variant model is presented in Figure 42 and was used for all subsequent analyses reported below.
Figure 42

Time-Variant Regression Model Applied to CLAS Scores

Moderation of the Intercept and by Demographic Variables (i.e., Sex, Sector, Language)

A series of moderation analyses were undertaken to check that the whole dataset could be analysed simultaneously to preserve statistical power. To determine whether participant characteristics were related to the initial CLAS score at test-time 1 (i.e., the intercept, $i$) or the average rate of growth per month (i.e., the slope, $s$), a moderation analysis was undertaken by regressing the intercept and slope on three dichotomous demographic variables i.e., sex (male, female), sector (public, private) and first language (English, non-English).

Sex did not moderate the intercept ($i$ on $\text{SEX}=.005, SE=0.013, p=.68$) but did moderate the average rate of growth per month ($s$ on $\text{SEX}=.002, SE=0.001, p=.01$). Sex ($n=914, n_{\text{male}}=561, n_{\text{female}}=353$) shared an intercept of 11.29 ($SE=0.022, p<.001$), but given the same initial score at test-time 1 females grew significantly more quickly than males (0.008 vs 0.006 of a complexity level per month, or 0.09 vs 0.07 of a complexity level per year).

Industry sector did not moderate the intercept ($i$ on $\text{SECTOR}=.063, SE=0.037, p=.08$) but did moderate the average rate of growth per month ($s$ on $\text{SECTOR}=-.003, SE=0.001, p=.01$). Both
sectors \( (n=1083, n_{\text{public}}=834, n_{\text{private}}=249) \) shared an intercept of 11.29 (SE=0.018, \( p<.001 \)), but given the same initial score at test-time 1, people working in the public sector grew more quickly than people working in the private sector (0.007 vs 0.004 of a complexity level per month, or 0.08 vs 0.05 of a complexity level per year).

First language did not moderate the intercept \( (i \text{ on LANGUAGE}=.008, \text{SE}=0.026, \text{p}=.77) \), nor the average rate of growth per month \( (s=.001, \text{SE}=0.001, \text{p}=.47) \). First language \( (n=870, n_{\text{English}}=767, n_{\text{nonEnglish}}=103) \) shared an intercept of 11.30 (SE=0.019, \( p<.001 \)), and given this same initial score at test-time 1, both groups grew at an average rate of 0.007 of a complexity level per month or 0.08 of a complexity level per year.

**Predictors of Growth in the Combined Sample**

To determine whether pedagogical characteristics of the development programs predicted growth rates, the slope, \( s \), was regressed on three variables: a) the total number of cognitive-developmental assessments completed by each leader (a simple count of assessments), b) the average amount of instruction time received (average hours/month), and c) formative versus summative application of cognitive-developmental assessments. The intercept, \( i \), was not employed in the regression model as test-takers were only exposed to pedagogical characteristics after test-time 1, so if they predicted CLAS scores at test-time 1 these findings would be uninterpretable.

Number of assessments completed \( (n=1023, M=2.24, s^2=1.21, \text{min}=1, \text{max}=6) \) significantly predicted the average rate of growth per month \( (s \text{ ON NUMBER ASSESS}=.002, \text{SE}=0.001, \text{p}=.03) \). This indicated that for every additional assessment taken, there was an increase in the average rate of growth per month by 0.002 of a complexity level (or 0.02 of a complexity level per year).

Average instruction time received per month \( (n=1023, M=9.92, s^2=112.79, \text{min}=1.43, \text{max}=47.62) \) did not significantly predict the average rate of growth per month \( (s \text{ ON} \)
HOURS_PER_MONTH=.00, SE= 0.00, p=.27). This indicated that developmental interventions which provided more contact time were not associated with significantly more growth.

Formative versus summative application of Lectical Assessments (n=1023, M=0.17, $s^2=0.14$) significantly predicted the average rate of growth per month (s on APPLICATION=-.01, SE=0.003, p<.001). On average, formative applications were associated with 0.011 of a complexity level more growth per month than summative applications (or 0.132 complexity level per year).

Relative Rates of Growth Between Developmental Providers

Even though causal attributions could not be made about growth, relative rates of growth between providers were compared. This analysis could only be undertaken at the scale of providers but not for programs or cohorts because the sample sizes were insufficient to produce meaningful results at lower levels of the data hierarchy. This was undertaken via two analyses. In Analysis 1, a formal statistical comparison was conducted between the two providers with sufficiently large samples i.e., Providers 5 and 12. In Analysis 2, an exploratory analysis was undertaken between all providers in which statistical constraints were introduced so that providers with considerably smaller samples could be included for exploratory purposes only.

Formal Statistical Comparison Between Two Providers with Larger Samples. Provider 5 administered assessments at test-times 1, 2, 3 and 4, whereas Provider 12 mostly administered assessments at test-times 1 and 2. As a result, the most meaningful way to compare rates of growth was to compare the average rate of growth per month between test-times 1 and 2 for the two providers. A total of 781 cases were available for this purpose: 546 from Provider 5 and 235 from Provider 12. Descriptive statistics for CLAS scores at test-times 1 and 2 are presented in Table 55.
Table 55

Descriptive Statistics for CLAS Scores at Test-Times 1 and 2 for Provider 5 and 12

<table>
<thead>
<tr>
<th>Provider</th>
<th>Test-time 1</th>
<th>Test-time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>546</td>
<td>11.30</td>
</tr>
<tr>
<td>12</td>
<td>235</td>
<td>11.29</td>
</tr>
</tbody>
</table>

For Provider 5, the intercept at test-time 1 occurred at 11.30 (SE=0.02, p<.001) and the average rate of growth per month was 0.007 of a complexity level (SE=0.002, p=.004) or 0.084 of a complexity level per year. There was a significant amount of variance around the intercept ($s^2_i=0.019$, SE=0.005, p<.001), but there was not a significant amount of variance around the slope ($s^2_s=0.00$, SE=0.00, p=.66). This suggests that participants entered the program with reasoning skills at different amounts of hierarchical complexity but experienced relatively similar growth trajectories. The covariance between the intercept and slope was nonsignificant (COV$_{i,s}=0.001$, SE=0.001, p=.51).

For Provider 12, the intercept at test-time 1 occurred at 11.30 (SE=0.017, p<.001), and the average rate of growth per month was 0.02 of a complexity level (SE=0.008, p=.02) or 0.24 of a complexity level per year. There was a significant amount of variance around the intercept ($s^2_i=0.013$, SE=0.004, p=.001), but there was not a significant amount of variance around the slope ($s^2_s=0.001$, SE=0.000, p=.06). As noted above, this suggests that participants entered the program with reasoning skills at different amounts of hierarchical complexity but experienced relatively similar growth trajectories. The covariance between the intercept and slope was nonsignificant (COV$_{i,s}=0.00$, SE=0.001, p=.68).

Analyses indicated no significant differences between the intercepts (Diff$_{i1,i2}=0.005$, SE=0.02, p=.83) and the slopes (Diff$_{s1,s2}=-0.013$, SE=0.01, p=.13) between providers. This was likely because of the size of the standard errors (SEs) relative to the size of the parameter estimates.
**Exploratory Comparison Between All Providers.** The sample size for some providers was more limited and could not be included in the formal statistical comparisons above. However, an exploratory analysis was undertaken in which the variance of the slope between test-times 1 and 2 was set to 0 following the suggestions provided by Muthén (1999). Given that the variance of the slopes for Providers 5 and 12 was not significantly different from 0, the application of this statistical constraint is not unreasonable. Because of these statistical constraints and high variability between group sizes, statistical comparisons between the slopes were not undertaken. Exploratory comparisons were informed by identifying non-overlapping confidence intervals between slopes for each provider (i.e., 1.96 x SE) (refer to Tables 56 and 57).

As can be seen from Tables 56 and 57, some providers appeared to be associated with growth (e.g., Providers 8 and 12), some providers appeared to be associated with stats (e.g., Providers 3 and 9), and some providers appeared to be associated with regression (e.g., Provider 10). Non-overlapping confidence intervals constructed around slopes suggest the possibility of different rates of growth between providers and/or different leader development cohorts.
Table 56

Descriptive statistics for CLAS Scores at Test-Times 1 and 2 by Provider

<table>
<thead>
<tr>
<th>Provider</th>
<th>CLAS score at time 1</th>
<th>CLAS score at time 2</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M</td>
<td>s²</td>
</tr>
<tr>
<td>Provider 1</td>
<td>43</td>
<td>11.39</td>
<td>0.03</td>
</tr>
<tr>
<td>Provider 3</td>
<td>24</td>
<td>11.37</td>
<td>0.02</td>
</tr>
<tr>
<td>Provider 4</td>
<td>7</td>
<td>11.55</td>
<td>0.02</td>
</tr>
<tr>
<td>Provider 5</td>
<td>546</td>
<td>11.30</td>
<td>0.04</td>
</tr>
<tr>
<td>Provider 6</td>
<td>20</td>
<td>11.40</td>
<td>0.03</td>
</tr>
<tr>
<td>Provider 7</td>
<td>70</td>
<td>11.27</td>
<td>0.04</td>
</tr>
<tr>
<td>Provider 8</td>
<td>57</td>
<td>11.41</td>
<td>0.04</td>
</tr>
<tr>
<td>Provider 9</td>
<td>12</td>
<td>11.43</td>
<td>0.03</td>
</tr>
<tr>
<td>Provider 10</td>
<td>40</td>
<td>11.33</td>
<td>0.03</td>
</tr>
<tr>
<td>Provider 11</td>
<td>85</td>
<td>11.15</td>
<td>0.03</td>
</tr>
<tr>
<td>Provider 12</td>
<td>235</td>
<td>11.29</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 57

Relation Between 95% Confidence Intervals Constructed Around Slopes Between Providers

<table>
<thead>
<tr>
<th>Provider No.</th>
<th>95% CI around the slope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Provider 1</td>
</tr>
<tr>
<td>Provider 1</td>
<td>0.005 - 0.005</td>
</tr>
<tr>
<td>Provider 3</td>
<td>0.000 - 0.000</td>
</tr>
<tr>
<td>Provider 4</td>
<td>0.007 - 0.007</td>
</tr>
<tr>
<td>Provider 5</td>
<td>0.003 - 0.011</td>
</tr>
<tr>
<td>Provider 6</td>
<td>0.001 - 0.013</td>
</tr>
<tr>
<td>Provider 7</td>
<td>0.005 - 0.009</td>
</tr>
<tr>
<td>Provider 8</td>
<td>0.007 - 0.011</td>
</tr>
<tr>
<td>Provider 9</td>
<td>0.000 - 0.000</td>
</tr>
<tr>
<td>Provider 10</td>
<td>-0.015 - 0.005</td>
</tr>
<tr>
<td>Provider 11</td>
<td>0.001 - 0.005</td>
</tr>
<tr>
<td>Provider 12</td>
<td>-0.003 - 0.029</td>
</tr>
</tbody>
</table>

Note. * Denotes non-overlapping confidence intervals between slopes
Hypothesis Tests for Study 4

- Hypothesis 1 was supported. The time-variant model appeared to be the best fit and had a significant slope, \( s \). This model indicated an average rate of growth of \( 0.006 \) of a complexity level per month, or \( 0.07 \) of a complexity level per year (i.e., \( 0.006 \times 12=0.07 \)).

- Hypothesis 2 was partially supported. The rate of growth in hierarchical complexity as determined by the slope, \( s \), was significantly moderated by some, but not all, pedagogical characteristics. Rate of growth was moderated by the number of cognitive-developmental assessments completed because for every additional assessment taken, there was an increase in the average rate of growth by \( 0.002 \) of a complexity level per month, or \( 0.02 \) of a complexity level per year (\( s \) ON NUMBER_ASSESS=\( .002 \), SE=\( .001 \), \( p=.03 \)). Rate of growth was also moderated by formative versus summative applications of cognitive-developmental assessments. Formative applications were associated with \( 0.011 \) of a complexity level more growth per month than summative applications, or \( 0.132 \) of a complexity level per year (\( s \) ON APPLICATION=\( -.011 \), SE=\( .003 \), \( p<.001 \)). Rate of growth was not moderated by the average instruction time received per month (\( s \) ON HOURS_PER_MONTH=\( .000 \), SE=\( .000 \), \( p=.27 \)).

- Hypothesis 3 was not supported. There was no significant difference in rates of growth between Providers 5 and 12 (\( \text{Diff}_{s1,s2}=-0.013 \), SE=\( .009 \), \( p=.13 \)). The possibility of different rates of growth between providers with smaller samples was explored via 95% confidence intervals, but this could not be tested statistically.

Discussion

Findings provide mixed evidence for the hypotheses. Findings suggest that the hierarchical complexity of leaders’ reasoning skills grew at an average rate of just under one-tenth of a complexity level per year during leader development programs. This rate of growth was shown to be moderated by some pedagogical characteristics, including the number of assessments taken and formative use of assessments, but not average monthly instruction time. Two of the largest
programs were positively related to growth, but no significant differences were found between rates of growth due to the size of the standard errors. Exploratory analyses on smaller samples suggested that different programs may be associated with different rates of growth in larger samples, but this could not be confirmed empirically. If real differences were found, this may have significant implications for our understanding of cognitive development over time as it could be inferred that either pedagogical design, participant characteristics, or the interaction between them, may impact growth trajectories.

The mean CLAS score at test-time 1 was 11.31 which occurs within Lectical Phase 11b, i.e., 11.25-11.49. As discussed in Chapter 2, CLAS scores are based on a discriminant function applied to the density of concepts at successive Lectical Phases. As a result, some leaders may have received higher CLAS scores at successive assessment administrations because of an increase in the density of concepts used within Lectical Phase 11b and preceding phases (i.e., 11a, 10d, etc.), but not necessarily because they started to use concepts at higher orders of abstraction.

Findings suggest that the process of reflective abstraction may have been activated. This explanation is consistent with the nature of the two pedagogical characteristics that were found to moderate growth. First, more frequent completion of assessments may have provided more reflective opportunities and, therefore, more opportunities to stimulate reflective abstraction. Second, the formative use of assessment may have introduced leaders to disequilibrating material that also stimulated reflective abstraction. Instruction time was not found to significantly predict rates of growth potentially because of a conflation between instruction time and practice time. It was assumed that increased instruction time may be associated with increased practice time, at least in a classroom setting. However, these variables may need to be considered separately in future research.

One of the main limitations of this study is the inability to attribute growth to the pedagogical characteristic of the programs rather than general life conditions or participant characteristics. Martin et al. (2021) outline six criteria which need to be satisfied by interventional
research to make causal attributions. First, the design must include an appropriate control
condition. If the growth of an experimental group is compared with a non-training group, it is not
possible to establish whether growth is causally attributable to the learning content or process.
Experimental groups should be compared with a control group in which the learning process and
learning content are varied independently. Second, the sample must be representative of the
population from which it is drawn. Third, leaders should be randomly allocated to experimental and
control conditions to ensure that conditions are equivalent and pedagogical and participant
characteristics are not conflated. Fourth, experimental and control conditions should be
independent in the sense that leaders allocated to one condition should not interact with leaders
allocated to another condition, otherwise, learning might be transferred from one group to another.
Fifth, consideration must be given to the appropriate timing of measurement intervals. Finally,
consideration must be given to mitigate researchers’ involvement in the delivery of development
processes to control for the potential impact of positive expectation bias.

Many, but not all, of these criteria were satisfied in Study 4. The design did not include a
non-training group, and growth was only compared between different providers. In a sense, each
provider functioned as a control condition against which other providers could be compared.
However, there may be reasons to believe that taking a cognitive-developmental assessment acts as
treatment because it requires operating on one’s knowledge, potentially transforming tacit
knowledge into declarative knowledge, which may impact scores at subsequent test-times.
Development processes and content were not varied independently, so it is not possible to make
causal attributions about the primary cause of growth. The sample was fairly representative of the
population from which it was drawn, given the sample size, cultures, and languages represented.
Since Lectica previously collected the sample, leaders were not randomly allocated to conditions,
meaning growth cannot be attributed to program or participant characteristics. Despite this,
conditions were independent because there was no interaction between leaders participating in
different programs. Timing of measurement intervals was accounted for by the time-variant model,
which analysed the average rate of growth per month even though the actual time between successive administrations varied from several months to several years between providers. The author of this thesis was not involved in the design or delivery of any development programs included in this sample which provides support for an appropriate level of objectivity.

Despite these limitations, it seems reasonable to suggest that in the absence of developmental catalysts such as disequilibrating learning material, cognitive development either slows down or arrests in adulthood. As a result, development programs may have played a role in leaders’ growth. Broader implications, limitations, and future research directions are discussed in Chapter 6.
Chapter 6: General Discussion

As discussed in Chapter 1, this thesis takes an interdisciplinary approach to pure and applied research. It is interdisciplinary in the sense that it draws on insights from the fields of leadership, adult development, and complexity science. It consists of pure research in the sense that it may contribute to our understanding of fundamental aspects of adult development and leadership phenomena. It consists of applied research in the sense that it applies a particular approach to adult development (i.e., cognitive-developmental theory and an associated form of measurement) to various aspects of leadership. As a result of this approach, this thesis aims to contribute to theory, empirical research, and practice pertaining to leaders’ ability to navigate organisational complexity. This aim was underpinned by three research questions which were addressed by four empirical studies.

These research questions are reiterated for readers’ convenience. **Research Question 1**: Do cognitive-developmental scores awarded by the Lectical Assessment System (LAS), and ego development scores awarded by the ego development scoring system, satisfy hard stage requirements? **Research Question 2**: To what extent does the hierarchical complexity of leaders’ reasoning skills satisfy the task demands of their roles? **Research Question 3**: To what extent do leaders develop the hierarchical complexity of their reasoning skills during participation in various leader development programs?

Given the number of studies undertaken to address these research questions, key findings are also summarised for readers’ convenience. Findings suggest that cognitive-developmental scores (operationalised by Lectical Phase Scores), but not ego development scores (operationalised
by SCT scores), satisfy hard stage requirements. Cognitive-developmental scores were provisionally demonstrated to be unidimensional, invariantly sequenced, qualitatively distinct, and structured wholes that exhibit spurts and plateaus of development, suggesting a succession of hierarchical integrations. By way of contrast, ego development scores were found to be invariantly sequenced. However, they appeared to violate all other hard stage requirements as tested on various samples and via different analytical approaches. These analytical approaches included Rasch analysis, multivariate statistics, scoring ego development scoring exemplars with the LAS, and scoring SCT protocols with the ego development scoring system and CLAS. Given the cumulative body of evidence presented, ego development scores may be a stronger reflection of a soft stage model which increases cumulatively in the number of perspectives taken or the number of unique words used by test-takers. Notwithstanding the need to rule out competing explanations for the findings on cognitive-developmental scores, they may be better suited to making inferences about leaders’ ability to navigate complexity. Table 58 provides a direct comparison of findings.

Given the properties displayed by cognitive-developmental scores, CLAS scores were used to compare the hierarchical complexity of leaders’ reasoning skills to the task demands of their roles and also to analyse longitudinal growth. Findings suggest a significant increase in the hierarchical complexity of reasoning skills from mid-leaders to upper leaders to senior leaders, but not from senior leaders to executive leaders. Findings also suggest a significant gap between the hierarchical complexity of leaders’ reasoning skills and the task demands of their roles, particularly for senior and executive-level leaders. Despite the plausible existence of the complexity gap, the findings suggest that leaders develop the hierarchical complexity of their reasoning skills during participation in leader development programs, even though direct causal attributions could not be made. The rate of development was shown to increase with more frequent and formative use of cognitive-developmental assessments but not with increased contact time. Given the size of the complexity gap and the rate at which hierarchical complexity was shown to develop in the sample analysed, leader development may need to be augmented via collective leadership development processes to reduce the impact of the complexity gap in individuals.
Table 58

Comparison of Findings Between Cognitive Development and Ego Development

<table>
<thead>
<tr>
<th>Step</th>
<th>Purpose</th>
<th>Analysis</th>
<th>Analysis name</th>
<th>Cognitive-Development scores operationalised by Lectical Phase Scores</th>
<th>Ego development scores operationalised by SCT scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Confirm if the whole dataset could be analysed simultaneously</td>
<td>1a</td>
<td>Facet analysis of scorer severity</td>
<td>Different scorers awarded equivalent scores to test-takers of equal ability (Study 1A)</td>
<td>Different scorers awarded equivalent scores to test-takers of equal ability (Study 1B)</td>
</tr>
<tr>
<td>1b</td>
<td>Invariance between versions of the assessment</td>
<td>Not applicable</td>
<td></td>
<td>Different versions of the SCT awarded similar scores to test-takers of equal ability (Study 1B)</td>
<td></td>
</tr>
<tr>
<td>1c</td>
<td>Retention of phases/stages items</td>
<td>All 5 LDMA items and 8 Lectical Phases from 10c through 12b could be retained for full Rasch analysis (Study 1A)</td>
<td></td>
<td>24% of stems (n=13) had to be deleted and 9 stages had to be collapsed to 7 stages before running full Rasch analysis (Study 1B)</td>
<td></td>
</tr>
</tbody>
</table>

2. Summary statistics

| 2    | Summary statistics | 2 | Summary statistics | Persons were broadly separated over a range of 33.14 logits, and average item difficulties were distributed over 1.35 logits (Study 1A). The similarity in item difficulties may result in compromised measurement properties in some contexts | Persons were narrowly separated over a range of 5.79 logits, and item difficulties were distributed over 1.53 logits (Study 1B) |

3. Direct tests of hard stage requirements

<p>| 3    | Direct tests of hard stage requirements | 3a | Unidimensionality | The Lectical Phase Scores exhibited negligible multidimensionality as demonstrated by a family approach to analysis (Study 1A). Age was not significantly correlated with CLAS scores (Study 2C). Unique number of perspectives did not significantly contribute to predicting CLAS scores (Study 2C). Unique word count made a statistically significant contribution to predicting CLAS | The SCT exhibited moderate multidimensionality in some subsamples as demonstrated by a family approach to analysis (Study 1B). Age was not significantly correlated with SCT Rasch scores (Study 2C). Unique word count and unique number of perspectives did significantly contribute to predicting SCT Rasch scores over and above that predicted by CLAS scores (Study 2C). SCT scores are unlikely to reflect a single |</p>
<table>
<thead>
<tr>
<th>Step</th>
<th>Purpose</th>
<th>Analysis</th>
<th>Analysis name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cognitive-Developmental scores</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>operationalised by Lectical Phase Scores</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>scores, but this was not practically significant (Study 2C). Scores along the Lectical Scale appear to reflect a single factor, hierarchical complexity</td>
</tr>
<tr>
<td>3b</td>
<td>Invariant sequence</td>
<td>Item thresholds and, therefore, Lectical Phase Scores 10c through 12b occurred in the theoretically predicted sequence (Study 1A)</td>
<td>Stem thresholds and, therefore, SCT scores occurred in the predicted sequence but only after considerable data cleaning (Study 1B). Mean scores for successive TPR distributions were statistically distinct and correctly sequenced (Study 2A). However, some findings were theoretically questionable because mean scores for some pairwise comparisons between successive TPRs occurred within the same stage (Study 2A)</td>
</tr>
<tr>
<td>3c</td>
<td>Qualitatively distinct</td>
<td>Most item thresholds were well aligned across LDMA items and created clear partitions between Lectical Phase Scores 10c through 12b. Some misalignment was noted for Lectical Phases 12a and 12b (Study 1A). However, this has already been resolved by Lectica. The reliability index (PSI= .97) was consistent with the claim that LAS can distinguish between 8 distinct Lectical Phases (Study 1A)</td>
<td>Most item thresholds were misaligned across stems and did not create clear separation between stages which meant that the amount of complexity reflected by each stage could not be identified (Study 1B). The reliability index (PSI=.93) suggested that the ego development scoring system can distinguish between 4 or 5 distinct stages (Study 1B). Two-sample Kolmogorov-Smirnov Tests indicated a significant difference between the location of successive TPR distributions (Study 2A). However, the Coefficient of Overlap (OLV) between successive TPR distributions demonstrated overlap in the range of 74-84%, which is uncharacteristic for scores based on hard stage models (Study 2A). LAS scoring the scoring exemplars for one SCT stem demonstrated that the ego development</td>
</tr>
<tr>
<td>Step</td>
<td>Purpose</td>
<td>Analysis</td>
<td>Analysis name</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>----------</td>
<td>---------------</td>
</tr>
<tr>
<td>3d</td>
<td>Structure of the whole</td>
<td>Category probability curves for Lectical Phase Scores exhibited apexes in the probability range .80–.90, which suggests that a single Lectical Phase organises each test-taker’s response to LDMA items (Study 1A)</td>
<td>Category probability curves for SCT scores exhibited apexes in the probability range .40–.60, which suggests that a single stage does not organise each test-taker’s response to SCT stems (Study 1B). Successive TPR distributions demonstrated that the modal stage accounted for 27-47% of the response distribution, which does not support the claim that the centre of gravity organises the whole worldview (Study 2A). Also, 95% confidence intervals constructed around the modal stage indicated that significantly less than 80% of the response distribution was accounted for by the modal stage (Study 2A). Finally, 95% confidence intervals constructed around the min and max value for successive TPR distributions showed that distributions exceeded a range of three which suggests that a single stage plus adjacent stages do not account for the response distribution (except for Expert) (Study 2A)</td>
</tr>
<tr>
<td>3e</td>
<td>Hierarchical integrations</td>
<td>Item equating procedures on Lectical Phase Scores displayed spurts and plateaus which may suggest a succession of hierarchical integrations between phases (Study 1A). Age was not significantly correlated with CLAS scores (Study 2C). Unique number of perspectives did not significantly contribute to predicting CLAS scores (Study 2C).</td>
<td>Item equating procedures on SCT scores did not demonstrate spurts and plateaus but did demonstrate smoothly monotonic growth. This does not suggest a succession of hierarchical integration between stages (Study 1B). Historical arguments that SCT scores operationalise hierarchically integrated stages appeared to rest upon a logical fallacy (Study 2B)</td>
</tr>
<tr>
<td>Step</td>
<td>Purpose</td>
<td>Analysis</td>
<td>Analysis name</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>----------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Implications for Theory and Empirical Research

As described in Chapter 1, Kohlberg’s conception of hard stages is a refinement of the formal properties that Piaget (1960, 1968, 1972) originally attributed to cognitive-developmental stages (Colby & Kohlberg, 1987; Kohlberg, 1969, 1975; Kohlberg & Armon, 1984). As discussed in Chapter 1, Stage S+1 is considered to be a hierarchical integration of Stage S if and only if Stage S+1 is defined in terms of two or more elements of Stage S, it organises and transforms the elements of Stage S, and does so in a non-arbitrary way (Commons & Pekker, 2008). This results in an emergent new structure, which enables performance that cannot be accomplished at Stage S alone. As discussed in Chapter 2 and Appendix A, cognitive developmentalists and ego developmentalists contend that their respective stages are hierarchically integrated.

Implications for ego development

From the perspective of theory, ego development theory was primarily developed through a combination of inductive and bootstrapping methods because the definitional attributes of stages emerged from the SCT (Cook-Greuter, 1999; Loevinger & Wessler, 1970), and other qualitative data such as biographical descriptions and first-person phenomenological experience (Torbert, 1972, 1987c, 2004). To reiterate, “The most distinctive feature of this theory, and of the instrument, is that the first evolved out of the second” (Cook-Greuter, 1999, p. 38) and “it is thus hard to separate the theory from the efforts to measure aspects of personality” (Cook-Greuter, 1999, p. 37). As discussed in Chapter 2, inductive methods were supplemented by bootstrapping processes via which theory and measurement mutually informed each other in order to improve the overarching account of ego development over time (Cook-Greuter, 1999; Loevinger, 1976). Perhaps one exception to these approaches may be Cook-Greuter’s (1999) use of abductive reasoning to the development of ego development theory by testing conceptions of the Construct-Aware and Unitive stages against other stage theories. On closer inspection, however, it is unclear that the theories employed by
Cook-Greuter have been demonstrated to comprise hierarchically integrated stages. For example, the Model of Hierarchical Complexity (Commons & Pekker, 2008) has been shown to comprise such stages. However, it is unclear whether this property can be extended to the six other theories Cook-Greuter used i.e., ego development theory (Loevinger, 1976), dialectical thinking (Basseches, 1984), constructive-developmental theory (Kegan, 1982), action-logics (Torbert, 1987c), and conceptions of systems and unitary stages (Koplowitz, 1984). It is possible that Cook-Greuter’s process influenced some versions of ego development theory and some versions of the ego development scoring system to approximate soft stages. Notwithstanding this potential exception, the inductive and bootstrapping methodologies that have been used to construct ego development theory may result in a uniquely tight coupling between theory and measurement. As a result, the findings reported on measurement of ego development (refer Table 58) may imply that the central tenets of ego development theory may need to be revisited. Perhaps the theoretical conceptions of ego development stages might benefit from being tested against other theories that have been empirically shown to consist of hierarchically integrated stages. If the findings reported on the measurement of ego development also impact ego development theory, then these findings may not only imply limited relations between ego development and cognitive-developmental scores, but they may also imply limited relations between ego development theory and cognitive-developmental theory.

From the perspective of measurement, it is underdetermined whether the hard stage postulates of ego development theory have been unsuccessfully operationalised for the purpose of measurement, or whether soft ego development stages have been operationalised for the purpose of measurement with a moderate degree of success. In either scenario, ego development scores do not seem to be well suited to making inferences about leaders’ ability to navigate complexity. Given that ego development scores seem to be a stronger reflection of the number of perspectives taken or the number of unique words used by test-takers, perhaps a more parsimonious approach to measuring ego development that may also strengthen the unidimensionality of the SCT, might focus
on directly counting the breadth of perspectives that test-takers consider when projecting their worldviews on sentence stems.

There may be several other implications related to measuring ego development via the SCT. Irrespective of the particular construct measured by the SCT, unidimensionality seems to be significantly compromised in some samples (refer Studies 1B and 2C). This may present difficulties when using total scores (e.g., TPRs) in the context of empirical research and/or workplace practice because a total score may not adequately reflect the construct in question. Furthermore, ego development stages do not seem to reflect a consistent magnitude of development due to their inconsistent location along the logit scale (refer Study 1B). This may suggest that ego development stages have different meanings in different contexts. For example, getting Expert on one set of SCT stems may have little in common with getting Expert on a different set of stems. This may also present difficulties in empirical research and/or workplace practice because this may imply that stages cannot be consistently used to signify an agreed amount of development.

Notwithstanding these implications for measurement, the fact that ego development scores predict some complexity-related leadership outcomes needs to be accounted for. Given that ego development scores have been shown to be a stronger reflection of randomness (80%) than ego development (20%) in earlier studies (Blasi, 1971; Kishton et al., 1984; Lambert, 1972a, 1972b; Loevinger & Wessler, 1970; Lorr & Manning, 1978), perhaps randomness itself predicts complexity-related leadership outcomes. In light of the chaotic ways that social systems such as teams and organisations behave in VUCA conditions (Helbing, 2013; Snowden & Boone, 2007), the possibility that somewhat random responses demonstrated by leaders may achieve positive outcomes for broader social systems may not be unreasonable. A second explanation is that if the ego develops towards a broader range of perspectives rather than more hierarchically complex worldviews, then leaders’ capacity to hold multiple perspectives may impact their ability to lead in complex conditions. For example, perhaps the ability to work with multiple competing perspectives in a team environment might contribute to leaders’ collaborative capacity and other leadership skills. A third
explanation is derived from the definitional attributes of ego development stages in which the later stages are characterised by warmer, more aware, and more sagely interpersonal styles (Cook-Greuter, 2000; Torbert, 2013). Perhaps there is something unique about the presence of such a leader that may create conditions in which teams can perform more effectively.

In general terms, there are two significant considerations that may be brought to bear on ego development. The first, as discussed above, is an apparent commitment to inductive and bootstrapping approaches to theory construction. The second, as discussed in Chapter 2, is the logic underpinning the sequence of ego development stages. To briefly recap, Cook-Greuter (1999) contends that stages are sequenced according to the acquisition and deepening of more cognitively oriented grammatical perspectives called perspectives on the self. Torbert (1994, 2004) contends that stages are sequenced according to the acquisition and coordination of phenomenological perspectives that arise in our awareness called territories of experience. Both considerations may converge on the potential conclusion that ego development more strongly resembles a perspectival theory rather than a hierarchically integrated set of stages that are sequenced in an ascending order of complexity. This potential implication may corroborate earlier research that suggests that perspective taking may have less in common with hierarchical complexity than what has typically been assumed in the adult development literature (Fuhs, 2016). This does not preclude the possibility that some forms of perspective acquisition may result in a sequence of hierarchically integrated stages (Selman, 1980). However, the manner in which perspectives are acquired and deepened in the context of ego development does not seem to result in stages that consistently increase in hierarchical complexity.

Considering the implications for ego development theory and measurement jointly, it is argued that ego development cannot clearly be considered an instance of ‘vertical development’ because there seems to be relatively little evidence to suggest that these stages are organised vertically, see Petrie (2014) for example. Hard stage claims about ego development may not meet reasonable standards of epistemic adequacy because they do not appear to be adequately represent
the information that is actually available about ego development. For ego developmentalists to mount a counter-argument against this conclusion, they would need to justify what they consider to be vertical about a set of stages that have not consistently been shown to be hierarchically integrated. This is not to suggest that ego development is inferior to cognitive development as a general approach to adult development. It is to suggest, however, that ego development stages may not exhibit a particular property and therefore may be ill-suited to applications related to complexity in particular. While ego development seems to account for some form of sequential development, perhaps the ego develops towards a broader range of perspectives rather than more hierarchically complex worldviews.

**Implications for cognitive development**

As discussed in Chapter 2, the patterns identified in scores that reflect hierarchically integrated stages are not typically found in scores that reflect other types of ordinal scales. Scores at the scale of whole complexity levels have been shown to satisfy hard stage requirements (Commons, Li, et al., 2014; Dawson et al., 2005), but the properties of scores at a lower level of granularity have not been consistently evaluated in the published literature. Based on the findings reported in this thesis, Lectical Phase Scores have been provisionally demonstrated to exhibit these patterns and this may result in significant implications for cognitive-developmental theory and empirical knowledge.

Needless to say, one psychometric study is insufficient to clearly establish implications for theory. This is a particularly important point because unlike ego development theory, the construction of cognitive-developmental theory does not primarily depend on inductive methods. As a result, inferences from measurement to theory are offered more tentatively. Patterns identified in scores could be a function of cognitive-developmental theory, the scoring system employed, scoring protocols, the properties of items, and other related factors. If the findings reported in Study 1A are supported by additional evidence and also if the three competing explanations for findings discussed in Study 1A could be ruled out, then perhaps patterns identified
in scores could be used to identify potential implications for theory. It may be tentatively suggested that the unique patterns identified in cognitive-developmental scores may imply that hierarchical integrations—or processes akin to hierarchical integrations—inhere at a lower level of granularity than whole complexity levels. This may be consistent with the postulates of related theories in which smaller increments of development have been proposed to account for the transformation of performance from one complexity level to the next (Commons et al., 2005; Commons & Richards, 2002; Fischer, 1980; Fischer & Bidell, 2006; Lahey et al., 2001; Ross, 2008a). This may also be consistent with the contention that hierarchical complexity is the kind of construct that may exhibit fractal-like patterns of repeating self-similarity at multiple levels of granularity, as discussed in Chapter 2 (Commons & Richards, 2002; Ross, 2008a).

From the perspective of measurement, identification of these unique patterns may suggest that the theoretical postulate that cognition develops hierarchically has been successfully operationalised for the purpose of measurement. Not only this, but it appears that cognitive-developmental scores may be well suited to making inferences about leaders’ ability to navigate complexity and that increments of development that are smaller than whole complexity levels can be detected. In the context of empirical research, this may imply that finer within-person and between-person distinctions can be made in cognitive development and these distinctions may be predictive of other leadership phenomena and/or may be used to identify differences in interventional and longitudinal studies.

Given these implications for cognitive development, findings reported for Study 3 may suggest that the increase in 21st century complexity has outstripped the pace at which leaders’ reasoning has developed. This may provide confirmatory evidence of the unmanaged systemic complexity that Bell (1973) and other scholars predicted some 50+ years ago. If the complexity gap were shown to generalise across industries, geographies, and leadership roles, this might also support Bennis’s (2007, p. 2) contention that “... the four major threats to world stability are a nuclear/biological catastrophe, a world-wide pandemic, tribalism, and the leadership of human
institutions. Without exemplary leadership, solving the problems stemming from the first three threats will be impossible”. The need to address the complexity gap is underscored by the observations that reasoning skills are one of the strongest predictors of workplace performance (Schmidt et al., 2016; Schmidt et al., 2008) and also that social systems exhibit a general tendency to increase in complexity over time, which may widen the complexity gap even further (Nolan & Lenski, 2008; Stewart, 2014).

The findings reported in Study 3 may imply that cognitive-developmental approaches can be used to operationalise nascent theorising about leader complexity by conceptualising and measuring relations between leaders’ cognitive abilities and the requirements of their social contexts. For example, scales of hierarchical complexity could be used to operationalise multiple aspects of the Leaderplex Model (Hooijberg et al., 1997), leader complexity (Hannah et al., 2009), requisite complexity (Hannah et al., 2011; Lord et al., 2011), and other models of leadership complexity (Day & Lance, 2004). In the context of the Leaderplex Model, social complexity and cognitive complexity are predicted to impact behavioural complexity, which is predicted to affect leader effectiveness and organisational performance (Hooijberg et al., 1997). As Day and Lance (2004) argued, the Leaderplex Model fails to specify the content of these constructs. Scales of hierarchical complexity could be used to operationalise cognitive and social complexity (Commons et al., 1993) to study relations between constructs included in the model. In the context of leader complexity, leaders’ ability to navigate organisational complexity increases when leaders’ self-concept is complex and positive (Hannah et al., 2009). Scales of hierarchical complexity do not reflect the positivity of leaders’ self-concept, but they could measure the complexity of leaders’ self-concept. In the context of the requisite complexity model, leader effectiveness is predicted to emerge when leaders’ responses are commensurate with the complexity of their organisational challenges (Hannah et al., 2011; Lord et al., 2011). Scales of hierarchical complexity could be used to operationalise general cognitive complexity, social complexity, self-complexity, and dynamic complexity. There are
significant advantages to be gained, considering the limited empirical research informed by this line of theorising.

Finally, the findings reported in Study 4 may have implications for our understanding of the theoretical mechanisms that explain cognitive growth in leaders. As indicated in Study 4, leader development programs based on formative applications of cognitive-developmental assessments were included in the sample, and these programs were more likely to result in growth because they incorporated learning material that is aimed at a higher level of hierarchical complexity. While methodological limitations prevented direct causal attributions from being made, Study 4 may imply that reflective abstraction played a role in catalysing statistically significant growth in hierarchical complexity. As discussed in Chapter 2, reflective abstraction is a hypothetical process originating in cognitive-developmental theory via which lower-order cognitive structures are transformed into qualitatively distinct higher-order structures, when lower-order cognitive structures are disequilibrated by increasingly complex information (Piaget, 2000). A considerable amount of literature on leader development recommends that leaders should be provided with ‘challenging experiences’ to catalyse growth (Day et al., 2009; Day & Sin, 2011; Manners & Durkin, 2000; Van Velsor & Drath, 2004). Study 4 may imply that the efficacy of such programs may be partially attributable to reflective abstraction, and that leader development programs may not necessarily need to be challenging with respect to their volume, pace, risk, or profile but specifically with respect to their ability to disequilibrate leaders’ existing cognitive structures. This may also imply reflection abstraction should play a more prominent role in the field of leadership so that empirical research can contribute to our understanding of the learning conditions that may foster optimal rates of growth.

**Implications for Practice**

In general terms, the findings on ego development suggest that it positively violates most hard stage requirements (refer to Table 58). This may suggest that ego development is poorly suited to
workplace applications that are specifically related to complexity. Continuing to use ego
development for such applications may surface some potential ethical implications. If additional
evidence confirmed that the SCT fails to measure a unidimensional construct and also that the SCT
fails to consistently yield information about hierarchical complexity, then using the SCT to provide
workplace leaders with feedback about their ability to navigate complexity, may result in providing
them with inaccurate information about their own development. If, as discussed in the preceding
section, findings on the measurement of ego development have implications for the central tenets
of the theory, then using ego development theory to help workplace leaders make-sense of their
ability to navigate complexity, may also result in the provision of inaccurate information, even if the
SCT is not administered. It may be tentatively suggested that both cases could result in practice that
is ineffective or harmful to leaders and/or their organisations. This may occur if leaders are offered
formal roles or informal opportunities that require a level of complexity that they are not equipped
to handle. Alternatively, if leaders are offered inaccurate feedback about their ability to function in
complexity, it is possible that this may adversely impact their self-concept and performance. This is
not to suggest, however, that a revised version of ego development theory and the SCT may not be
useful for a variety of other leadership applications that are less related to complexity in particular
e.g., fostering interpersonal maturity.

In general terms, the findings on cognitive development suggest that cognitive-developmental
approaches are well suited to workplace applications that are specifically related to complexity. If
this were supported by additional evidence, then this may imply that cognitive-developmental
approaches may be applied to identify and/or mitigate the complexity gap in workplace contexts
through effective human capital management.

In comparing leaders’ reasoning to the task demands of their roles, a considerable
proportion of senior leaders appear to have been appointed to roles that are one or two
management layers beyond their cognitive ability. The findings presented in Study 3 may imply that
leaders who exhibit the complexity gap may be downward assimilating the inherent complexity of
their roles to the level of hierarchical complexity at which they can reason. This may be analogous to Piaget’s experiment on the conservation of volume that was discussed in Chapter 4, because some leaders may be unaware of the fact that their reasoning skills do not allow them to satisfy the task demands that are constitutive of their roles. If the finding of a statistically significant complexity gap were supported by the weight of additional evidence, this may go some way towards explaining why ‘wicked organisational problems’ often resist leadership solutions (Edmonstone et al., 2019; Rittel & Webber, 1973). Practical examples of how these wicked problems may be perpetuated include sending employees to training courses rather than creating cultural conditions for optimal learning (Kjellström & Mitchell, 2019), implementing organisational restructures rather than holistically addressing problematic business models, and trying to change the behaviour of individuals rather than addressing the systemic dynamics that lead to the emergence of organisational culture. Potential symptoms of the complexity gap from the empirical literature may include the high stress levels reported by many leaders (Harms et al., 2017), the limited number of HR professionals who rate their organisational leadership favourably (DDI, 2021; Gahan et al., 2016), and the common occurrence of failed organisational change and culture transformation (Decker et al., 2012; Smith, 2003). Dóci and Hofmans (2015) provide an instructive example in which participants were asked to assume the role of a CEO and lead a simulated business meeting. As task complexity was increased experimentally, followers’ perceptions of transformational leadership significantly decreased.

Given the size of the complexity gap and the rate at which cognition was shown to develop, a more expedient way of addressing the complexity gap might be to mitigate its presence before it emerges through effective human capital management (Noe et al., 2017). The high level of hierarchical complexity characterising the task demands of leaders’ roles may reduce the pool from which suitable candidates can be drawn. Notwithstanding this challenge, considering leaders’ reasoning skills may contribute to recruitment, onboarding, talent identification, and succession planning processes. Candidates with reasoning skills within a role complexity range could be short-
listed for further evaluation in a recruitment process. Small complexity gaps might inform onboarding plans for the first six-to-twelve months of leaders’ incumbency. The hierarchical complexity of leaders’ reasoning might play a critical role in identifying future talent or successors for particularly complex roles. Given the findings from Study 3, there appears to be a non-significant difference in the cognitive ability of senior versus executive leaders, and the complexity gap seems to emerge at senior leadership layers. This may imply that some candidate pools lack senior leaders with the requisite cognitive skills and/or that some senior appointments are made on the basis of considerations that are unrelated to leaders’ cognitive skills. From the perspective human capital management, individual leader development or collective leadership development processes may become increasingly relevant as a result. In this context, collective leadership development refers to a form of development that focuses on socially distributed leadership and may include practices such as building social networks, various job assignments, action learning projects, collaborative decision-making, and so on (Day, 2000).

Meta-analyses suggest that leader development processes typically focus on achieving behavioural and psychological outcomes (Avolio et al., 2009; Burke & Day, 1986; Collins & Holton, 2004; Day & Sin, 2011; Lacerenza et al., 2017; Reyes et al., 2019; Rosch et al., 2017). Behavioural outcomes include adopting more constructive behavioural styles and psychological outcomes include understanding personality preferences or strengthening self-efficacy. In the context of leader development, it may be possible to design practices which integrate behavioural change with the development of hierarchical complexity. For example, as leaders hold coaching conversations (Peláez et al., 2020), they could practise new behavioural styles while simultaneously practising advanced reasoning skills. These practices could include suspending personal attributions about challenging interpersonal dynamics while seeking a range of perspectives that may inform an authentic understanding of root causes. For example, biological, psychological, behavioural, interpersonal, systemic, cultural, contextual, epistemic, and ethical causes could be considered to avoid making attribution errors. These practices could be selected on the basis of relatively granular
cognitive-developmental scores given the properties they exhibited in Study 1A. Stålne et al. (2016) provide an instructive example of how the Model of Hierarchical Complexity can be used to evaluate the complexity of developmental curricula and learning outcomes. As a result, practitioners may create more optimal conditions for leader development by providing leaders with practices tailored to their learning needs. These practices may be experienced as more palatable and developmentally efficacious than practices aimed one whole complexity level above their current level of reasoning, which may be too challenging for them to grasp. This line of reasoning may have significant implications for the pedagogical characteristics of developmentally efficacious leader(ship) development programs. Rather than simplifying curricula in order to engage busy workplace leaders who are too time-poor to participate in leader development processes, these considerations may imply that curricula must be sufficiently complex to engage and disequilibrate leaders’ existing cognitive structures to foster their long-term growth.

Study 3 did not suggest the presence of a complexity gap in more junior leaders, however, Karlton et al. (2021) have started to explore ways that first-line managers can develop resilient action strategies in order to deal with unstable and variable organisational conditions in specific industry sectors. Resilient action strategies included defining the focus, making a diagnosis, and defining objectives while establishing activity perspectives, identifying various perspectives on work, and favouring different perspectives. From the perspective of practice, it may be useful to determine if, and how, resilient action strategies could be applicable to more senior leaders who are experiencing a complexity gap and if these strategies may play a role in fostering growth and/or mitigating the impact of the complexity gap.

Given the rate at which cognition was shown to develop in Study 4, the complexity gap might only be closed for leaders who exhibit a relatively minor gap or are adept at learning new reasoning skills, and only after several years of practice. Furthermore, there are likely to be neurological and contextual constraints that impose limits on the level of hierarchical complexity that individuals might attain (Collins et al., 2010; Commons et al., 1993; Fischer, 2008; Fischer et al.,
For this reason, leader development may be a necessary but not sufficient condition to close the complexity gap and could be augmented by other collective leadership development practices such as collaborative decision-making (Owen & Buck, 2020), the use of big data, and applications of artificial intelligence (McAfee & Brynjolfsson, 2012).

Eva et al. (2021) describe a multi-perspectival framework of collective leadership development which integrates five main theoretical perspectives on collective forms of leadership: person-centred, social network, social-relational, sociomaterial, and institutional. Person-centred perspectives focus on the attributes of individual leaders (e.g., mindsets, behaviours, etc.) and the way these attributes foster the emergence of leadership. Social network perspectives focus on the structure of interpersonal dynamics between two or more leaders and the way these dynamics contribute to collective leadership. Social-relational perspectives focus on intersubjective processes of shared leadership that construct meaning, relationships, and influence within a broader collective. Sociomaterial perspectives focus on leadership as a process that emerges through the dynamic interaction between social, material, and discursive relations (e.g., practices, routines, etc.) Institutional perspectives focus on leadership as a process that emerges as leaders iteratively negotiate relations and polarities that emerge in organisations as a result of temporal changes that unfold over time e.g., the evolution of core values. Eva et al. (2021) describe how this framework can provide a theoretical foundation to identify practices that may develop collective forms of leadership. For example, the person-centred approach is associated with training individuals in relevant competencies, the social network approach is associated with structural interventions that foster a stronger sense of the collective e.g., collaborative technology, and so on. An example of an interpersonally oriented practice that is gaining popularity in the practitioner literature is sociocracy which includes the integration of diverse perspectives, the construction of psychologically safe environments, consent-based decision-making, and applications of principles such as dynamic steering (Buck & Villines, 2007). Some of these collaborative forms of leadership development may
play a role in mitigating the impact of the complexity gap in individuals because they transcend practices that focus on individual leader development (Cullen et al., 2012).

Whether leader or leadership development practices are employed, Study 4 suggests that growth may be accelerated by providing leaders with meaningful reflective opportunities that enable them to operate on increasingly complex learning material. Reflective opportunities may be implemented via the formative use of well-constructed assessments or commonly used reflective practices such as journaling. Journaling may include engaging with prompts that invite leaders to reflect on their observations about their organisations, professional relationships, and internal psychological experiences (Raymer et al., 2018; Torbert & Fisher, 1992). The finding that instruction time was not significantly related to rates of growth suggests that labour-intensive programs with high amounts of contact time may be unnecessary—particularly if they include opportunities for deliberate daily practices (D. Day, 2010).

**Broader Implications for the Fields of Adult Development and Leadership**

As discussed in Chapter 1, there are three main points of intersection between the fields of leadership and complexity science: complexity leadership theory (Uhl-Bien et al., 2007), leader complexity theory (Hannah et al., 2009; Hooijberg et al., 1997), and adult development theory (McCauley et al., 2006). In the context of this thesis, adult development theory has been used as primary a lens through which research questions have been addressed (Kjellström & Stålne, 2017). Notwithstanding the affordances of adult development theory, the published literature (McCauley et al., 2006) and the general discourse in adult development communities (Petrie, 2014) may have implicitly accepted several ‘big assumptions’ that may have limited the applicability of adult development theory to leadership studies:

- **Big Assumption 1:** Adult development stages reflect an underlying construct of complexity
- **Big Assumption 2:** Stages from one theory directly correspond to stages from another theory (this may be a corollary of assumption 1)
• Big Assumption 3: Choosing between one theory and another is a matter of personal preference (this may be a corollary of assumption 1)

• Big Assumption 4: Adult development assessments are directly measuring complexity or constructs related to complexity

For example, Big Assumption 2 is implicitly expressed in tables of correspondence which assume a direct mapping between one set of stages and another (Commons et al., 1990; Cook-Greuter, 1990; Fisher et al., 1987; McCauley et al., 2006; Wilber, 2000). Tables of correspondence have proliferated throughout the adult development literature, but they may depend on a questionable evidence base.

A particular developmental practice that has been adopted by practitioner communities emphasises the importance of surfacing and explicitly testing the big assumptions that drive different courses of action (Kegan & Lahey, 2009). There is a meaningful sense in which the arguments presented in this thesis have explicitly tested some of the big assumptions listed above, and the findings may provide fertile ground for fostering the growth of the field of adult development itself. As discussed in Chapter 4, some forms of growth are contended to occur through the processes of differentiation and integration (Akrivou, 2008; Johnson, 2000; Kjellström & Sjölander, 2014). Perhaps the field of adult development could be benefit from submitting itself to iterative cycles of differentiation and integration. While it is broadly recognised that the field comprises different adult development theories (Demick & Andreoletti, 2003; Robinson, 2013), it may not be so broadly recognised that fundamentally different structures (e.g., nominal, cumulative, hierarchical, etc.) may account for the distinctions between different theories. An adapted version of the transition steps described by the Model of Hierarchical Complexity may illustrate how the processes of differentiation and integration may contribute to the growth of the field (Commons & Richards, 2002; Ross, 2008a):
• Transition step 0 (thesis): Theory 1 is best suited to making sense of leadership in general

• Transition step 1 (deconstruction of thesis): Theory 1 is limited in its ability to make sense of complexity-related leadership phenomena

• Transition step 2 (antithesis): Theory 2 may account for complexity-related leadership phenomena

• Transition step 3 (relativism): In the context of some other leadership outcomes, Theory 1 may be a sound functional fit. In the context of complexity-related leadership phenomena, Theory 2 may be a sound functional fit

• Transition step 4 (smash): By simultaneously coordinating Theories 1 and 2, a wide variety of leadership outcomes can be accounted for, and these outcomes include, but are not limited to, complexity-related leadership phenomena

• Transition step 5 (hierarchical integration): The field enquires into the emergent property of positive adult development. We can make sense of this emergent property by coordinating meaningful relations between adult development theories that differ with respect to their structure (e.g., nominal, cumulative, hierarchical, etc.) and not only with respect to their content.

Ushering a dialogue about adult development theory through these transition steps may create a more intellectually spacious conversation at the scale of the field. If this were considered a tenable path of exploration, then cognitive-developmental theory and ego development theory could be used in tandem to account for fundamentally different leadership phenomena in a more inclusive scholarly manner, rather than being pitted against each other with respect to claims about complexity. For example, cognitive-developmental theory could be used to account for leaders’ ability to navigate complexity while ego development theory could be used to account for distinct worldviews, thereby yielding a more integrative psychological profile. For example, organisational leaders may benefit from conventional worldviews and high complexity levels in order to reason in
sophisticated ways about organisational performance, while philosophers may benefit from postconventional worldviews and high complexity levels in order to reason in sophisticated ways about existential issues. If these considerations were shown to generalise to other adult developmental theories that were not considered in this thesis, then perhaps some theories may have relatively little in common apart from including ordinal categories that happen to be referred to as ‘stages’. If post-formal stages of hierarchical complexity enable leaders to navigate wicked social and environmental issues as contended by some scholars (Commons & Ross, 2008c; Inglis, 2008), then recognition of these structural distinctions may be important.

These considerations may have significant implications for the field of leadership because they may yield insight into fundamentally different types of development that leaders are subject to, and how these types of development may account for different leadership phenomena. These considerations may also imply that Big Assumptions 1 through 4 listed above, are rooted in questionable intellectual ground. If Assumptions 1 through 4 were challenged by additional evidence, then this may provide a context in which to make sense of several findings reported in this thesis. For example, structural distinctions between ego development stages and cognitive-developmental stages may reconcile tensions between the claim that ego development accounts for broader personality changes across the lifespan (Cook-Greuter, 1990, 1999; Loevinger, 1976; Torbert, 2004) and findings reported in Study 2B. For example, if the relation between ego development and cognitive development was contended to be relatively orthogonal rather than isomorphic, then it would make sense that growth hierarchical complexity occurs within rather than between worldviews; that ego development stages span a narrow range of complexity levels; and that the latest stages of ego development may reflect highly deconstructed egos, but only moderate levels of hierarchical complexity (refer Study 2B). This line of reasoning is consistent with earlier research which suggested that two people who perform at the same level of hierarchical complexity can be classified at distinctly different development stages in the context of other types of stage models (Dawson & Heikkinen, 2009).
As discussed in Chapter 1, the field of leadership typically conceives of leadership as a process of social influence, which has a positive impact on the social systems in which it is nested (Antonakis & Day, 2018; Bass, 2008; Day, 2012). However, a negligible amount of research has focused on contextual, adaptive, and complexity-related leadership theory (Gardner et al., 2020). In general terms, this thesis provides support for arguments that adult development affords a unique lens on leadership studies (Kjellström & Stålne, 2017). More importantly, this thesis surfaces new implications about distinctions between fundamentally different types of adult development, and that only some of these types of development may be used to foster insights into leaders’ ability to navigate the complexity that is inherent in their social contexts.

Limitations and Future Research Directions

General considerations

The conclusions drawn in this thesis are constrained by several methodological limitations, which highlight possibilities for future research. The early literature attempted to measure ego development via alternative approaches (Hansell et al., 1984). However, the SCT appears to be the one of the few options for measuring ego development in the contemporary literature. This does not preclude the possibility of developing new—and potentially more effective—ways of measuring ego development in the future. For example, some scholars have developed surveys about value systems that consist of items that may be conceptually related to aspects of ego development and data yielded by these surveys may exhibit patterns which are similar to those observed in the SCT (Kjellström & Sjölander, 2014; Kjellström et al., 2017; Sjölander et al., 2014). By contrast, Lectica’s scoring systems are among several options for measuring cognitive-developmental complexity levels. The generalisability of the findings reported in this thesis should be tested by performing similar analyses on scores yielded by related scoring systems. Some of these scoring systems can only award scores at the scale of whole stages. This may present a challenge because only scores at a lower level of granularity may detect complexity gaps and growth over short-to-medium
timescales. After undertaking further psychometric validation, future research may take advantage of the seven transition steps and substeps awarded by Common’s revised version of the Hierarchical Complexity Scoring System (HCSS) (Commons et al., 2005; Commons & Richards, 2002; Ross, 2008a), or the five transition steps awarded by the subject-object interview scoring system (Lahey et al., 2001).

As discussed in Chapter 5, the cognitive literature indicates a clear relation between cognition and leadership phenomena (M. D. Mumford et al., 2015). However, many of these studies have relied upon cumulative conceptions of cognition via traditional IQ tests and domain-specific problem-solving tasks (Hooijberg et al., 1997; Mumford et al., 2000; Schmidt & Hunter, 1998; Zaccaro et al., 2013). As discussed in Chapter 1, hard stages are not necessarily the only or best way of conceptualising and measuring complexity. As a result, future research should address the predictive validity of cumulative versus hierarchical conceptions of complexity relative to criterion variables that are typically considered in the leadership literature. As also discussed in Chapter 1, there are sound conceptual arguments as to why hard stages may account for a unique proportion of variability in leadership phenomena, but the empirical argument has been more challenging to establish. Torbert’s (2013) and Commons’ (2006) research on the external and predictive validity of developmental assessments relative to leadership-related criterion variables are notable exceptions.

Clear relations between hierarchical complexity and leadership phenomena, such as leader emergence, leader role occupancy, leader performance, leader effectiveness, and other adaptive forms of leadership are critical to establish. The importance of establishing relations between these variables is underscored by the possibility that more traditional approaches to measuring reasoning skills, such as IQ, may not be able to detect forms of reasoning that go beyond the structural equivalent of Lectical Level 11 (Leite et al., 2016).

In the context of leadership research, several of the criterion variables listed above (e.g., leader role occupancy) have been shown to be predicted by attributes, experiences, and opportunities earlier in life (Day, 2011; Liu et al., 2021). For example, academic intrinsic motivation in childhood
and adolescence has been shown to predict motivations to lead in adulthood (Gottfried et al., 2011); a temperamental approach/withdrawal in childhood and extraversion in adolescence has been shown to predict leadership potential in adulthood (Guerin et al., 2011); and positive family functioning during adolescence has been shown to predict the emergence of transformational leadership in adulthood (Oliver et al., 2011). In the context of research on hierarchical complexity, education has been shown to be a sound predictor of cognitive attainment later in life (Dawson et al., 2005; King & Kitchener, 1994; Kitchener & King, 1981). Some preliminary evidence may suggest that clarity of reasoning may also predict the rate of growth in cognitive development over time (Dawson & Seneviratna, 2015). However, it would be useful for future research to investigate whether other experiences and opportunities might predict the complexity level at which adult development may stabilise, the rate at which hierarchical complexity may grow in adulthood, and how these factors relate to leadership outcomes later in life.

In addition to exploring relations between hard stages and leadership, future research should be directed towards understanding the boundaries that apply to claims based on adult development theories and measures. Adult development assessments are typically used to make inferences about leaders’ psychological processes or conditions occurring in their mind-brains (Leite et al., 2016). As a result, these assessments are typically interpreted to reflect leaders’ ability to navigate complexity, their internal meaning-making systems, their internal worldviews, and so on. However, it does not follow that because leaders are able to navigate complexity, that they actually demonstrate a behavioural propensity to do so. For example, meta-analyses have demonstrated relatively weak relations between moral reasoning and moral action ($r=.24$) and also between moral reasoning and immoral action ($r=-.19$) (Wu & Liu, 2014). It seems unlikely that a single variable could predict leaders’ behavioural propensity to navigate complexity successfully, particularly given that such a propensity would need to be enacted in VUCA conditions which may introduce a variety of contextual factors. Perhaps a cluster of psychological and contextual variables that include—but is
not limited to—hierarchical complexity may be jointly necessary and collectively sufficient to
account for leaders’ behavioural propensity to navigate complexity in workplace contexts.

In terms of research methodology, empirical research methods are predominantly applied
throughout this thesis. Some ego developmentalists advocate for supplementing empiricism with
more inclusive research methods that enquire into first, second, and third-person phenomena via
first, second, and third-person research practices in the past, present, and future (Chandler &
Torbert, 2003; Torbert, 2013). While these methods may illuminate facets of adult development
such as leaders’ subjective experience of working with different developmental models, it is unclear
whether they can contribute to addressing the types of research questions addressed in this thesis.
While ego developmentalists often regard empirical science as a manifestation of the Expert stage of
ego development (Cook-Greuter, 2013; Torbert, 2013), the empirical origins of ego development
theory and measurement is often emphasised in the literature (Cook-Greuter, 2013; Loevinger,
1976). As a result, the research methods employed in some of the ego development literature seem
consistent with those employed in the context of this thesis.

In terms of approaches to sampling, samples were drawn from Western nations, which may
compromise generalisability. Future research should test a more diverse range of cultures to
determine whether similar patterns are obtained.

Finally, the relationship between the author of this thesis and Lectica is worthy of comment.
The author of this thesis undertook an internship at Lectica with the goal of understanding Lectica’s
particular approach to cognitive development and measurement. Through this process, the author
of this thesis eventually became a Certified Senior Lectical Analyst™ with over eight years of training
in applications of the LAS and CLAS. At no point was the author of this thesis an employee of Lectica
and there was no commercial exchange for undertaking any of the analyses reported in this thesis.
To mitigate against potential bias, the author of this thesis had no influence over the samples
analysed for Studies 1A, 1B, 2A, 3, and 4 as they were scored by other suitably qualified parties, as
discussed in each study. The author of this thesis did have direct influence on the cognitive-
developmental scores but not on the ego development scores analysed for Study 2B, and did have indirect influence on the cognitive-developmental scores for Study 2C but not on the ego development scores. For Studies 2B and 2C, bias was mitigated by ensuring that samples were analysed in a manner that is consistent with the manner in which they would have been analysed in other contexts. This analytical process consisted of determining the amount of hierarchical complexity that is associated with the meaning of concepts that were identified in ego development samples. As a result of this process, the Lectical Phase Scores attributed to these concepts not only affected the LAS and CLAS scores awarded in the context of Studies 2B and 2C, but they would also affect the scores awarded to other samples of reasoning if the same concepts were employed in those samples of reasoning. The author of this thesis is not formally certified in scoring ego development but as discussed in Chapter 2, after completing various self-training exercises achieved an inter-rater agreement of greater than 90% in a sample of several hundred SCTs. It is possible that if the author of this thesis had been trained in the ego development scoring system rather than a cognitive-developmental scoring system, findings may have been interpreted more strongly through the lens of meaning-making and less strongly through the lens of hierarchical complexity. While the possibility of bias cannot be ruled out, appropriate steps were taken to mitigate the impact of bias on analyses and interpretations of findings.

**Considerations about Studies 1 and 2**

Study 1A analysed scores awarded by the LAS. Lectica recently implemented revisions to CLAS (Dawson, 2022a) and given the relation between Lectica’s two scoring systems there may be reasons to hypothesise that these revisions may also impact the LAS. As a result, it needs to be determined whether findings from this thesis can be used to justify scores awarded by Lectica’s current scoring system(s). Strictly speaking, reliability and validity are properties of samples of scores, but are not enduring properties of scoring systems or assessments (Andrich & Marais, 2019; Messick, 1989; Traub & Rowley, 1991). As a result, a similar suite of analyses may need to be undertaken to re-
establish the reliability and validity of Lectical Phase Scores and Lectical Decimal Scores. This may include consideration whether scores yielded by the revised scoring systems (e.g., Lectical Decimal Scores) are shown to exhibit interval, or near-interval, measurement properties (Dawson-Tunik, 2006; Stein & Heikkinen, 2009). This may be particularly important if Lectical Decimal Scores are used for high-stakes human capital decisions such as executive recruitment. Study 1A also presented three competing explanations for findings: determining whether structured wholeness is an artefact of scoring procedures, revisiting the rationale for using items with similar difficulties, and further validating whether the sequence of Lectical Phases reflects an increase in hierarchical complexity or an elaboration of concepts. Future research may be directed at ruling out these competing explanations to further validate the properties of Lectical Scores. Ruling out these competing explanations is non-trivial and may play a role in shifting the status of the findings about cognitive development from provisional to confirmatory.

Study 1B analysed the scores awarded by the ego development scoring system. Several opportunities were identified to improve the SCT, including item analysis, correcting for multidimensionality, addressing the inconsistent location of stages along the logit scale, and so on. Given that Rasch analyses test scores against the fundamental requirements of measurement, future research may consider implementing some of the recommended changes, gathering more data, and then conducting a similar suite of analyses to determine whether scores perform more favourably. It may be possible to identify if, and how, worldviews and complexity are related in a more systematic manner by administering a strengthened version of the SCT, other types of worldview measures (Lilienfeld et al., 2000; Obasi et al., 2009), and an assessment that evaluates the reasoning for holding a particular worldview, to a common sample of test-takers.

Studies 1 and 2 instantiate a research methodology that could be applied to a broader range of adult development assessments. A full methodology might elaborate on those employed in past research, including Rasch Analyses on developmental scores (Commons, Li, et al., 2014; Dawson et al., 2005), analysing relations between developmental scores (Dawson, 2001), scoring the scoring
criteria of other developmental scoring systems (Dawson & Gabrielian, 2003), and qualitative analyses of the hierarchical relations between successive complexity levels (Dawson-Tunik, 2004). Additional methods that have not been consistently used in the literature include evaluation of the broader epistemic paradigms in which particular adult development approaches emerged (Kuhn, 2012), scoring the definitional attributes of stages with a domain-general scoring system, qualitative analysis of the logic underpinning stages, confirming that the general requirements of ordinal scales are satisfied, and so on. Given that relatively few of these procedures have been undertaken on the measurement approaches associated with Kegan’s constructive-developmental theory (Kegan, 1982; Lahey et al., 2001), Gravesian life conditions (Beck & Cowan, 1996; Cowan & Todorovic, 2003, 2005; Laloux, 2014), and more contemporary versions of ego development theory (Murray, 2020) it is possible that some inferences about these theories and approaches to measurement may need to be revisited (Messick, 1989, 1995). Some work on the conceptual relations between Kegan’s constructive-developmental stages and stages of hierarchical complexity is underway, however this work tends to rely upon qualitative analyses of the definitional attributes of stages rather than quantitative empirical approaches (Hagström & Stålne, 2015; Heikkinen, 2011). In the recent past, some researchers have developed questionnaires and surveys which are contended to measure constructive-developmental stages (Brownlow, 2022) and other forms of vertical development (Parker, 2022). Perhaps the research methodologies described above may play a role in continuing to explore the psychometric properties of these new instruments by determining whether hierarchically integrated constructs are being measured.

**Considerations about Studies 3 and 4**

One of the most serious limitations of Study 3 was the inability to determine how large a complexity gap must be to impact leader effectiveness. Perhaps a complexity gap of one Lectical Phase may be acceptable if other forms of social scaffolding are available such as collaborative decision-making. However, a complexity gap of two or more Lectical Phases may be insurmountable
irrespective of the organisational resources provided. Future research may be directed towards evaluating the relations between the size of the complexity gap and leadership phenomena. Study 3 may also have been limited by analysing a particular sample of task demands that are constitutive of each management layer and scoring them with a particular cognitive-developmental scoring system. Future research may also be directed towards scoring a different sample of task demands for each management layer and/or scoring those task demands with another domain-general scoring system.

Study 4 presents limitations to interpreting the annual rate of growth identified. First, it was not possible to establish whether growth of under one-tenth of a complexity level has practical implications for leader effectiveness and organisational performance. Second, it was not possible to establish whether different growth rates to those identified may be more optimal for leaders over longer timescales. Third, findings revealed the rate at which hierarchical complexity happened to develop, but not necessarily the rate at which it might develop in the context of increasingly well-designed leader development initiatives. Future research may include additional variables so that longitudinal modelling can examine relations between these types of considerations.

As discussed in Chapter 5, design considerations limited the extent to which causal attributions could be made. Also, sample sizes limited the ability to identify potential differences in rates of growth between leadership providers. If these design considerations could be addressed by future research, this may have significant implications for our understanding of cognitive development over time, as it could be inferred that either pedagogical design, participant characteristics, or the interaction between them may impact growth trajectories.

A final limitation of Studies 4 and 5 was a function of CLAS. CLAS only awards a single score to each assessment to reflect the hierarchical complexity exhibited in leaders’ reasoning but does not award a unique score to each item. For this reason, only regression models on manifest variables could be applied, rather than full structural equation models that include measurement and structural models. This approach imposes some limitations on the ability to account for error
variance and should be addressed by future research as CLAS becomes capable of scoring individual items.

**Concluding Reflections**

Ego developmentalists may be correct that the “obdurate vagary of human beings” presents challenges to measuring and making clear inferences about human development (Loevinger, 1993a, p. 57). Notwithstanding the significance of these challenges, when aspects of leaders’ performance are considered in certain ways, we are provided with a unique window of insight into their capacity to navigate the complex conditions which currently beset our civilisation. For the field of leadership to make a serious contribution to bolstering the types of leadership that are required in these complex conditions, it must focus on approaches which are theoretically robust, empirically rigorous, and ethically defensible.
References


Commons, M., & Richards, F. (1984). A general model of stage theory. In M. L. Commons, F. A. Richards, & C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development* (pp. 120-140). Praeger Publishers.


Dawson, T. (2021). *Lectical levels: Structure and origins* (1 January, 2021). Copyright Lectica® Inc. [https://lecticalive.org/about/skill-levels#gsc.tab=0](https://lecticalive.org/about/skill-levels#gsc.tab=0)


IBM. (2010). *Capitalizing on complexity: Insights from the Global Chief Executive Officer study*.


Johnson, V. (2000). *The concept of differentiation as a contribution to the psychology of maturity* [Doctoral dissertation, Harvard University, Harvard Graduate School of Education].


Li, G., Commons, M., Duong, T., & Going, E. (2019). Predicting criminality with social perspective taking, attachment relationship, anger, impulsivity and depression.


research. *Frontiers in Human Neuroscience*, 9(58).
https://doi.org/https://doi.org/10.3389/fnhum.2015.00058


OpenAI. (2023). *ChatGPT (Mar 14 version) [Large language model]*. In https://chat.openai.com/chat


RUMM Laboratory. (2009b). *Multidimensionality and subtests in RUMM.* RUMM Laboratory Pty Ltd.


The jamovi project. (2021). *Jamovi.* In (Version 1.6) [Computer Software]. [https://www.jamovi.org](https://www.jamovi.org)


Van Rossum, Z. (2013). The development of social perspective taking and leadership decision-making in city government managers (Publication Number ERIC Number: ED559538) [Doctoral Dissertation, Columbia University]. New York, USA.


https://doi.org/10.1002/bdm.1751


https://doi.org/10.1023/A:1016052026435


https://www.rasch.org/rmt/rmt94n.htm


Appendix A: Hard Stage Claims Made by Ego Developmentalists

Hard Stage Claims Made by Loevinger

“Although not clearly recognizing ego development as an example, Piaget and Inhelder (Tanner & Inhelder, 1956, 1960) have abstracted the properties of the hierarchic model somewhat as follows: There is an invariable order to the stages of development; no stage can be skipped; each stage is more complex than the preceding one; each stage is based on the preceding one and prepares for the succeeding one. One can speak of stages of development, they assert, to the extent that these conditions are met. There is an inner logic that determines the sequence. In determining a person's current behavior, Piaget says, this inner logic of development is as important as his history or his heredity or his current environment” (Loevinger, 1966, p. 201).

“4. There is an inner logic (to borrow Piaget's term) of ego development, an invariable sequence. Each stage builds on, incorporates, and transmutes the previous one. This inner logic is not created by our definitions or by our reason; it constitutes the claim of this construct to preempt the term ego development, in contrast with other usages of the term in child psychology and psychoanalysis. No one, however, has succeeded in capturing fully this inner logic as yet. 5. The attempt to measure ego development must be based on a hierarchic model. Every person, protocol, or behavior sample is given just one rating, whatever diversity contained. The alternative embryonic model seems to imply as many ratings for every person or protocol as there are stages of development; evidence for fixation at one stage is not evidence against fixation at another. The embryonic model may have merit for psychosexual development but seems less good than the hierarchic model for ego development” (Loevinger, 1966, p. 204).

“Its development, consequently, is seen as a transformation of structures, a formulation that suggests both the importance of the cognitive element and the influence of Jean Piaget (Chapter Three)” (Loevinger, 1976, p. 4).
“In extending structuralism to other areas, in particular to personality and social development, the same conceptions and insights have been used. The most ambitious and the best articulated among such attempts has come to be known as cognitive developmentalism, an approach that is not a single coherent theory but includes researchers of varying backgrounds. They share two general tendencies: first, to adopt the most general developmental concepts advanced by Piaget, such as assimilation, accommodation, stage, and hierarchical and universal sequence; second, to hypothesize that the development of cognitive structures is the fundamental factor in psychological development as a whole, affecting in particular the acquisition of personality structures” (Loevinger, 1976, p. 40).

“These issues are central to theory of ego development. Ego development is presented as the "master trait" in personality, as the frame that provides more specific traits with their meaning and around which the whole edifice of personality is constructed. This theory shares with cognitive developmentalism the Piagetian notion of stage: ego stages are conceptualized as equilibrated structures, related to each other in an invariant hierarchical sequence” (Loevinger, 1976, p. 41).

“Werner’s statement of the orthogenetic law of development applies as well to ego as to cognitive development: "The development of biological forms is expressed in an increasing differentiation of parts and an increasing subordination, or hierarchization. Such a process of hierarchization means for any organic structure the organization of the differentiated parts for a closed totality, an ordering and grouping of parts in terms of the whole organism" (Werner, [1940] 1964, p. 41)” (Loevinger, 1976, p. 58).

“To the extent that ego development is a hierarchical variable, the difficulty we have of understanding, describing, and judging the highest stage is plausible. In principle, a person does not fully understand the relevant thinking of people much above his or her own level. Even if people of the highest ego level became psychologists, would they really want to spend their time scoring SCTs?” (Loevinger, 1993a, p. 59).
“The SCT is one of several stage-type theories, of which the best known is Kohlberg’s moral judgment interview (Colby & Kohlberg (1987). Kohlberg’s scale of developmental stages for moral judgment is recognizably about the same human species as the scale of ego development. The stages of the two conceptions can be set in approximate correspondence, although they concern somewhat different topics or aspects of personality (Blasi, 1998; Loevinger, 1976)” (Loevinger, 1998b, p. 82).

Loevinger asserted that there was a direct one-to-one correspondence between her stages, Piaget’s stages of moral reasoning (Loevinger, 1976, pp. 79-85) and Kohlberg’s stages of moral reasoning (Loevinger, 1976, pp. 118-122). Both models have been shown to satisfy hard stage requirements. This correspondence could only exist if ego development stages also exhibited, or were strongly related to, hard stage requirements.

**Hard Stage Claims Made by Torbert**

“In brief, developmental theory holds that persons and organizations evolve through a definite sequence of stages. Each stage change represents a fundamental transformation in which the assumptions that framed the previous stage are dethroned and become variables within the more inclusive framework surrounding the new stage. The assumptions of the earlier stage are not wrong; they are simply not absolutely and always the primary truth” (Torbert, 1987b, p 31)

“An important feature of developmental theory, consistently supported by research, is that there is a natural ordering or progression of these worldviews. As a person grows or matures, his or her worldview tends to go through a predictable sequence of changes (Kegan, 1982; Kohlberg, 1969; Piaget, 1967). In addition, structural developmental theory includes the following key propositions:

1. The order of development implies an invariant hierarchical sequence in which each more evolved worldview represents a more adequate understanding of the world than prior worldviews (Kohlberg, 1969).
2. Individuals holding more evolved worldviews tend to have developed greater cognitive abilities and conceptual complexity than those holding earlier worldviews (Harvey, Hunt & Schroeder, 1961; Loevinger, 1976).

3. As one matures developmentally, one becomes increasingly able to (a) accept responsibility for the consequences of one’s actions, (b) empathize with others who hold conflicting or dissimilar worldview, and (c) tolerate higher levels of stress and ambiguity (Bartunek, Gordon & Weathersby, 1983).

4. The person holding a more evolved worldview tends to be more attuned to his or her own inner feelings and outer environment than the person holding an earlier worldview (Loevinger, 1976)” (Fisher et al., 1987, p. 262).

“Developmental theory makes a series of strong claims about how development occurs. It claims that:

1. Development consists of a series of fundamental transformations. Each transformation reduces the "worldview" in which one was embedded to a part of a more inclusive worldview. In each transformation, what has been subject becomes object: the worldview that the person was (controlled by) becomes a capacity that the person has (control of).

2. This series of transformations occurs in a definite sequence.

3. This sequence is irreversible - in other words, a person does not regress once having reached a given stage.

4. A person’s development can cease at any stage - in other words, persons do not necessarily progress.

5. Persons at later stages of development can follow the logic of earlier stages, but persons at earlier stages tend to reinterpret later stage actions and logics into their own terms.

6. Development is caused, neither by the person nor by the environment, neither by "nature" or by "nurture," but by an interaction between the two” (Torbert, 1987c, p. 226-227).
“Constructive-developmental psychology identifies an ordered series of developmental stages, each of which is governed by a unique logic, set of assumptions, and overall framework. Developmental change from each stage to the next involves second-order transformation. The initial limiting framework is dethroned and becomes a strategic option, or variable, within a more inclusive assumptive framework (Kegan, 1982; Kohlberg, 1969; Loevinger, 1976; Torbert, 1987a; Trevino, 1986; Wilber, 1980). At the late stages of development, the person becomes increasingly aware that there are alternative frames, that perceptions (including one's own) are always framed by assumptions, and that such assumptions can be tested and transformed” (Torbert, 1989, p. 86).

“In seems a very important step in helping students to know one another and to write and share a paper on their own development that they have previously written the paper on their work-colleagues’ development and that the descriptive/prescriptive value of the developmental theory has been repeatedly highlighted as opposed to the evaluative/hierarchical quality (I use the more descriptive words ‘earlier’ and ‘later’ stages of development rather than ‘lower’ and ‘higher’ in this spirit)” (Torbert, 1991b, p. 6-7).

“Consequently, just as the practice of transforming power through action inquiry is unfamiliar to most managers, so too the concept of transforming power and the entire flavor of qualitative, developmental transformations to later perspectives that include one's entire previous perspective as a variable is by and large unfamiliar to the American management research community” (Rooke & Torbert, 1998, p. 14).

“Developmental Action Inquiry (DAI) is an approach to social science and to personal and social life that posits a sequential series of increasingly complex, inclusive, and mutually transforming action-logics through which individual persons, conversations, relationships, organizations, and scientific paradigms may evolve as they intertwinedly act and inquire” (Torbert & Livne-Tarandach, 2009, p. 133).
“The statistical technique of cluster analysis is designed to assess the extent to which a measurement tool is able to capture qualitative differences that are conceptualized in the underlying theory that the tool is trying to measure” (Torbert & Livne-Tarandach, 2009, p. 145).

“In spite of their deep yin-yang complementarity, the field tends to glorify the formal hierarchical ordering of stages, and to avoid the chaotic, fluid messiness of how and why people develop” (Palus et al., 2020, p. 62).

Torbert asserted that there was a direct one-to-one correspondence between his stages and Kohlberg’s stage of moral reasoning (Fisher et al., 1987, p. 260; Lichtenstein et al., 1995, p. 98) and also between Commons’ Model of Hierarchical Complexity (Torbert, 1994). Both models have been shown to satisfy hard stage requirements. Torbert also asserted that there is a direct one-to-one correspondence between ego development and Kegan’s stages of the evolving self (Fisher et al., 1987, p. 260; Lichtenstein et al., 1995, p. 98; Torbert, 1994) and between Selman’s stages of socio-moral perspective taking (Fisher et al., 1987, p. 260). Both models are likely to satisfy hard stage requirements. This correspondence could only exist if ego development stages also exhibited, or were strongly related to, hard stage requirements.

**Hard Stage Claims Made by Cook-Greuter**

“The evolving perspective on the self, as measured by the SCT, can be shown to satisfy the demands for a hard-stage theory. It shows a qualitative difference in structure that serves the same basic function of meaning-making throughout the stages. The stages form an invariant sequence, they represent structured wholes, and they constitute hierarchical integrations” (Cook-Greuter, 1990, p. 86).

“In summary, ego development theory postulates an invariant, hierarchical sequence of distinct view of reality and subject-object integrations which comprise operative, cognitive and emotional aspects of living” (Cook-Greuter, 1994, p. 121).

“In contrast to Commons’ and Richards’ model which is heavily theory-driven, Loevinger’s ego development theory (1976) and especially her measuring instrument *The Washington
University Sentence Completion Test (SCT; Loevinger, Wessler and Redmore, 1978) are based on empirical evidence. Ego development theory proposes an invariant hierarchical sequence of cohesive meaning perspectives with which humans face life. It is a psychological system which interrelates action, affect and cognition and therefore unlike Commons’ and Richards GSM, does attempt a more full-bodied explanation of human meaning-making” (Cook-Greuter, 1995, p. 21).

“This thesis investigates mature ego development in response to perceived high-end problems with Loevinger’s Ego Development Theory (1976) and measuring instrument, The Washington University Sentence Completion Test (SCI) (1970; Hy & Loevinger, 1996). In addition to a continuing differentiation-integration at higher levels of hierarchical complexity, postconventional development is here conceived of as a conscious, stepwise deconstruction of the habits of mind, or an increasing awareness of the constructed nature of reality. Two new stages, a Construct-aware and a Unitive stage are described which replace Loevinger’s Integrated stage. Nine unique, empirically derived characteristics for identifying these two postautonomous ego stages are also introduced” (Cook-Greuter, 1999, p. vii).

“Like other developmental stage theories, ego development theory posits that the stages form an invariant sequence and represent hierarchical integrations of greater and greater complexity, cognitive differentiation, and integration (Piaget, 1952; Kegan, 1982; 1994; Kohlberg, 1984; Basseches, 1984a). A higher stage does not merely contain a more complex matrix of the content and structure of the prior stage, but transforms the previous way of looking at reality and offers a broader or deeper, more integrated perspective” (Cook-Greuter, 1999, p. 38).

“In response to his critique, I suggest that the evolving perspective on the self demonstrates the inner logic of ego development from the Symbiotic stage through the proposed postautonomous stages. I argue that spelling out the structure of an underlying regularly expanding self-perspective moves ego development theory from a "soft" stage theory to one more akin to a "hard" stage type according to Kohlberg’s distinction” (Cook-Greuter, 1999, p. 52).
“Each of the two new stages hierarchically integrates the previous way of meaning making. The Construct-aware stage is the highest observed self-view in the rational, symbolically mediated realm of cognition. The Unitive stage, with its fluid, deconstructed self-view, seems to be at the threshold to the realm of postsymbolic, metaphysical or transpersonal consciousness.” (Cook-Greuter, 1999, p. 68).

“First, by introducing the construct of the ‘perspective on the self’, I have sought to demonstrate that ego development theory can be seen as akin to a hard structural theory. Unlike the ego as presented in Loevinger’s theory, ‘the perspective on the self’ can be logically mapped and figuratively represented throughout the scale. Thus, ego development more clearly fulfills the structural requirements for being a stage theory: (a) The stages follow each other in a hierarchical sequence, (b) the same construct is used to show its transformation throughout the sequence” (Cook-Greuter, 1999, p. 69).

“Thus, although the Unitive stage is more uniquely defined than Loevinger’s Integrated stage, and can be considered a hierarchical integration according to the tenets of constructivist developmental theory, it is best considered as a threshold position between the personal and the transpersonal tier” (Cook-Greuter, 1999, p. 70).

“I have argued in the beginning of this chapter that postconventional development can be looked at from two seemingly contradictory perspectives. First, postconventional development continues the pattern of alternating stages of differentiation and integration toward greater and greater hierarchical complexity accommodating an ever-expanding experiential universe. At the same time, postconventional development can be understood as a stepwise deconstruction of the previously unconsciously constructed permanent object world with its closed boundaries, linear causality and subject/object separation. I suggested that both views provide useful lenses into aspects of postconventional personality development, but that only the latter is adequate to describe the insights of people at the most advanced ego stages as seen on rare SCTs.” (Cook-Greuter, 1999, pp. 126-127).
“As can be seen in Figure 4.82 below, the overall distributions are consistent with predictions as well as the gradual shift of the center of the distributions toward the right with increasing TPR. This reflects the shifting center of meaning making at successively higher ego stages. The graph shows that the distribution curves for the different personality types (ego stages) are clearly discontinuous and distinct from each other as each type of TPR peaks in a different area of the distribution space. The prediction that the distribution patterns for different ego stages are distinct is supported by the evidence. For a stage theory to be hierarchical, the stages have to be distinct and sequential. Stages 5(E8), 5/6 (C9) and 6 (C10) follow each other in a sequential order and are discontinuous according to the percentage frequency distributions for the different groups of TPRs. Therefore, the Construct aware and the Unitive stage seem to fulfil the stage requirements for being higher-order hierarchical integrations” (Cook-Greuter, 1999, p. 160).

Various claims that hierarchical complexity is a necessary but non-sufficient condition to account for the underpinning logic of ego development stages (Cook-Greuter, 1999, pp. 90-93, 175).

“According to constructivist, bottom-up developmental stage theory (Piaget, 1938, 1954; Kohlberg & Armon, 1984; Basseches, 1984; Commons & Richards, 1984) mental growth follows a predictable trajectory. People move through an invariant, hierarchical sequence of increasingly complex and coherent stages of reasoning from birth to adulthood. Each higher stage follows the next lower in a predictable order. Each higher stage not only follows, but also transcends and integrates the content of the lower into a more complex mental model of reality. In other words, each higher stage can understand and coordinate the stage below it while the reverse is not true. Instead, the lower tends to reduce whatever it gleans from above to its own level (Fisher, Hand & Russel, 1984)” (Cook-Greuter, 2000, pp. 2-3).

“Central to the constructivist developmental model are the claims that (a) the stage sequence is unidirectional and that the stages constitute hierarchical integrations; (b) people evolve from the least differentiated to ever more differentiated ways of knowing and relating to the world and their inner experience; and (c) development moves from simple to complex in regard to all
possible contents, lines and domains, and their interconnections. In the most global sense, development can be described as the gradual unfolding of people’s capacity of embrace ever-vaster horizons and to plumb ever-greater depths of heart, mind, body, and soul” (Cook-Greuter, 2005, p. 26).

Cook-Greuter asserted that there was a direct correspondence between her stages and the Model of Hierarchical Complexity (Cook-Greuter, 1990, p. 104; 2002, pp. 7-28; 2013, pp. 22-76). The Model of Hierarchical Complexity has been shown to satisfy hard stage requirements (Commons et al., 2008; Dawson et al., 2010). This correspondence could only exist if ego development stages also exhibited, or were strongly related to, hard stage requirements.

**Hard Stage Claims Made by Other Ego Development Scholars**

"Loevinger’s model of ego development formulates a series of sequentially ordered stages (seven stages and three transitional phases). Although they may be correlated with chronological age, they are defined independently of age. As in Piaget's system, these stages comprise an invariant hierarchical order. Each stage is more complex than the last; none can be skipped in the course of development. However, different individuals may not develop beyond various stages. Among adults there are representatives of each stage, who can then be characterized in terms of the features specific to the stage at which they have stopped. Consequently, in addition to being an invariant sequence of stages, Loevinger’s system generates a typology of individual differences in “character styles” (Hauser, 1976, p. 930).

"Loevinger's model of ego development (Loevinger, 1966), which the SCT was developed to assess, draws from the work of Piaget (1932) in postulating a series of developmental stages that are assumed to form a hierarchical continuum and to occur in an invariant sequence” (Browning, 1987, p. 113).

"Development proceeds jerkily, the time occupied by each stage varying, so that periods of rapid change alternate with plateaus of indefinite length during which slow or negligible change takes place” (Holt, 1998, p. 78).
“In summary, the research conducted in the 20 years since the last review (Loevinger, 1979) has provided further support for the conceptual soundness of ego development theory. The findings indicate that ego development may be regarded as a complex, but unitary, construct, with the ego developing in a hierarchical, invariant, and sequential manner” (Manners & Durkin, 2001, p. 562).
Appendix B: The Lectical Scale

Table B 1

*The Lectical Scale*
<table>
<thead>
<tr>
<th>Lectical Level</th>
<th>Name of Lectical Level</th>
<th>Hierarchical order of abstraction</th>
<th>Logical structure of arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Primary order of abstraction</td>
<td>Secondary order of abstraction</td>
</tr>
<tr>
<td>0</td>
<td>Single reflexes</td>
<td>Reactive actions</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;-order reflexes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Typically accounts for development pre-birth to 3 months. At this order, infants process and innately respond to basic sensory stimuli such as sights and sounds.</td>
<td>0b – unelaborated</td>
</tr>
<tr>
<td>1</td>
<td>Reflexive mappings</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;-order reflexes which coordinate 1&lt;sup&gt;st&lt;/sup&gt;-order reflexes</td>
<td>1a – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1d – highly elaborated</td>
</tr>
<tr>
<td>2</td>
<td>Reflexive systems</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;-order reflexes which coordinate 2&lt;sup&gt;nd&lt;/sup&gt;-order reflexes</td>
<td>2a – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2d – highly elaborated</td>
</tr>
<tr>
<td>3</td>
<td>Single sensorimotor schemes</td>
<td>Sensorimotor schemes</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;-order sensorimotor schemes which coordinate 3&lt;sup&gt;rd&lt;/sup&gt;-order reflexes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Typically accounts for development from approximately 3 to 18 months. At this order, infants learn to act intentionally on the environment to achieve certain outcomes such as shaking an object to hear a sound.</td>
<td>3b – unelaborated</td>
</tr>
<tr>
<td>4</td>
<td>Sensorimotor mappings</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;-order sensorimotor schemes which coordinate 1&lt;sup&gt;st&lt;/sup&gt;-order sensorimotor schemes</td>
<td>4a – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4d – highly elaborated</td>
</tr>
<tr>
<td>5</td>
<td>Sensorimotor systems</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;-order sensorimotor schemes which coordinate 2&lt;sup&gt;nd&lt;/sup&gt;-order sensorimotor schemes</td>
<td>5a – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5d – highly elaborated</td>
</tr>
<tr>
<td>6</td>
<td>Single representations</td>
<td>Representations</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;-order representations which coordinate 3&lt;sup&gt;rd&lt;/sup&gt;-order sensorimotor schemes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Typically accounts for development from approximately 18 months to 8 years. At this order, children use concrete representations of the world which grow in complexity from gestures to words to sentences to entire stories.</td>
<td>6b – unelaborated</td>
</tr>
<tr>
<td>7</td>
<td>Representational mappings</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;-order representations which coordinate 1&lt;sup&gt;st&lt;/sup&gt;-order representations</td>
<td>7a – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7d – highly elaborated</td>
</tr>
<tr>
<td>8</td>
<td>Representational systems</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;-order representations which coordinate 2&lt;sup&gt;nd&lt;/sup&gt;-order representations</td>
<td>8a – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>Name of Lectical Level</td>
<td>Hierarchical order of abstraction</td>
<td>Logical structure of arguments</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------</td>
<td>-----------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Primary order of abstraction</td>
<td>Secondary order of abstraction</td>
</tr>
<tr>
<td>9</td>
<td>Single abstractions</td>
<td>Abstractions</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;-order abstractions which coordinate 3&lt;sup&gt;rd&lt;/sup&gt;-order representations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9&lt;sup&gt;b&lt;/sup&gt; – unelaborated</td>
</tr>
<tr>
<td>10</td>
<td>Abstract mappings</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;-order abstractions which coordinate 1&lt;sup&gt;st&lt;/sup&gt;-order abstractions</td>
<td>10&lt;sup&gt;a&lt;/sup&gt; – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10&lt;sup&gt;d&lt;/sup&gt; – highly elaborated</td>
</tr>
<tr>
<td>11</td>
<td>Abstract systems</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;-order abstractions which coordinate 2&lt;sup&gt;nd&lt;/sup&gt;-order abstractions</td>
<td>11&lt;sup&gt;a&lt;/sup&gt; – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11&lt;sup&gt;d&lt;/sup&gt; – highly elaborated</td>
</tr>
<tr>
<td>12</td>
<td>Single principles</td>
<td>Principles</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;-order principles which coordinate 3&lt;sup&gt;rd&lt;/sup&gt;-order abstractions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12&lt;sup&gt;b&lt;/sup&gt; – unelaborated</td>
</tr>
<tr>
<td>13 *</td>
<td>Principled mappings</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;-order principles which coordinate 1&lt;sup&gt;st&lt;/sup&gt;-order principles</td>
<td>13&lt;sup&gt;a&lt;/sup&gt; – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13&lt;sup&gt;d&lt;/sup&gt; – highly elaborated</td>
</tr>
<tr>
<td>14 *</td>
<td>Principled systems</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;-order principles which coordinate 2&lt;sup&gt;nd&lt;/sup&gt;-order principles</td>
<td>14&lt;sup&gt;a&lt;/sup&gt; – transitional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14&lt;sup&gt;d&lt;/sup&gt; – highly elaborated</td>
</tr>
</tbody>
</table>

* Note. Levels 13 and 14 are currently hypothetical possibilities. Adapted from Dawson-Tunik (2006).
Appendix C: Lectica’s Privacy and Legal Statements

“Privacy policy

Lectica, Inc., Northampton, MA 01060

(Revised June 25, 2020)

Please read the following Privacy Policy carefully. This Policy contains general information about the policy of this website in its collection, protection, use and disclosure of the personal information gathered on this site. If you are under 18 years of age, please be sure to read this privacy statement with a parent or guardian and ask questions about anything you may not understand.

This website, lecticalive.org, (the "Site"), is owned and operated by Lectica, Inc, ("Lectica"). Lectica recognizes the importance of explaining our privacy policies to users of this Site and protecting the privacy of Personal Information we obtain through the Site. Please note that this Privacy Policy may be revised by Lectica at any time, particularly to address changes in our business, the law, or technology. If revisions are made, we will post the revised Policy here. Using this Site following the posting of changes to this Policy shall constitute your acceptance of the revised Policy.

SUMMARY

This summary is not intended as a replacement for the entire policy.

- We never share Personal Identifying information with third parties without your consent. You can withdraw your consent simply by contacting us with a request.
- You can delete your account at any time simply by contacting us with a request.
- We never knowingly allow a Child (age 12 and under) access to assessments & reports without the authorization of a teacher, parent, or guardian.
- We use demographic information only to (1) populate reports, (2) conduct research on learning, and (3) monitor and eliminate bias.
- We use software that helps us detect cheating.

SPECIAL NOTICE TO PARENTS, TEACHERS, AND CHILDREN
PARENTS AND TEACHERS: We encourage parents, guardians, and teachers to spend time with their Children (Children 12 and under) when their Children engage in the online activities on this Site. We never knowingly allow a Child direct access to our assessments without the authorization of a teacher, parent, or guardian. We never request Personal Identifying Information directly from a Child. We encourage parents to help us protect their Children's privacy by instructing them never to provide Personal Information of any kind on this Site or any other website without parental permission.

CHILDREN: We do not knowingly permit children to register on our site, and have taken measures to make this very difficult.

Children may from time to time voluntarily submit Personal information through a "Contact Us" link on the Site or a public posting area on the Site, if such feature is available. If a Child e-mails us via a "Contact Us" link on any child-directed or public area of this Site and we can identify the sender as a Child, any Personal Information submitted will only be used to respond on a one-time basis to such Child. In addition, if we have actual knowledge that a Child is posting Personal Information on any area of this Site, we will use commercially reasonable efforts to delete such Personal Information before it is made public. Please note that any online store that may be accessed from the Site is intended for parents, teachers and other adult purchasers over the age of 18 and requires the submission of Personal Information to fulfill and service any transactions.

We may disclose a Child's Personal Information if we have reason to believe that disclosing this information is necessary to identify, contact or bring legal action against someone who may be causing injury to or interference with (either intentionally or unintentionally) our rights or property, other users, or anyone else that could be harmed by such activities. We may also disclose a Child's Personal Information when we believe in good faith that the law requires it.

LINKS TO OTHER SITES

This Site may include external links to other sites of possible interest to users of this Site. In addition to lecticalive, Lectica operates World Wide Web sites for the Developmental Testing Service (DTS), and our Moodle site. You may access these sites directly or through lecticalive.org.
Links from this site to external websites are provided solely for your convenience. You understand that should you leave this site via a link, the content of any third-party site is not provided or endorsed by Lectica, and Lectica has not reviewed or approved the terms of use or other policies governing such sites. Lectica does not monitor, make any representation with respect to or assume any liability for any third-party sites, including, without limitation, any products or services that are advertised or made available for purchase through such sites.

**PERSONAL INFORMATION WE COLLECT AND USE**

We collect and store personal information that can identify you as a specific individual (“Personal Information”), which you provide voluntarily when using our Site. Personal Information may include, but is not limited to: your full name, email address, home address, phone number, date of birth, credit card information, and certain other information if it is linked with your Personal Information. We strongly recommend that you not use your Social Security Number or any other important personal information as a user ID or password for this Site or others.

Lectica does not collect Personal Information about a user unless it is voluntarily provided to our Company. We use personal information to serve the user's needs and our legitimate business purposes as stated here. For example, when a user sends us a message or makes an online purchase, we need to retain certain Personal Information on this Site in order to respond. We may also use Personal Information to help us make improvements to this Site. We never use the Personal Information you provide for marketing or promotional purposes. The information will be treated confidentially within our company, as we protect other confidential information. You can choose not to provide certain information, but then you might not be able to take advantage of certain Site features.

**PUBLIC POSTINGS**

No unauthorized individual is permitted access to users’ assessment results, including essays, scores, comments, etc. However, any Personal Information that is voluntarily posted by a user to a PUBLIC area of a Lectica Web site (for example, a posting to a bulletin board or chat room) might be collected and used by others. We cannot prevent such uses. We advise users to exercise caution in public spaces on our sites.
INFORMATION SHARING AND DISCLOSURE

Except as specifically stated in this statement, our policy is not to share the Personal Information collected through this Site with third parties. We may disclose Personal Information (a) in response to a subpoena, court order or legal process, to the extent permitted and required by law; (b) to protect user security or the security of other persons, consistent with applicable laws; (c) in connection with a sale, joint venture or other transfer of some or all of Lectica the assets of Lectica; or (d) in order to enforce the Site’s Terms of Use. We exercise commercially reasonable care otherwise not to share, disclose, or sell Personal Information of users to third-parties, except with the prior approval of the user.

Lectica may use a third-party service to collect anonymous visitor information like IP addresses, browser types (such as Internet Explorer or Netscape), referring pages, pages visited, and time spent on a particular site. Lectica collects this information (1) for statistical analysis of web page traffic patterns; (2) to administer our Site and servers; (3) to allow for auditing of our services by some third parties who have that right; (4) for internal purposes to make marketing decisions; and to detect cheating on our assessments.

For pages that provide materials and services to instructors and students for purposes of specific courses or testing, the Site may display the names, email addresses, scores, and log-in names of instructors and students and make them available to school administrators. Administrators may store personal data generated through the assessment interface for the purpose of maintaining educational records or other legitimate educational purposes.

Lectica may use assessment information to compile aggregate data for research purposes. We may share data from assessments with third parties who are engaged in research that we believe in good faith is legitimate and who agree to use the information for research purposes only. This information will be studied and/or shared only in aggregate, and never in connection with your personal identifying information.

COLLECTION AND USE OF INFORMATION FROM COOKIES

Our Site uses "cookies" to obtain certain types of information when your Web browser accesses the Site. Cookies are textual identifiers that our system transfers to your computer's hard drive through
your Web browser to enable our Site to recognize your browser, and to optimize and sometimes customize your use of our Site. If you want to disable cookies, you can. The Help portion of the toolbar on most browsers and "add-on" programs will tell you how to prevent your browser from accepting new cookies, how to have the browser notify you when you receive a new cookie, or how to disable cookies altogether. It is your choice, but cookies allow you to take full advantage of this Site's features, and we recommend that you leave them turned on.

We do not use cookies to get users' Personal Information before a user registers with our Site and affirmatively opts-in. Once a user opts-in and provides personal information for purposes of receiving more information from us, we may collect and use that information obtained through the cookie technology for purposes of marketing to you, in accordance with this policy. If you do not want to receive e-mail or other marketing communications from us, you can opt out of further marketing to the addresses you have provided.

Lectica reserves the right to monitor your use of the site to determine if you are complying with our terms and the right, in its sole discretion, to terminate your access to the Site.

**OBTAINING ADDITIONAL INFORMATION ABOUT PRIVACY**

If you have any concerns or questions about privacy on this Site, please [Contact Us].

(Lectica, 2020).
Appendix D: Additional Rasch Analyses Conducted on the LDMA

Step 4: Test for Other General Measurement Properties

The purpose of this step was to undertake additional analyses on general measurement properties of the LDMA.

Differential Item Functioning (DIF)

Differential item functioning (DIF) occurs when two or more groups of persons with identical ability have non-identical expected responses. Uniform DIF occurs when one group is consistently awarded higher scores across all ability levels. Non-uniform DIF occurs when there is an interaction between group membership and ability (Andrich & Hagquist, 2012). Six demographic variables were analysed for DIF: sex, ethnicity, first language, age at test time, educational attainment, and management layer. Statistical analyses were conducted via two-way ANOVAs (refer Table D1), and visual analyses were conducted via Item Characteristic Curves (ICCs). In the context of two-way ANOVAs, a significant main effect for class interval indicates a general lack of fit to the ICC irrespective of group classification, a significant main effect for group indicates uniform DIF, and a significant class interval x group interaction indicates non-uniform DIF. A Bonferroni correction yielded a critical alpha of .003. Analyses of ICCs were used as a supplementary method of identifying DIF patterns which may not have reached statistical significance. In all cases, random sampling was used to obtain groups of comparable size to avoid bias. RUMM performed calculations based on the default of 10 class intervals.
Table D 1

Differential Item Functioning Analyses Conducted on the LDMA

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Category</th>
<th>n</th>
<th>Main effect of class intervals</th>
<th>Main effect of groups</th>
<th>Interaction effect</th>
<th>Uniform DIF</th>
<th>Non-uniform DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex</td>
<td>Male</td>
<td>260</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Female</td>
<td>260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Ethnicity</td>
<td>Caucasian</td>
<td>149</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Caucasian</td>
<td>149</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Language</td>
<td>Native English speakers</td>
<td>221</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Native English speakers</td>
<td>221</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Age at test time</td>
<td>18 to 27 years</td>
<td>50</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28 to 37 years</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>38 to 47 years</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>48 to 57 years</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>58 to 68 years</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Educational attainment</td>
<td>High School</td>
<td>50</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Undergraduate</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Masters</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Management layer</td>
<td>Layer 2</td>
<td>90</td>
<td>Item 4, p=0.002</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Layer 3</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Layer 4</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Layer 5</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Neither statistical nor visual analysis identified any instances of DIF, except for a class interval effect for item 4 on management layer. As an illustrative example, item 1 for the DIF analysis on language was selected, as its ICC was representative of the pattern observed in all other ICCs (refer Figure D1). It was not deemed necessary to resolve DIF or take further action.
Response Dependence

If responses are independent, each item makes a related, but unique, contribution to the total score. Local dependence Type 2, or response dependence, is an inter-item effect which occurs when, for the same person and therefore for the same value of $\beta$, the location of one item $j$ is governed by its relation to another item $i$ (Marais & Andrich, 2008a, 2008b). In this way, the responses are only not a function of the parameters of the model, but also a function of the response to the first item. The algebraic representation of response dependence for polytomous items is beyond the current scope. It is sufficient to say that $d$ characterises the magnitude of response dependence resulting from a change in location to item $j$ caused by its dependence on item $i$. Response dependence occurs when pairs of items are more related to one another than they are to other items due to shared variance that cannot be attributed to the Rasch factor. Such dependence is revealed when correlations between standardised item residuals exceed $+0.3$ (Andrich & Kreiner, 2010). As it can be seen from Table D2, highest correlations between residuals did not reach the threshold (min= -0.34, max= -0.12). As a result, evidence in support of local dependence Type 2, or response dependence, was not found.
In summary, while some items do appear to be more related to others in terms of content (i.e., items 1 and 2, and items 3, 4 and 5) these analyses do not provide support for significant violations to local independence Types 1 and 2.

**Measurement Precision**

The measurement precision of thresholds, \( \tau_n \), is critical in the context of a development assessment because it determines how accurately thresholds, and therefore Lectical Phases, can be located on the logit scale. Standard Errors (SE) for thresholds appeared to be low (\( M=0.28, \ SD=0.25, \min=0.11, \max=1.13 \)). The range of a 95% confidence interval for thresholds is 1.10 logits. As a result, thresholds can be confidently located within a 1.10 logit range which appears to be relatively precise in the context of a 31.87 logit range for thresholds (i.e., \( 2 \times 1.96 \times 0.28=1.10 \)).

The measurement precision of persons is also critical in the context of a development assessment because it determines how confidently leaders can be placed into Lectical Phases. For persons, SEs appeared to be somewhat higher than for thresholds (\( \text{mean}=1.04, \ SD=0.13, \min=0.88, \max=1.37 \)). The range of a 95% confidence interval for person ability is 4.08 logits (i.e., \( 2 \times 1.96 \times 1.04=4.08 \)). In the context of a 33.14 logit range for persons, estimates appeared to be relatively precise since approximately eight confidence intervals can fit within this range (\( 33.14/4.08 \approx 8.12 \)). This is consistent with the claim that leaders can be placed into eight statistically distinct Lectical

---

**Table D 2**

*Residual Correlations Between LDMA Items*

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-0.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-0.29</td>
<td>-0.26</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-0.27</td>
<td>-0.34</td>
<td>-0.12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-0.34</td>
<td>-0.29</td>
<td>-0.26</td>
<td>-0.15</td>
<td>-</td>
</tr>
</tbody>
</table>
Phases. In summary, evidence supports a sound level of measurement precision for both thresholds and persons.
# Appendix E: Studies on the Reliability of the SCT

Table E1

Coefficient Alpha for the SCT with 18 Stems (Listed from Lowest to Highest Alpha)

<table>
<thead>
<tr>
<th>Study</th>
<th>Stems</th>
<th>Alpha</th>
<th>Comments</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>.70</td>
<td>-</td>
<td>Newman (2005)</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>.76</td>
<td>Second split-half of 18 stems</td>
<td>Novy et al. (1997)</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>.77</td>
<td>-</td>
<td>Camberis et al. (2014)</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>.81</td>
<td>Second split-half of 18 stems</td>
<td>Novy and Francis (1992)</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>.82</td>
<td>First split- half of 18 items</td>
<td>Novy et al. (1997)</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>.84</td>
<td>-</td>
<td>Morros et al. (1998)</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>.84</td>
<td>First split- half of 18 items</td>
<td>Novy and Francis (1992)</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>.91</td>
<td></td>
<td>Daniels et al. (2018)</td>
</tr>
</tbody>
</table>

Note. Median alpha from Table E1 is 0.82
### Table E2

*Coefficient Alpha for SCT with 36 Stems (Listed from Lowest to Highest Alpha)*

<table>
<thead>
<tr>
<th>Study</th>
<th>Stems</th>
<th>Alpha</th>
<th>Sample</th>
<th>Comments</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36</td>
<td>.76</td>
<td></td>
<td></td>
<td>von der Lippe (1986)</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>.78</td>
<td></td>
<td></td>
<td>Pierre et al. (1999)</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>.80</td>
<td>Females only</td>
<td>Study 1 – seminal study on reliability of the SCT</td>
<td>Redmore and Waldman (1975)</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>.83</td>
<td>Adolescent males</td>
<td></td>
<td>Lindfors et al. (2007)</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
<td>.85</td>
<td></td>
<td></td>
<td>Frank and Quinlan (1976)</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
<td>.86</td>
<td>Males only</td>
<td>Study 1 – seminal study on the reliability of the SCT</td>
<td>Redmore and Waldman (1975)</td>
</tr>
<tr>
<td>8</td>
<td>36</td>
<td>.86</td>
<td>.86 was the higher of two estimates of alpha provided</td>
<td>Kurtz and Tiegreen (2005)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>36</td>
<td>.87</td>
<td>Males only</td>
<td>Study 2 – seminal study on the reliability of the SCT</td>
<td>Redmore and Waldman (1975)</td>
</tr>
<tr>
<td>10</td>
<td>36</td>
<td>.87</td>
<td></td>
<td></td>
<td>Novy et al. (1997)</td>
</tr>
<tr>
<td>11</td>
<td>36</td>
<td>.88</td>
<td>For the first test, not for retest</td>
<td></td>
<td>Weiss et al. (1989)</td>
</tr>
<tr>
<td>12</td>
<td>36</td>
<td>.88</td>
<td></td>
<td></td>
<td>Newman et al. (1998)</td>
</tr>
<tr>
<td>13</td>
<td>36</td>
<td>.89</td>
<td>Females only</td>
<td>Study 2 – seminal study on the reliability of the SCT</td>
<td>Redmore and Waldman (1975)</td>
</tr>
<tr>
<td>14</td>
<td>36</td>
<td>.89</td>
<td>Young adults</td>
<td></td>
<td>Browning (1986)</td>
</tr>
<tr>
<td>15</td>
<td>36</td>
<td>.90</td>
<td>University students</td>
<td></td>
<td>Holt (1980)</td>
</tr>
<tr>
<td>16</td>
<td>36</td>
<td>.90</td>
<td></td>
<td></td>
<td>Novy and Francis (1992)</td>
</tr>
<tr>
<td>17</td>
<td>36</td>
<td>.90</td>
<td></td>
<td></td>
<td>Novy (1993)</td>
</tr>
<tr>
<td>18</td>
<td>36</td>
<td>.90</td>
<td></td>
<td></td>
<td>Novy et al. (1994)</td>
</tr>
<tr>
<td>Study</td>
<td>Stems</td>
<td>Alpha</td>
<td>Sample</td>
<td>Comments</td>
<td>Source</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>----------</td>
<td>--------</td>
</tr>
<tr>
<td>19</td>
<td>36</td>
<td>.91</td>
<td>Women and adolescent girls</td>
<td></td>
<td>Loevinger and Wessler (1970)</td>
</tr>
<tr>
<td>20</td>
<td>36</td>
<td>.91</td>
<td></td>
<td></td>
<td>Torbert and Livne-Tarandach (2009); Torbert et al. (2010)</td>
</tr>
</tbody>
</table>

*Note. Median alpha for Table E2 is 0.88*
Table E 3

Coefficient Alpha SCT with Various Numbers of Stems (Organised by SCT Length and Lowest to Highest Alpha)

<table>
<thead>
<tr>
<th>Study</th>
<th>Stems</th>
<th>Alpha</th>
<th>Sample</th>
<th>Comments</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>.70</td>
<td>Early adolescents</td>
<td></td>
<td>Marsh et al. (2006)</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>.74</td>
<td>Females only</td>
<td></td>
<td>Dubow et al. (1987)</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>.76</td>
<td>Males only</td>
<td></td>
<td>Holt (1980)</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>.77</td>
<td>Females only</td>
<td></td>
<td>Holt (1980)</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>.78</td>
<td>Males only</td>
<td></td>
<td>Dubow et al. (1987)</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>.58</td>
<td>Adolescents</td>
<td></td>
<td>Loeb et al. (2021)</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>.81</td>
<td>Early adolescents</td>
<td>For stems 1 through 16</td>
<td>Westenberg et al. (2001)</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>.81</td>
<td>Early adolescents</td>
<td>For stems 17 through 32</td>
<td>Westenberg et al. (2001)</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>.80</td>
<td>Adolescents</td>
<td></td>
<td>Drewes and Westenberg (2001)</td>
</tr>
<tr>
<td>12</td>
<td>22</td>
<td>.87</td>
<td>Adolescent girls</td>
<td></td>
<td>von der Lippe (2000)</td>
</tr>
<tr>
<td>13</td>
<td>32</td>
<td>.95</td>
<td>Youth</td>
<td></td>
<td>Westenberg et al. (2004)</td>
</tr>
</tbody>
</table>

Note. Median alpha from Table E3 is 0.77.
Appendix F: Relation Between SCT Scores and Scores Based on Various Cognitive-Developmental Models

Table F 1

Correlations between SCT Scores and Scores Based on Various Cognitive-Developmental Models
(Listed by Size of Correlation)

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size</th>
<th>Sample age (where available)</th>
<th>Measure</th>
<th>Domain of Reasoning</th>
<th>Correlation and probability value (p)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>160</td>
<td>18 - 83</td>
<td>Good life interview</td>
<td>The good life</td>
<td>0.00, n.s.</td>
<td>Commons et al. (1989)</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>High school &amp; doctoral students</td>
<td>Defining Issues Test</td>
<td>Moral judgement</td>
<td>0.02, n.s. (pre-test only)</td>
<td>Kitchener et al. (1984)</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>16-28</td>
<td>Reflective judgement interview</td>
<td>Reflective judgement</td>
<td>0.05, n.s. (test-time 1 only)</td>
<td>King et al. (1989)</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>16-28</td>
<td>Defining Issues Test</td>
<td>Moral judgement</td>
<td>0.09, n.s. (test-time 1 only)</td>
<td>King et al. (1989)</td>
</tr>
<tr>
<td>5</td>
<td>160</td>
<td>18-83</td>
<td>Multi-systems assessment</td>
<td>Logico-mathematical</td>
<td>0.13, n.s.</td>
<td>Commons et al. (1989)</td>
</tr>
<tr>
<td>6</td>
<td>236</td>
<td>Early adolescents</td>
<td>Defining Issues Test</td>
<td>Moral judgement</td>
<td>0.19, p&lt;.05 (pre-test only)</td>
<td>Gfellner (1986a)</td>
</tr>
<tr>
<td>7</td>
<td>144</td>
<td>15-48</td>
<td>Defining Issues Test</td>
<td>Moral judgement</td>
<td>0.20, p&lt;.02</td>
<td>Skoe and von der Lippe (2002)</td>
</tr>
<tr>
<td>Study</td>
<td>Sample size</td>
<td>Sample age (where available)</td>
<td>Measure</td>
<td>Domain of Reasoning</td>
<td>Correlation and probability value (p)</td>
<td>Source</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>-----------------------------</td>
<td>---------</td>
<td>---------------------</td>
<td>--------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>8</td>
<td>193</td>
<td>First-year students through to seniors</td>
<td>Defining Issues Test</td>
<td>Moral judgement</td>
<td>0.21, p&lt;.002</td>
<td>Dunston and Roberts (1987)</td>
</tr>
<tr>
<td>9</td>
<td>160</td>
<td>18-83</td>
<td>Moral judgement interview</td>
<td>Moral judgement</td>
<td>0.22, n.s.</td>
<td>Commons et al. (1989)</td>
</tr>
<tr>
<td>10</td>
<td>58</td>
<td>16-35</td>
<td>Moral judgement interview</td>
<td>Moral judgement</td>
<td>0.30, n.s.</td>
<td>Haan et al. (1973)</td>
</tr>
<tr>
<td>11</td>
<td>60</td>
<td>17-30</td>
<td>Reflective interviews</td>
<td>Reflective judgement</td>
<td>0.36, p&lt;.01</td>
<td>Alishio and Schilling (1984)</td>
</tr>
<tr>
<td>12</td>
<td>125</td>
<td>Average of 19 years</td>
<td>Defining Issues Test</td>
<td>Moral judgement</td>
<td>0.38, p&lt;.01</td>
<td>Connelly and Lilienfield (2006)</td>
</tr>
<tr>
<td>13</td>
<td>120</td>
<td>12-17</td>
<td>Moral judgement interview</td>
<td>Moral judgement</td>
<td>0.40 (with age partialled out)</td>
<td>Sullivan et al. (1970)</td>
</tr>
<tr>
<td>14</td>
<td>34</td>
<td>12-25</td>
<td>Defining Issues Test</td>
<td>Moral judgement</td>
<td>0.41</td>
<td>Foster and Sprinthall (1992)</td>
</tr>
<tr>
<td>15</td>
<td>21</td>
<td>Graduate students</td>
<td>Moral development interview</td>
<td>Moral judgement</td>
<td>0.42</td>
<td>Liberman et al. (1983)</td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>27-42 prison inmates</td>
<td>Moral judgement interview</td>
<td>Moral judgement</td>
<td>0.42, p&lt;0.05 (pre-test only)</td>
<td>MacPhail (1989)</td>
</tr>
<tr>
<td>17</td>
<td>517</td>
<td>12-21</td>
<td>Defining Issues Test</td>
<td>Moral judgement</td>
<td>0.44, p&lt;.0001</td>
<td>Gfellner (1986b)</td>
</tr>
</tbody>
</table>
Appendix G: Illustrative Example of an SCT in the Public Domain

An illustrative example of two versions of the SCT in the public domain are presented below. Other versions of the SCT used in the context of this thesis are likely considered proprietary intellectual property, and therefore are not reproduced.

Washington University Sentence Completion Test (WU-SCT), Form 81 Part A For Men

“Instructions:
The table on the next 2 pages contains 18 incomplete sentences (sentence stems). Please finish each sentence. There are no right or wrong answers.

Please print your responses legibly.

The form should be completed within approximately 20 minutes.

1. When a child will not join in group activities ...
2. Raising a family ...
3. When I am criticized ...
4. A man’s job ...
5. Being with other people ...
6. The thing I like about myself is ...
7. My mother and I ...
8. What gets me into trouble is ...
9. Education ...
10. When people are helpless ...
11. Women are lucky because ...
12. A good father ...
13. A girl has a right to ...
14. When they talked about sex, I ...
15. A wife should ...
16. I feel sorry ...

17. A man feels good when ...

18. Rules are …”

(Hy & Loevinger, 1996, p. 28).

Washington University Sentence Completion Test (WU-SCT), Form 81 Part B For Men

“Instructions:

The table on the next 2 pages contains 18 incomplete sentences (sentence stems). Please finish each sentence. There are no right or wrong answers.

Please print your responses legibly.

The form should be completed within approximately 20 minutes.

19. Crime and delinquency could be halted if ...

20. Men are lucky because ...

21. I just can’t stand people who ...

22. At times he worried about ...

23. I am ...

24. A woman feels good when ...

25. My main problem is ...

26. A husband has a right to ...

27. The worst thing about being a man ...

28. A good mother ...

29. When I am with a woman ...

30. Sometimes he wished that ...

31. My father ...

32. If I can’t get what I want ...

33. Usually he felt that sex ...

34. For a woman a career is ...
35. My conscience bothers me if …

36. A man should always …”

(Hy & Loevinger, 1996, p. 29).
Appendix H: Participant Information and Consent Forms

Copy of the Participant Information and Consent Form (from the Author of this Thesis)

THE UNIVERSITY OF WESTERN AUSTRALIA

7 July, 2016

Participant Information Form

Dear Potential Research Participant,

You are being invited to participate in a research project conducted by Aiden Thornton (PhD Candidate) under the supervision of Professor David Day from the Business School at the University of Western Australia. The project is being undertaken in collaboration with both MetaIntegral™ and Lectica™.

Global complexity has increased significantly since the industrial revolution (circa 1760) and has rapidly become a key priority for senior leaders. This is often experienced via complex decisions which leaders make on a daily basis to support their organisations to thrive. Addressing this complexity requires leaders to have advanced capabilities to think, feel and act more wisely. However, this can be a challenge because leader development programs do not consistently support senior leaders to grow these advanced capabilities. Instead, they sometimes focus on changing behavior. As a result, more and more leaders report being overwhelmed by complexity and are sometimes feel “in over their heads”.

One way to support leaders to build these capabilities is to draw on developmental theory which describes how adult capabilities grow through levels of increasing complexity. Leader development programs based on these theories may be more likely support leaders to develop the capabilities required to thrive in a complex world.

The aim of this research project is to compare two developmental theories (and their associated assessment tools) to determine if one of them is better placed to inform the design and measurement of leader development programs which are focused on thriving in complexity. This has the potential to make a significant contribution to our understanding of leadership and adult development process and practice.

While there are a number of developmental theories available, arguably the two most commonly used in leadership are:

1. ego development (associated assessment methods are the Sentence Completion Test Integral (SCTI) and the Global Leadership Profile (GLP);
2. neo-Piagetian theory (associated assessment methods include Lectica™ Assessments).

The research project will involve a number of comparative analyses, including:

- ego development and neo-Piagetian approaches as a theoretical lens; and
- an exploration of how complexity is used to define developmental levels in both theories; and
- scores awarded both the SCTI scoring system and the Lectica Assessment System

This will allow us to determine whether both theories and assessments are actual tapping into “complexity”, and therefore are suited to inform leader development programs which focus on thriving in complexity.

You are being invited to participate in this study because you have recently taken an ego development metric (either the SCTI or the GLP, or both). Providing permission to participate in this research means that you are consenting to the primary researcher (Aiden Thornton, PhD candidate) accessing:

- your SCTI / GLP responses (that is, your responses to the actual sentence stems)
- your SCTI / GLP scores (which were scored by an expert ego developmentalist)

If you do provide consent, then your data will be used to:

- perform a range of mathematical and statistical analyses; and
- if you have also taken one or more Lectica Assessments (e.g. LDMA), your SCTI / GLP score will be mapped to your Lectica scores for comparative purposes.
How will your name be used?

- If you have only taken a SCLT / GLP and not a Lectical Assessment, your name will be completely removed from the primary researcher’s data file.
- If you have taken both types of assessments (i.e. SCLT / GLP and Lectical Assessment), then your name will only be used to map your SCLT / GLP score (from Metasimigra’s database) to your Lectical score(s) (from Lectica’s database). Once this mapping has taken place, your name will be completely removed from the primary researcher’s data file, so you cannot be personally identified.

By following this protocol, complete confidentiality and anonymity is assured. On no condition will any confidential data be disclosed to other persons or institutions. Data will only be used for research purposes in the hope of making a contribution to our understanding of leadership and adult development through publishing in scientific journals and professional conferences, etc.

We do not anticipate any risks associated with participation in this research project.

Participation in this project is completely voluntary and does not prejudice any right to compensation which you may have under statute or common law. If you choose to participate you are free to withdraw from the research at any time without prejudice in any way, and you need give no reason or justification for withdrawing.

Approval to conduct this research has been provided by The University of Western Australia, in accordance with its ethics review and approval procedures. Any person considering participation in this research project, or agreeing to participate, may raise any questions or issues with the researchers at any time.

- Aidan Thornton (PhD Candidate) (aidan.thornton@research.uwa.edu.au)
- Professor David Day (david.day@uwa.edu.au)

Any person not satisfied with the response of researchers may raise ethics issues or concerns, and may make any complaints about this research project by contacting the Human Research Ethics Office at The University of Western Australia on +61 (08) 6486 3703 or by emailing to hres-research@uwa.edu.au.

All research participants are entitled to retain a copy of the Participant Information Sheet and/or Participant Consent Form relating to this research project.
7 July, 2016

Participant Consent Form

I (the participant) have read the information provided and any questions I have asked have been answered to my satisfaction. I consent to my data being used for this research project, and future research being undertaken by Mataliogol Academy, realising that I may withdraw at any time without reason and without prejudice.

I understand that all identifiable (attributable) information that I provide is treated as strictly confidential and will not be released by the investigator in any form that may identify me. The only exception to this principle of confidentiality is if documents are required by law.

I have been advised as to what data is being collected, the purpose for collecting the data, and what will be done with the data upon completion of the research.

I agree that research data gathered for the study may be published provided my name or other identifying information is not used.

Once signed, please return directly to aiden.thornton@research.uwa.edu.au

Participant ___________________________ Date __________

Approval to conduct this research has been provided by The University of Western Australia, in accordance with its ethics review and approval procedures. Any person considering participation in this research project, or agreeing to participate, may raise any questions or issues with the researchers at any time.

In addition, any person not satisfied with the response of researchers may raise ethics issues or concerns, and may make any complaints about this research project by contacting the Human Research Ethics Office at The University of Western Australia on (08) 6488 3703 or by emailing to hrec-research@uwa.edu.au

All research participants are entitled to retain a copy of the Participant Information Sheet and/or Participant Consent Form relating to this research project.
INFORMATION SHEET FOR RESEARCH PARTICIPANTS

FACTORS AND PROCESSES INVOLVED IN ADULT CONSCIOUSNESS DEVELOPMENT

Please read and consider the following information before deciding whether to participate in this research.

PURPOSE OF THE STUDY

The focus and aim of this research is to better understand some of the cognitive and personality factors associated with the development of people's ways of understanding the world, as well as to increase understanding about the types of experiences that can trigger changes in the way people make sense of the world (including whether programs – such as the one you are currently participating in, can trigger such development).

WHAT PARTICIPANTS WILL BE ASKED TO DO AND HOW LONG IT WILL TAKE

Study participants will be asked to complete two psychological instruments upon commencement in their program and one at the end of the program. Completion of the first two instruments will take approximately 40 minutes. Completion of the final instrument at the end of the program will take approximately 20 minutes. In some cases participants may be asked to undertake a follow-up phone interview after completion of the last instrument (after the end of your program). If you agree to an interview then this will take approximately 30-45 minutes and will be undertaken at a time that is convenient to you.

The instruments that you will be asked to complete are as follows:

• The Washington University Sentence Completion Test (WUSCT) – a measure of your stage of consciousness;
• The Myers Briggs Personality Type Indicator (MBTI) – a measure of your personality preferences.

POSSIBLE BENEFITS FROM THE STUDY

Study participants will be provided with a brief report of their personal results on the instruments and a summary report of the outcomes of the research if they request this. As such, there may be some direct benefits to participants in the form of increased self-awareness and understanding from involvement in this research.

Senior staff of participating programs will be offered a summary report of the outcomes of the research and the pooled results for their particular program. Program staff will not be provided with the results of any individual participants in their program. However, they may be able to utilise the information from this research to improve the programs for which they have oversight.

The growth of adult consciousness is associated with many adaptive advantages for the individual and society. This research may therefore be able to make an important contribution to individuals as well as researchers and practitioners in the field of adult development.

The short to medium-term aim of this research is to inform the development of programs (such as the one you are currently participating in) to assist individuals to reach their maximum mental potential (should they wish to do so).

The longer term aim of this research is to assist in the development of a critical mass of people in influential leadership roles to higher levels of wisdom. It is envisioned that this may ultimately create a ‘snowball’ effect of higher level thinking that rolls throughout the institutions, systems and structures of our societies and eventually lifts everyone’s capacity.
POSSIBLE DISCOMFORT OR RISKS

There are no anticipated risks associated with participation in this study. The instruments you will be completing are in common use and the researcher (Niki Vincent) is accredited in their administration/scoring and trained to provide feedback to participants in relation to the results. If you are asked to participate in an interview at the end of your program, then the interview process will be transparent, with no trick questions. The interview will be recorded and the recording will be destroyed once it has been transcribed.

Some participants may experience minor discomfort in the form of anxiety associated with completing the instruments and participants may feel they have been inconvenienced by the time taken to complete the instruments (although this will be carried out during program time) and interview. All data will be kept confidential.

In some cases, program managers may elect to have the researcher conduct a workshop on understanding the Myers-Briggs Type Indicator for use within their program. If so, then participants may wish to disclose their MBTI results for the purposes of the workshop. However, this will be entirely voluntary and those who choose not to disclose their results will not be disadvantaged in any way and will still be included in the workshop.

WITHDRAWAL FROM THE STUDY

Your participation in this study is completely voluntary and you may withdraw at any time without penalty. Withdrawal will not prejudice your continued involvement in the leadership program that you are enrolled in at all.

CONFIDENTIALITY

The results of the tests that you take and all information provided in the interview (if you are interviewed) will be kept strictly confidential. Only the researcher (Niki Vincent) will have access to your results and your name. Information about your identity will not be stored with the data from your completed questionnaires. You will be assigned a unique identification number that you may use in completing all tests. The list connecting your name and this code number will be kept in a locked filing cabinet separate from the data and will be destroyed once all the data from this and
any possible follow-up studies have been collected and analysed. If you are selected for a follow-up interview and you agree to volunteer, you will have the option of choosing a pseudonym that will be used in any publications based on this study. If you do not choose a pseudonym then the researcher will choose one for you.

The results of this research reported in publications and a thesis will not identify any individual participants.

CONTACT INFORMATION

If you have questions or problems associated with the practical aspects of your participation in the project, or wish to raise a concern or complaint about the project, then you should consult the researcher:

Nicola (Niki) Vincent
Mobile: 0439 493 303
Email: nicola.vincent@adelaide.edu.au.

If you prefer, you may consult the supervisor of this research:
Dr Lynn Ward
Phone 08 8303 3182
Email: lynn.ward@adelaide.edu.au

You may also wish to consult Associate Professor Paul Delfabbro, who oversees ethical approval for research in the School of Psychology at the University of Adelaide. His phone number is 8303 4936 and email is paul.delfabbro@psychology.adelaide.edu.au

INDEPENDENT COMPLAINTS PROCEDURE

This research has been reviewed and approved by the University of Adelaide Human Research Ethics Committee. You may contact the Human Research Ethics Committee’s Secretary on (08) 8303 6028 if you wish to discuss with an independent person matters related to:

- Making a complaint, or
• Raising concerns on the conduct of the project, or
• The University policy on research involving human participants, or
• Your rights as a participant
THE UNIVERSITY OF ADELAIDE HUMAN RESEARCH ETHICS COMMITTEE
STANDARD CONSENT FORM
FOR PEOPLE WHO ARE PARTICIPANTS IN A RESEARCH PROJECT

1. I, ___________________________________________ (please print your name)
   consent to take part in the research project entitled: FACTORS AND PROCESSES INVOLVED IN ADULT
   CONSCIOUSNESS DEVELOPMENT

2. I acknowledge that I have read the attached Information Sheet entitled INFORMATION SHEET FOR RESEARCH
   PARTICIPANTS - FACTORS AND PROCESSES INVOLVED IN ADULT CONSCIOUSNESS DEVELOPMENT

3. I have had the project, so far as it affects me, fully explained to my satisfaction by the research worker. My
   consent is given freely.

4. Although I understand that the purpose of this research project is to improve programs to help people develop
   to their fullest mental potential, it has also been explained that my involvement may not be of any benefit to
   me.

5. I have been informed that, while information gained during the study may be published, I will not be identified
   and my personal results will not be divulged.

6. I understand that I may be contacted by the researcher and asked to participate in a follow-up interview or
   follow up studies once I have completed all of the tests at the end of the program.

7. I understand that if I am asked and agree to an interview, this interview will be recorded. I understand that the
   interview recording will be transcribed (typed up into a document) and the recording will be destroyed once
   the transcription has taken place.

8. I understand that I am free to withdraw from the project at any time and that this will not affect my
   participation in the leadership program in which I am enrolled now or in the future.

9. I am aware that I should retain a copy of this Consent Form, when completed, and the attached Information
   Sheet.

10. I give my consent to be contacted for future studies within the same area as the current study (please tick
    appropriate box) YES □ NO □

    ____________________________ (signature) ____________________________ (date)

WITNESS

I have described to: ___________________________________________ (name of participant)
the nature of the research to be carried out. In my opinion she/he understood the explanation.

Status in Project:

______________________________________________________________

Name: _______________________________________________________

_________________________ (signature) ____________________________ (date)
I, ________________________ (please print your name)
consent to take part in the research project entitled: FACTORS AND PROCESSES INVOLVED IN ADULT CONSCIOUSNESS DEVELOPMENT

2. I acknowledge that I have read the attached Information Sheet entitled INFORMATION SHEET FOR RESEARCH PARTICIPANTS - FACTORS AND PROCESSES INVOLVED IN ADULT CONSCIOUSNESS DEVELOPMENT.

3. I have read the project, so far as it affects me, fully explained to my satisfaction by the research worker. My consent is given freely.

4. Although I understand that the purpose of this research project is to improve programs to help people develop to their fullest mental potential, it has also been explained that my involvement may not be of any benefit to me.

5. I have been informed that, while information gained during the study may be published, I will not be identified and my personal results will not be divulged.

6. I understand that I may be contacted by the researcher and asked to participate in a follow-up interview or follow-up studies once I have completed all of the tests at the end of the program.

7. I understand that if I am asked and agree to an interview, this interview will be recorded. I understand that the interview recording will be transcribed (typed up into a document) and the recording will be destroyed once the transcription has taken place.

8. I understand that I am free to withdraw from the project at any time and that this will not affect my participation in the leadership program in which I am enrolled now or in the future.

9. I am aware that I should retain a copy of this Consent Form, when completed, and the attached Information Sheet.

10. I give my consent to be contacted for future studies within the same area as the current study (please tick appropriate box) YES ☐ NO ☐

_____________________________ (signature) __________________________ (date)

WITNESS

I have described to ________________________ (name of participant) the nature of the research to be carried out. In my opinion she/he understood the explanation.

Status in Project:

_____________________________ __________________________

Name: ____________________________

_____________________________ __________________________

(signature) (date)
2. Informed Consent JFKU

Edit Question  Move  Copy  Delete

INFORMED CONSENT
John F. Kennedy University

TITLE OF STUDY: iTeach
PRIMARY INVESTIGATOR: Sean Esbjörn-Hargens, PhD
CONTACT PHONE NUMBER: 925-969-3517

Purpose of the Study:
You are invited to participate in a research study. The purpose of this project is to conduct a longitudinal study using methods from all eight zones of Integral Methodological Pluralism to assess the transformative effects of integral education. In collaboration working with Theo Dawson of Developmental Testing Service and Susanne Cook-Greuter of Cook-Greuter and Associates this project will ask you to complete a variety of assessments that will examine your development in different domains and different aspect of your background and personality. All information will be used for several purposes including first and foremost how to improve educational programs to increase growth in target areas, 2) to determine what asset may help or hinder success of students, and 3) to share information obtained with a larger community to model research and educational improvement for other programs.

Participants:
You are being asked to participate in the study because you are at least 18 years old and a member of a target educational program at a University that is participating in this study. We believe that the information you provide will help us to obtain a better understanding about the ways in which educational programming can be improved.

Procedures:
If you volunteer to participate in this study, you will be asked to complete various questionnaires, assessments, and report general demographic information over an extended period of time. Depending on the portion of the research, it may take a minimum of one hour or up to ten to fifteen hours to complete sets of questionnaires. Depending on the length of the assessment, you will not be asked to complete the questionnaire(s) in one sitting, but may
have up to several weeks for completion. We ask that you answer each question carefully and honestly. Questionnaires administered may address cognitive abilities, emotional capacity, interpersonal skills, and/or knowledge of Integral Theory.

Benefits of Participation:
There are no guaranteed benefits to participating in this study other than adding to the general body of knowledge on this subject. In some cases you will be provided the results, which can be used for your own inquiry and reflection. In addition, this process is designed to be educational thereby providing you direct engagement with forms of testing and assessment so that you can learn directly about these and the kind of knowledge that they can produce.

Risks of Participation:
There are risks involved in all research studies. This study may include only minimal risks. You may experience some mild and temporary discomfort when completing the questions. If for any reason discomfort as a result of completing the questions is prolonged, you will be emailed with a list of occupational referrals to seek appropriate services at your request.

Cost /Compensation:
There will be no financial cost to you to participate in this study. The time the study will take will vary from one hour to 10-15 hours. Participants will not be penalized for withdrawing from the study or any part of it. There is no other compensation provided for time.

Voluntary Participation:
Your participation in this study is voluntary. You may refuse to participate in this study or in any part of this study. You may withdraw at any time without prejudice to your relations with the university. You are encouraged to ask questions about this study at any time by calling or emailing a member of the research team.

Confidentiality:
All information gathered in this study will be kept completely confidential. No reference will be made in written or oral materials that could link you to this study. No identifying information from your computer will be tracked or requested. Individual names are collected to consolidate and link individual data. Once material is consolidated, names will be kept in a separate data set to increase levels of confidentiality. Only main researchers will have access to connect specific data with names and that information will be maintained to provide individual feedback when possible if desired by the student, but will never be revealed in research presentations or without further permission from a participant. In some cases you will be provided the results of various assessments. Data will be stored for at least five years after which it may be destroyed.
Copy of Participant Consent Form Employed by Other Researchers (3 of 3)

Thank you for completing the GLP. You will shortly receive a Self-Estimate document, and our GLP assessment and report will reach you within 14 working days, unless otherwise agreed. As part of our ongoing research, development and training, it will be helpful if we may have your permission to use, with full anonymity, some of your responses. If you do not wish us to do so, please check this box. □
### Appendix I: SCT Stems Analysed in This Thesis

#### Table I 1

**SCT Stems Analysed in This Thesis**

<table>
<thead>
<tr>
<th>SCT Stem</th>
<th>Version 1 of SCT</th>
<th>Version 2 of SCT</th>
<th>Version 3 of SCT</th>
<th>Version 4 &amp; 5 of SCT</th>
<th>Naming Convention Used in this Thesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raising a family</td>
<td>stem_01</td>
<td>stem_01</td>
<td>stem_06</td>
<td>stem_02</td>
<td>stem_01</td>
</tr>
<tr>
<td>When I'm criticized</td>
<td>stem_02</td>
<td>stem_02</td>
<td>stem_03</td>
<td>stem_03</td>
<td>stem_02</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>stem_03</td>
<td>not used</td>
<td>not used</td>
<td>stem_03</td>
</tr>
<tr>
<td>When a child will not join in group activities</td>
<td>stem_03</td>
<td>not used</td>
<td>stem_01</td>
<td>stem_01</td>
<td>stem_04</td>
</tr>
<tr>
<td>A man's job</td>
<td>stem_04</td>
<td>stem_04</td>
<td>stem_08</td>
<td>stem_04</td>
<td>stem_05</td>
</tr>
<tr>
<td>Being with other people</td>
<td>stem_05</td>
<td>stem_05</td>
<td>stem_04</td>
<td>stem_05</td>
<td>stem_06</td>
</tr>
<tr>
<td>The thing I like about myself is</td>
<td>stem_06</td>
<td>stem_06</td>
<td>stem_05</td>
<td>stem_06</td>
<td>stem_07</td>
</tr>
<tr>
<td>My mother and I</td>
<td>stem_29</td>
<td>stem_07</td>
<td>not used</td>
<td>stem_07</td>
<td>stem_08</td>
</tr>
<tr>
<td>What gets me into trouble is</td>
<td>stem_08</td>
<td>stem_08</td>
<td>stem_09</td>
<td>stem_08</td>
<td>stem_09</td>
</tr>
<tr>
<td>Education</td>
<td>stem_09</td>
<td>stem_09</td>
<td>stem_02</td>
<td>stem_09</td>
<td>stem_10</td>
</tr>
<tr>
<td>When people are helpless</td>
<td>stem_10</td>
<td>stem_10</td>
<td>stem_07</td>
<td>stem_10</td>
<td>stem_11</td>
</tr>
<tr>
<td>Women are lucky because</td>
<td>stem_11</td>
<td>stem_11</td>
<td>not used</td>
<td>stem_11</td>
<td>stem_12</td>
</tr>
<tr>
<td>Left blank</td>
<td>stem_12</td>
<td>stem_12</td>
<td>not used</td>
<td>not used</td>
<td>stem_13</td>
</tr>
<tr>
<td>SCT Stem</td>
<td>Version 1 of SCT</td>
<td>Version 2 of SCT</td>
<td>Version 3 of SCT</td>
<td>Version 4 &amp; 5 of SCT</td>
<td>Naming Convention Used in this Thesis</td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>----------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>A girl has a right to</td>
<td>stem_13</td>
<td>stem_13</td>
<td>stem_13</td>
<td>stem_13</td>
<td>stem_14</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>stem_14</td>
<td>not used</td>
<td>not used</td>
<td>stem_15</td>
</tr>
<tr>
<td>When they talked about sex, I</td>
<td>stem_14</td>
<td>stem_15</td>
<td>not used</td>
<td>stem_14</td>
<td>stem_16</td>
</tr>
<tr>
<td>A wife should</td>
<td>stem_15</td>
<td>not used</td>
<td>not used</td>
<td>stem_15</td>
<td>stem_17</td>
</tr>
<tr>
<td>I feel sorry</td>
<td>stem_16</td>
<td>stem_16</td>
<td>stem_16</td>
<td>stem_16</td>
<td>stem_18</td>
</tr>
<tr>
<td>Left blank</td>
<td>stem_17</td>
<td>stem_17</td>
<td>stem_14</td>
<td>not used</td>
<td>stem_19</td>
</tr>
<tr>
<td>Rules are</td>
<td>stem_18</td>
<td>stem_18</td>
<td>stem_18</td>
<td>stem_18</td>
<td>stem_20</td>
</tr>
<tr>
<td>Crime and delinquency could be halted if</td>
<td>stem_19</td>
<td>stem_19</td>
<td>not used</td>
<td>stem_19</td>
<td>stem_21</td>
</tr>
<tr>
<td>Men are lucky because</td>
<td>stem_20</td>
<td>stem_20</td>
<td>not used</td>
<td>stem_20</td>
<td>stem_22</td>
</tr>
<tr>
<td>I just can’t stand people who</td>
<td>stem_21</td>
<td>stem_21</td>
<td>stem_12</td>
<td>stem_21</td>
<td>stem_23</td>
</tr>
<tr>
<td>At times s/he worried about (&quot;S/he&quot; should be read as &quot;she&quot; by women, and &quot;he&quot; by men)</td>
<td>stem_22</td>
<td>stem_22</td>
<td>not used</td>
<td>stem_22</td>
<td>stem_24</td>
</tr>
<tr>
<td>I am</td>
<td>stem_23</td>
<td>stem_23</td>
<td>stem_30</td>
<td>stem_23</td>
<td>stem_25</td>
</tr>
<tr>
<td>Left blank</td>
<td>stem_24</td>
<td>stem_24</td>
<td>not used</td>
<td>not used</td>
<td>stem_26</td>
</tr>
<tr>
<td>My main problem is</td>
<td>stem_25</td>
<td>stem_25</td>
<td>stem_25</td>
<td>stem_25</td>
<td>stem_27</td>
</tr>
<tr>
<td>Left blank</td>
<td>stem_26</td>
<td>stem_26</td>
<td>stem_20</td>
<td>not used</td>
<td>stem_28</td>
</tr>
<tr>
<td>Left blank</td>
<td>stem_27</td>
<td>stem_27</td>
<td>stem_21</td>
<td>not used</td>
<td>stem_29</td>
</tr>
<tr>
<td>A husband has a right to</td>
<td>stem_28</td>
<td>stem_28</td>
<td>not used</td>
<td>stem_26</td>
<td>stem_30</td>
</tr>
<tr>
<td>Left blank</td>
<td>stem_07</td>
<td>stem_29</td>
<td>stem_10</td>
<td>not used</td>
<td>stem_31</td>
</tr>
<tr>
<td>SCT Stem</td>
<td>Version 1 of SCT</td>
<td>Version 2 of SCT</td>
<td>Version 3 of SCT</td>
<td>Version 4 &amp; 5 of SCT</td>
<td>Naming Convention Used in this Thesis</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>----------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>Usually s/he felt that sex (&quot;S/he&quot; should be read as &quot;she&quot; by women, and &quot;he&quot; by men)</td>
<td>stem_30</td>
<td>not used</td>
<td>not used</td>
<td>stem_33</td>
<td>stem_32</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>stem_30</td>
<td>not used</td>
<td>not used</td>
<td>stem_33</td>
</tr>
<tr>
<td>My father</td>
<td>stem_31</td>
<td>stem_31</td>
<td>stem_22</td>
<td>stem_31</td>
<td>stem_34</td>
</tr>
<tr>
<td>If I can't get what I want</td>
<td>stem_32</td>
<td>stem_32</td>
<td>not used</td>
<td>stem_32</td>
<td>stem_35</td>
</tr>
<tr>
<td>Left blank</td>
<td>stem_33</td>
<td>stem_33</td>
<td>not used</td>
<td>not used</td>
<td>stem_36</td>
</tr>
<tr>
<td>For a woman a career is</td>
<td>stem_34</td>
<td>stem_34</td>
<td>not used</td>
<td>stem_34</td>
<td>stem_37</td>
</tr>
<tr>
<td>My conscience bothers me if</td>
<td>stem_35</td>
<td>stem_35</td>
<td>stem_23</td>
<td>stem_35</td>
<td>stem_38</td>
</tr>
<tr>
<td>Sometimes s/he wished that (&quot;S/he&quot; should be read as &quot;she&quot; by women, &quot;he&quot; by men)</td>
<td>stem_36</td>
<td>stem_36</td>
<td>not used</td>
<td>stem_30</td>
<td>stem_39</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_11</td>
<td>not used</td>
<td>stem_40</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_15</td>
<td>not used</td>
<td>stem_41</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_17</td>
<td>not used</td>
<td>stem_42</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_19</td>
<td>not used</td>
<td>stem_43</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_24</td>
<td>not used</td>
<td>stem_44</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_26</td>
<td>not used</td>
<td>stem_45</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_27</td>
<td>not used</td>
<td>stem_46</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_28</td>
<td>not used</td>
<td>stem_47</td>
</tr>
<tr>
<td>Left blank</td>
<td>not used</td>
<td>not used</td>
<td>stem_29</td>
<td>not used</td>
<td>stem_48</td>
</tr>
<tr>
<td>A good father</td>
<td>not used</td>
<td>not used</td>
<td>not used</td>
<td>stem_12</td>
<td>stem_49</td>
</tr>
</tbody>
</table>
### SCT Stem Version 1 of SCT Version 2 of SCT Version 3 of SCT Version 4 & 5 of SCT Naming Convention Used in this Thesis

<table>
<thead>
<tr>
<th>SCT Stem</th>
<th>Version 1 of SCT</th>
<th>Version 2 of SCT</th>
<th>Version 3 of SCT</th>
<th>Version 4 &amp; 5 of SCT</th>
<th>Naming Convention Used in this Thesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man feels good when</td>
<td>not used</td>
<td>not used</td>
<td>not used</td>
<td>stem_17</td>
<td>stem_50</td>
</tr>
<tr>
<td>A woman feels good when</td>
<td>not used</td>
<td>not used</td>
<td>not used</td>
<td>stem_24</td>
<td>stem_51</td>
</tr>
<tr>
<td>The worst thing about being a wo/man (should be read as &quot;woman&quot; by females and &quot;man&quot; by males)</td>
<td>not used</td>
<td>not used</td>
<td>not used</td>
<td>stem_27</td>
<td>stem_52</td>
</tr>
<tr>
<td>A good mother</td>
<td>not used</td>
<td>not used</td>
<td>not used</td>
<td>stem_28</td>
<td>stem_53</td>
</tr>
<tr>
<td>When I am with a wo/man (should be read as &quot;man&quot; by females, and &quot;woman&quot; by males)</td>
<td>not used</td>
<td>not used</td>
<td>not used</td>
<td>stem_29</td>
<td>stem_54</td>
</tr>
<tr>
<td>A wo/man should always (should be read as &quot;man&quot; by males, and &quot;woman&quot; by females)</td>
<td>not used</td>
<td>not used</td>
<td>not used</td>
<td>stem_36</td>
<td>stem_55</td>
</tr>
</tbody>
</table>

**Note.** Some versions of the SCT used in the context of this thesis are considered proprietary intellectual property. Unique stems used in these versions have been removed from Table I1 and are indicated by *left blank*, so that only stems that are in the public domain are included.
Appendix J: Additional Rasch Analyses Conducted on the SCT

Step 4: Test for Other General Measurement Properties

The purpose of this step was to undertake additional analyses on general measurement properties of the SCT.

Differential Item Functioning (DIF)

Differential item functioning (DIF) occurs when two or more groups of persons with identical ability have non-identical expected responses. Uniform DIF occurs when one group is consistently awarded higher scores across all ability levels. Non-uniform DIF occurs when there is an interaction between group membership and ability (Andrich & Hagquist, 2012). Given the linked data structure, it was critical to identify and delete SCT stems exhibiting DIF from the analysis (Loover & Mulligan, 2009). Five demographic variables were analysed: leadership cohort, first language, sex, birth year and age at test time. Statistical analyses were conducted via two-way ANOVAs (refer to Table J1), and visual analyses were conducted via Item Characteristic Curves (ICCs). In the context of two-way ANOVAs, a significant main effect for class interval indicates a general lack of fit to the ICC irrespective of group classification, a significant main effect for group indicates uniform DIF, and a significant class interval x group interaction indicates non-uniform DIF. A Bonferroni correction yielded a critical alpha of .004. Analyses of ICCs were used as a supplementary method of identifying DIF patterns which may not have reached statistical significance. In all cases, random sampling was used to obtain groups of comparable size to avoid bias. RUMM performed calculations based on the default of 10 class intervals.
Table J1

Differential Item Functioning Analyses Conducted on the SCT

<table>
<thead>
<tr>
<th>#</th>
<th>Group</th>
<th>Category</th>
<th>n</th>
<th>Main effect of class intervals</th>
<th>Main effect of groups</th>
<th>Interaction effect</th>
<th>Uniform DIF</th>
<th>Non-uniform DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cohort</td>
<td>Cohort 1</td>
<td>115</td>
<td>n.s.</td>
<td>Stem 31, p=0.0003</td>
<td>Stem 25, p=0.0001</td>
<td>Stem 31</td>
<td>Stem 25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cohort 2</td>
<td>115</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cohort 3</td>
<td>115</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cohort 4</td>
<td>115</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Language</td>
<td>Native English speakers</td>
<td>135</td>
<td>Stem 14, p=0.0003</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Native English speakers</td>
<td>135</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Sex</td>
<td>Male</td>
<td>349</td>
<td>Stem 6, p=0.00002</td>
<td>Stem 5, p=0.0004</td>
<td>n.s.</td>
<td>Stem 5</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Female</td>
<td>349</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Birthyear</td>
<td>1940-1959</td>
<td>70</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1960-1979</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1980-1999</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Age</td>
<td>18–35 years</td>
<td>183</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36–45 years</td>
<td>180</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>46–71 years</td>
<td>180</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some ICCs did reveal irregularities, but because of the limited group sizes, it was difficult to draw clear inferences about DIF. Statistically, stem 5, 25 and 31 appeared to display DIF. However, inspection of the ICCs suggested that the magnitude of DIF was relatively mild, and its source was unclear based on stem content (refer to Figure J1). As a result, it was not deemed necessary to resolve DIF or take further action. In future studies, a larger sample of SCTs may be required to identify the presence of DIF.
**Figure J 1**

*Item Characteristic Curves for Stems 31, 25, and 5*

- **Item: Descriptor for Item 31 ([0031]) - 4 Levels for Person Factor: CCHORT**
  - Slope: 1.25
  - ExpV: 3.10

- **Item: Descriptor for Item 25 ([0025]) - 4 Levels for Person Factor: CCHORT**
  - Slope: 1.04
  - ExpV: 2.90

Legend:
- • AIA
- × EPC
- # JFKU
- + NV
Response Dependence

If responses are independent, each stem makes a related, but unique, contribution to the total score. Local dependence Type 2, or response dependence, is an inter-stem effect which occurs when, for the same person and therefore for the same value of $\beta$, the location of one stem $j$ is governed by its relation to another stem $i$ (Marais & Andrich, 2008a, 2008b). In this way, the responses are only not a function of the parameters of the model, but also a function of the response to the first stem. The algebraic representation of response dependence for polytomous items is beyond the current scope. It is sufficient to say that $d$ characterises the magnitude of response dependence resulting from a change in location to stem $j$ caused by its dependence on stem $i$. Due to the linked design, analysis of local independence could not be undertaken on the aggregated sample. Instead, analysis was conducted on subsamples, including the 11 stems shared by all versions of the SCT. This enabled the broadest range of stems to be tested. For all five subsamples, threshold disorder was resolved by rescoring categories or deleting stems if rescoring was unsuccessful. Response dependence occurs when pairs of stems are more related than they are to other stems due to shared variance that cannot be attributed to the Rasch factor. Such dependence is revealed when correlations between standardised stem residuals exceed +0.3 (Andrich & Kreiner, 2010). As can be seen from Table 24 in Chapter 2, highest correlations did not reach the threshold (i.e., $+0.05 \leq r_{\text{resid}} \leq +0.29$). For subsample 2, the correlation of +0.29
approached +0.30 and occurred between stem 34 ("For a woman, a career is ...") and stem 35 ("My conscience bothers me if ...") which are not clearly related in terms of content. As a result, evidence in support of local dependence Type 2, or response dependence, was not identified.

**Measurement Precision**

The measurement precision of thresholds, $\tau_n$, is critical in the context of a development assessment because it determines how accurately thresholds, and therefore ego development stages, can be located on the logit scale. Standard Errors (SE) for thresholds appeared to be low ($M=0.15$, $SD=0.18$, $min=0.07$, $max=2.35$). As a result, thresholds can be confidently located within a 0.59 logit range which appears to be relatively precise in the context of a 12.7 logit range for thresholds (i.e., $2 \times 1.96 \times 0.15 = 0.59$).

The measurement precision of persons is also critical in the context of a development assessment because it determines how confidently leaders can be placed into ego development stages. For persons, SEs appeared to be somewhat higher than for thresholds ($mean=0.27$, $SD=0.03$, $min=0.21$, $max=0.34$). The range of a 95% confidence interval for a person ability is 1.06 logits (i.e., $2 \times 1.96 \times 0.27 = 1.06$). In the context of a 5.79 logit range for person ability, estimates appeared to be moderately precise since approximately five confidence intervals can fit within this range ($5.79/1.06 = 5.46$). However, this finding is not consistent with the claim leaders be placed into seven (or more) statistically distinct ego development stages. This finding is consistent with the interpretation of the reliability indices which suggested that the ego development scoring system can distinguish between approximately five distinct stages.
Appendix K: Person-Item Threshold Distribution for a Particular Version of the SCT

Figure K 1

Note. Threshold locations are illustrated in the lower part of Figure K1. As it can be seen, thresholds, \( \tau_{ij} \), do not have a consistent location which means that ego development stages cannot be clearly identified along the logit scale. This suggests that stages do not reflect a consistent amount of complexity and/or development across different stems.
Exemplars for the stem “When a child will not join in group activities ... “ were scored in-context and out-of-context of the SCT sentence stem as discussed in Study 2B. Detailed findings for scoring the exemplars in context were presented in Study 2B. Findings for scoring the exemplars out-of-context are presented in an abbreviated format below.

**Figure L 1**

*Boxplot of Lectical Phase Scores Awarded to Exemplars for Each Ego Development Stage (Scored Out-of-Context of the SCT Stem)*
Table L 1

Descriptive statistics for Lectical Phase Scores Awarded to Exemplars (Impulsive to Unitive) (Scored Out of Context of the SCT Stem)

<table>
<thead>
<tr>
<th>Stage</th>
<th>No. Exemplars</th>
<th>Min Lectical Phase</th>
<th>Max Lectical Phase</th>
<th>Range of Lectical Complexity Levels</th>
<th>Range of Lectical Phases</th>
<th>Mean Lectical Phase</th>
<th>Median Lectical Phase</th>
<th>Modal Lectical Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impulsive</td>
<td>3</td>
<td>07b</td>
<td>08b</td>
<td>2</td>
<td>5</td>
<td>07c</td>
<td>07b</td>
<td>07b</td>
</tr>
<tr>
<td>Opportunist</td>
<td>13</td>
<td>07c</td>
<td>10a</td>
<td>4</td>
<td>11</td>
<td>08c</td>
<td>08c</td>
<td>08b</td>
</tr>
<tr>
<td>Diplomat</td>
<td>17</td>
<td>07c</td>
<td>10c</td>
<td>4</td>
<td>13</td>
<td>09a</td>
<td>09b</td>
<td>09c</td>
</tr>
<tr>
<td>Expert</td>
<td>64</td>
<td>07b</td>
<td>11a</td>
<td>5</td>
<td>16</td>
<td>09d</td>
<td>09d</td>
<td>10b</td>
</tr>
<tr>
<td>Achiever</td>
<td>47</td>
<td>07d</td>
<td>11a</td>
<td>5</td>
<td>14</td>
<td>10a</td>
<td>10b</td>
<td>10b</td>
</tr>
<tr>
<td>Individualist</td>
<td>19</td>
<td>09c</td>
<td>11a</td>
<td>3</td>
<td>7</td>
<td>10b/10c</td>
<td>10c</td>
<td>10c</td>
</tr>
<tr>
<td>Strategist</td>
<td>6</td>
<td>10c</td>
<td>11a</td>
<td>2</td>
<td>3</td>
<td>10d</td>
<td>10c/10d</td>
<td>10c</td>
</tr>
<tr>
<td>Integrated</td>
<td>2</td>
<td>10d</td>
<td>11a</td>
<td>2</td>
<td>2</td>
<td>10d/11a</td>
<td>10d/11a</td>
<td>10d</td>
</tr>
<tr>
<td>Construct-Aware</td>
<td>20</td>
<td>10a</td>
<td>11c</td>
<td>2</td>
<td>7</td>
<td>10d</td>
<td>10d</td>
<td>10d</td>
</tr>
<tr>
<td>Unitive</td>
<td>26</td>
<td>10c</td>
<td>11b</td>
<td>2</td>
<td>4</td>
<td>10d/11a</td>
<td>11a</td>
<td>11a</td>
</tr>
</tbody>
</table>

Figure L 2

Boxplot of Ego Development Stage Exemplars and Lectical Phase Scores: Three Scenarios (Scored Out-of-Context of the SCT Stem)
Note. Scenario 1: seven stages of growth from Impulsive to Construct-Aware and five Lectical Levels of growth from 07b to 11c. Scenario 2: six stages of growth from Opportunist to Construct-Aware and no Lectical growth from 10a. Scenario 3: four stages of growth from Expert to Construct-Aware and a regression by one Lectical Level from 11a to 10a.

A one-way Analysis of Variance (ANOVA) was conducted to identify whether there was significant difference in the mean Lectical Phase Scores awarded to exemplars between successive stages. The Impulsive (n=3) and Integrated (n=2) stages were removed from the analysis due to the small number of exemplars available. The ANOVA was significant (F(7, 47.37)=43.14, p<.001). Levene’s test of equality of error variances was also significant (F(7,204)=8.24, p<.001) so post hoc comparisons were conducted using the Games-Howell Test which does not assume equality of variances between groups. For all pairwise comparisons, there was a nonsignificant difference. The analysis was repeated using non-parametric procedures (i.e., Kruskal-Wallis and Mann-Whitney U tests) but yielded identical findings, so are not reported. Pairwise comparisons are presented in Table L2.

**Table L2**

*Means, Standard Deviations, Mean Differences Based on One-Way Analysis of Variance (ANOVA) on Lectical Phase Scores Between Successive Ego Development Stages (Scored Out of Context)*

<table>
<thead>
<tr>
<th>Stage</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Diplomat</td>
<td>17</td>
<td>09a</td>
<td>3.62</td>
<td>-</td>
<td>-2.85</td>
<td>-4.22**</td>
<td>-5.65***</td>
<td>-7.01***</td>
<td>-7.63***</td>
<td>-7.68***</td>
<td></td>
</tr>
<tr>
<td>4. Expert</td>
<td>64</td>
<td>09d</td>
<td>3.79</td>
<td>-</td>
<td>-1.37</td>
<td>-2.80***</td>
<td>-4.16***</td>
<td>-4.78***</td>
<td>-4.83***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Achiever</td>
<td>47</td>
<td>10a</td>
<td>2.75</td>
<td>-</td>
<td>-1.43</td>
<td>-2.79**</td>
<td>-3.41***</td>
<td>-3.46***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Individualist</td>
<td>19</td>
<td>10b/10c</td>
<td>1.68</td>
<td>-</td>
<td>-1.36</td>
<td>-1.98*</td>
<td>-2.03***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Strategist</td>
<td>6</td>
<td>10d</td>
<td>0.98</td>
<td>-</td>
<td>-</td>
<td>-0.62</td>
<td>-0.667</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Construct-Aware</td>
<td>20</td>
<td>10d</td>
<td>1.57</td>
<td>-</td>
<td>-</td>
<td>-0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Unitive</td>
<td>26</td>
<td>10d/11a</td>
<td>0.71</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. * p < .05, ** p < .01, *** p < .001*
Appendix M: Syntax for all Analyses Presented in Chapter 4

TITLE: Preliminary analysis on ICC to evaluate clustering in the data

DATA: FILE IS CGMPlus.dat;

VARIABLE: NAMES ARE
dtkey !unique identifier for each assessment
lakey !another unique identifier for each assessment
userkey !unique identifier for each leader/test-taker
instru !type of lexical assessment e.g., LDMA
assignt !leader test-time
clas !ICLAS score for this assessment
provider !leadership provider
programc !leadership program code
cohortc !leadership training cohort code
geo !geographical location
grade !number of years of education
birth !birth year
lang !first language code
manlo !original management layers
manlc !collapsed management layers
manla !adjusted management layers
manlj !management layers quadratic function
p1k4 !management layers piece 1 with knot at management layer 4
p2k4 !management layers piece 2 with knot at management layer 4
p1k3 !management layers piece 1 with knot at management layer 3
p2k3 !management layers piece 2 with knot at management layer 3
mo !management layer dummy variable 0
m1 !management layer dummy variable 1
m2 !management layer dummy variable 2
m3 !management layer dummy variable 3
sector !sector code
sex !sex code
ethnic !ethnicity code
wc !total word count
uwc !unique word count
lb !lower boundary for role complexity
mp !mid-point for role complexity
ub !upper boundary for role complexity
svsr !skill vs. role complexity i.e., score<LB, LB<score<MP, MP<score<UB, score>UB
;

USEVARIABLE ARE
clas
;

MISSING ARE all (-999);
USEOBSERVATIONS ARE
assign == 1 AND
instru ==1;

CLUSTER IS
provider;
programc;
cohortc;

ANALYSIS: TYPE IS TWOLEVEL BASIC;

OUTPUT: STDYX
;

TITLE:  Question 1. Testing for relation between management level and mean CLAS score

DATA:  FILE IS CGMPlus.dat;

VARIABLE: NAMES ARE !All variables are defined above
dtkey
lakey
userkey
instru
assignt
clas
provider
programc
cohortc
geo
grade
birth
lang
manlo
manlc
manla
manlq
p1k4
p2k4
p1k3
p2k3
mo
m1
m2
m3
sector
sex
ethnic
wc
uwc
lb
mp
ub
svr
;

USEVARIABLE ARE
clas
manla
manlq
p1k4
p2k4
p1k3
p2k3
m1
m2
m3
manl10 !define new variable as calculated below
manle !define new variable as calculated below
;

MISSING ARE all (-999);

USEOBSERVATIONS IS
SUBPOPULATION IS
  instru == 1 AND
  assignt == 1;

CLUSTER IS
  provider;
  programc
  cohor tc;

DEFINE:
  manl10 = log10(manlc); !log 10 of management layer
  manle = log(manlc); !natural log of management layer

ANALYSIS: TYPE IS COMPLEX;

MODEL:
  clas ON manla;
  clas ON manla manlq;
  clas ON p1k4 p2k4;
  clas ON p1k3 p2k3;
  clas ON m1 m2 m3;
  clas ON manl10;
  clas ON manle;

OUTPUT:
  SAMPSTAT
  STDYX
  RESIDUAL
  MODINDICES(ALL)
  TECH4
TITLE: Question 2. Testing for differences in CLAS scores between successive management layers

DATA: FILE IS CGMPlus.dat;

VARIABLE: NAMES ARE !All variables are defined above
    dtkey
    lakey
    userkey
    instru
    assignt
    clas
    provider
    programc
    cohortc
    geo
    grade
    birth
    lang
    manlo
    manlc
    manla
    manlq
    p1k4
    p2k4
    p1k3
    p2k3
    mo
    m1
    m2
    m3
    sector
    sex
    ethnic
    wc
    uwc
    lb
    mp
    ub
    svsr

USEVARIABLE ARE
    clas
    manlc

GROUPING ARE
    manlc (2=manlevel2 3=manlevel3 4=manlevel4 5=manlevel5);
MISSING ARE all (-999);

USEOBSERVATIONS ARE
instru == 1 AND
assignt == 1;

CLUSTER IS
provider;
programc;
cohortc;

ANALYSIS: TYPE IS COMPLEX;

MODEL: [clas];

MODEL manlevel2:
[clas] (a);

MODEL manlevel3:
[clas] (b);

MODEL manlevel4:
[clas] (c);

MODEL manlevel5:
[clas] (d);

MODEL CONSTRAINT:
NEW(diff1);
diff1=(a)-(b);

NEW(diff2);
diff2=(b)-(c);

NEW(diff3);
diff3=(c)-(d);

MODEL TEST:
diff1=0;
diff2=0;
diff3=0;

0=a-b;
0=b-c;
0=c-d;

OUTPUT: SAMPSTAT
STDYX
SAMP
;
TITLE: Question 3. Testing for gap between task demands of role and CLAS score at successive management layers

DATA: FILE IS CGMPlus.dat;

VARIABLE: NAMES ARE !All variables are defined above
    dtkey
    lakey
    userkey
    instru
    assignt
    clas
    provider
    programc
    cohortc
    geo
    grade
    birth
    lang
    manlo
    manlc
    manla
    manlq
    p1k4
    p2k4
    p1k3
    p2k3
    mo
    m1
    m2
    m3
    sector
    sex
    ethnic
    wc
    uwc
    lb
    mp
    ub
    svsr
    ;

    USEVARIABLE ARE
    clas
    ;

    MISSING ARE all (-999);

    USEOBSERVATIONS ARE
    instru == 5 AND
    assignt == 1 AND
manlc ==2
manlc ==3
manlc ==4
manlc ==5
;

CLUSTER IS
provider;
programc;
cohortc;

ANALYSIS: TYPE IS COMPLEX;

MODEL: [clas] (a);

MODEL CONSTRAINT:

NEW(diff1);
diff1=(a)-11.00; !LB for mgmt layer 2

NEW(diff1);
diff1=(a)-11.23; !MP for mgmt layer 2

NEW(diff1);
diff1=(a)-11.45; !UB for mgmt layer 2

NEW(diff1);
diff1=(a)-11.20; !LB for mgmt layer 3

NEW(diff1);
diff1=(a)-11.43; !MP for mgmt layer 3

NEW(diff1);
diff1=(a)-11.65; !UB for mgmt layer 3

NEW(diff1);
diff1=(a)-11.60; !LB for mgmt layer 4

NEW(diff1);
diff1=(a)-11.63; !MP for mgmt layer 4

NEW(diff1);
diff1=(a)-11.85; !UB for mgmt layer 4

NEW(diff1);
diff1=(a)-11.60; !LB for mgmt layer 5

NEW(diff1);
diff1=(a)-11.83; !MP for mgmt layer 5

NEW(diff1);
diff1=(a)-12.05; !UB for mgmt layer 5
MODEL TEST:
  diff1=0;

OUTPUT:    SAMPSTAT
  STDYX

TITLE:     Question 2. Extracting summary statistics and frequencies as input into chi-square test for complexity gap

DATA:      FILE IS CGMPlus.dat;

VARIABLE:  NAMES ARE !All variables are defined above
dtkey
lakey
userkey
instru
assignt
clas
provider
programc
cohortc
geo
grade
birth
lang
manlo
manlc
manla
manlq
p1k4
p2k4
p1k3
p2k3
mo
m1
m2
m3
sector
sex
ethnic
wc
uwc
lb
mp
ub
svsr

USEVARIABLE ARE
clas
manlc

GROUPING ARE
svsr (1=<LB 2=LB-MP 3=MP-UB 4=>UB);
svsr (1=<LB 2=<MP 3=<UB 4=>UB);
svsr (1=<LB 2=LB-MP 3=MP-UB);
svsr (1=<LB 2=LB-MP);

MISSING ARE all (-999);

USEOBSERVATIONS ARE
instru == 1 AND
assignt == 1 AND
manlc ==2
manlc ==3
manlc ==4
manlc ==5

OUTPUT: SAMPSTAT
STDX
SAMP;
Appendix N: Abbreviated Complexity Gap Analyses Undertaken on Various Lectical Assessments

Abbreviated results for Lectical Leadership Reasoning Assessment (LLRA)

The intraclass correlation (ICC) exceeded the critical value of 0.05, and the design effect (Deffs) exceeded the critical value of 2 (refer Table N1). As a result, an aggregated approach to multi-level modelling (MLM) was applied with cohort as the clustering variable.

Table N 1

Variability Attributable to Hierarchical Clustering in the Data Structure (LLRA)

<table>
<thead>
<tr>
<th>Clustering variable</th>
<th>ICCs for Mean CLAS score</th>
<th>Average cluster size (c)</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider</td>
<td>NA - all from same provider</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Program</td>
<td>NA - all from same provider</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Cohort</td>
<td>0.130</td>
<td>18.00</td>
<td>2.21</td>
</tr>
</tbody>
</table>

Fit statistics used to evaluate relative model fit were the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), sample-size adjusted BIC, and shared variance ($r^2$). Fit statistics for the linear, quadratic, piecewise (knot at management layer 3), and freed models are presented in Table N2.
Table N 2

*Fit Statistics for the Linear, Quadratic, Piecewise and Freed Models (LLRA)*

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>AIC</th>
<th>BIC</th>
<th>Sample adjusted BIC</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>306</td>
<td>-218.040</td>
<td>-206.869</td>
<td>-216.383</td>
<td>0.039</td>
</tr>
<tr>
<td>Quadratic</td>
<td>306</td>
<td>-216.087</td>
<td>-201.193</td>
<td>-213.879</td>
<td>0.039</td>
</tr>
<tr>
<td>Piecewise</td>
<td>306</td>
<td>-216.087</td>
<td>-201.193</td>
<td>-213.879</td>
<td>0.039</td>
</tr>
<tr>
<td>Freed</td>
<td>306</td>
<td>-216.087</td>
<td>-201.193</td>
<td>-213.879</td>
<td>0.039</td>
</tr>
</tbody>
</table>

The statistical evidence suggested that the linear model was the best fit, but this was not supported by confirmatory analyses which were undertaken with a Bonferroni correction (0.05/(2x2)=0.0125). The smaller samples may have impacted these findings for more senior management layers. There was a significant increase of 0.062 from management layer 2 (n=98, $M=11.297$, $s^2=0.028$) to management layer 3 (n=186, $M=11.359$, $s^2=0.030$) according to the Wald Test ($W=33.594$, df=1, $p=0.0000$) and $z$-test ($SE=0.018$, $p=0.000$). There was an increase of 0.052 in from management layer 3 (n=186, $M=11.359$, $s^2=0.030$) to management layer 4 (n=22, $M=11.410$, $s^2=0.017$) but this was not significant according to the Wald Test ($W=1.071$, df=1, $p=0.3008$) and the $z$-test ($SE=0.050$, $p=0.301$).

For the linear model (n=306), leaders at management layer 2 had a mean CLAS score of 11.298 ($SE=0.017$, $p=0.000$). There was an increase of 0.059 between successive management layers ($SE=0.011$, $p=0.000$). The residual variance in this model was significant ($s^2=0.028$, $SE=0.003$, $p=0.000$). The mean CLAS score relative to the role complexity range for each management layer is presented in Table N3. A Bonferroni correction was applied for 9 comparisons (0.05/9 $z$-tests = 0.006).
Table N 3

*Descriptive Statistics by Management Layer (LLRA)*

<table>
<thead>
<tr>
<th>Model</th>
<th>Management layer 2</th>
<th>Management layer 3</th>
<th>Management layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>98</td>
<td>186</td>
<td>22</td>
</tr>
<tr>
<td>M</td>
<td>11.297</td>
<td>11.359</td>
<td>11.410</td>
</tr>
<tr>
<td>s²</td>
<td>0.028</td>
<td>0.030</td>
<td>0.017</td>
</tr>
<tr>
<td>min</td>
<td>10.820</td>
<td>10.880</td>
<td>11.130</td>
</tr>
<tr>
<td>max</td>
<td>11.710</td>
<td>11.990</td>
<td>11.680</td>
</tr>
<tr>
<td>Z-test on difference between mean and lower boundary</td>
<td>z=0.297, SE=0.017, p=0.000</td>
<td>z=0.159, SE=0.017, p=0.000</td>
<td>z=0.010, SE=0.051, p=0.838</td>
</tr>
<tr>
<td>Z-test on difference between mean and midpont</td>
<td>z=0.067, SE=0.017, p=0.000</td>
<td>z=-0.071, SE=0.017, p=0.000</td>
<td>z=-0.220, SE=0.051, p=0.000</td>
</tr>
<tr>
<td>Z-test on difference between mean and upper boundary</td>
<td>z=-0.153, SE=0.017, p=0.000</td>
<td>z=-0.291, SE=0.017, p=0.000</td>
<td>z=-0.440, SE=0.051, p=0.000</td>
</tr>
</tbody>
</table>

For management layer 2, the mean CLAS score was located between the midpoint and the upper boundary. For management layer 3, the mean CLAS score was located between the lower boundary and the midpoint. For management layer 4, the mean CLAS score was located at the lower boundary.

To better understand whether CLAS scores were distributed as expected, a series of chi-square analyses were undertaken. For each management layer, the actual and expected distributions of CLAS scores relative to role complexity were compared and the results are presented in Table N4.
Table N 4

Comparison Between Expected and Actual Distribution of Scores (LLRA)

<table>
<thead>
<tr>
<th></th>
<th>Management Layer 2</th>
<th>Management Layer 3</th>
<th>Management Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=98</td>
<td>n=186</td>
<td>n=22</td>
</tr>
<tr>
<td>Less than lower</td>
<td>Expected</td>
<td>Actual</td>
<td>Expected</td>
</tr>
<tr>
<td>boundary</td>
<td>4.9 (5%)</td>
<td>4.0 (4.1%)</td>
<td>9.3 (5%)</td>
</tr>
<tr>
<td>Between lower</td>
<td>63.7 (65%)</td>
<td>35.0 (35.7%)</td>
<td>120.9 (65%)</td>
</tr>
<tr>
<td>boundary and midpoint</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between midpoint</td>
<td>24.5 (25%)</td>
<td>35.0 (35.7%)</td>
<td>46.5 (25%)</td>
</tr>
<tr>
<td>and upper boundary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greater than upper</td>
<td>4.9 (5%)</td>
<td>24.0 (24.5%)</td>
<td>9.3 (5%)</td>
</tr>
<tr>
<td>boundary</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One-sample chi-square analyses were conducted on each management layer. Analyses demonstrated significant differences between the expected and actual distributions for management layer 2 ($\chi^2(3, n=98)=92.05$, $p=0.00$), management layer 3 ($\chi^2(3, n=186)=45.22$, $p=0.00$), and management layer 4 ($\chi^2(3, n=22)=136.29$, $p=0.00$).

In summary, there is insufficient evidence to suggest the presence of a complexity gap at management layer 2. The mean score is located between the midpoint and upper boundary, and the actual distribution between the midpoint and upper boundary, and also above the upper boundary, appears to be somewhat higher than the expected distribution. There appears to be sufficient evidence to suggest a shift towards a complexity gap at management layer 3. While the mean CLAS score is located between the lower boundary and midpoint, the percentage of CLAS scores below the lower boundary is somewhat greater than expected. There is sufficient evidence to suggest the beginning of a complexity gap at management layer 4. The mean CLAS score is located
at the lower boundary and the percentage of CLAS scores below the lower boundary is considerably greater than expected.

**Abbreviated results for Lectical Self Understanding Assessment (LSUA)**

The intraclass correlation (ICCs) exceeded the critical value of 0.05, but the design effect (Deffs) did not exceed the critical value of 2 (refer Table N5). An aggregated approach to multi-level modelling (MLM) was still applied with cohort as the clustering variable.

**Table N 5**

*Variability Attributable to Hierarchical Clustering in the Data Structure (LSUA)*

<table>
<thead>
<tr>
<th>Clustering variable</th>
<th>ICCs for Mean CLAS score</th>
<th>Average cluster size (c)</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider</td>
<td>NA - all from same provider</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>NA - all from same provider</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohort</td>
<td>0.060</td>
<td>26.647</td>
<td>1.54</td>
</tr>
</tbody>
</table>

Fit statistics used to evaluate relative model fit were the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), sample-size adjusted BIC, and shared variance \( r^2 \). Fit statistics for the linear, quadratic, piecewise (knot at management layer 3), and freed models are presented in Table N6.
Table N 6

Fit Statistics for the Linear, Quadratic, Piecewise and Freed Models (LSUA)

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>AIC</th>
<th>BIC</th>
<th>Sample adjusted BIC</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>453</td>
<td>-236.998</td>
<td>-224.651</td>
<td>-234.172</td>
<td>0.039</td>
</tr>
<tr>
<td>Quadratic</td>
<td>453</td>
<td>-239.030</td>
<td>-222.567</td>
<td>-235.261</td>
<td>0.048</td>
</tr>
<tr>
<td>Piecewise</td>
<td>453</td>
<td>-239.030</td>
<td>-222.567</td>
<td>-235.261</td>
<td>0.048</td>
</tr>
<tr>
<td>Freed</td>
<td>453</td>
<td>-239.030</td>
<td>-222.567</td>
<td>-235.261</td>
<td>0.048</td>
</tr>
</tbody>
</table>

Statistical evidence could not distinguish between the quadratic, piecewise, or freed models.

The piecewise model appeared to be the best fit based on the confirmatory analyses which were undertaken with a Bonferroni correction (0.05/(2 x 2)=0.0125). However, the smaller samples may have impacted these findings for senior management layers. There was a significant increase of 0.085 from management layer 2 (n=171, $M=11.172$, $s^2=0.040$) to management layer 3 (n=252, $M=11.257$, $s^2=0.029$) according to the Wald Test ($W=18.910$, df=1, p=0.0000) and z-test (SE=0.020, p=0.000). There was a decrease of 0.002 from management layer 3 (n=252, $M=11.257$, $s^2=0.029$) to management layer 4 (n=30, $M=11.256$, $s^2=0.044$) but this was not significant according to the Wald Test ($W=0.001$, df=1, p=0.9782) and z-test (SE=0.063, p=0.978). As a result, the piecewise model was selected as the best fitting model.

For the piecewise model (n=453) with a knot at management layer 3, leaders at management layer 2 had a mean CLAS score of 11.172 (SE=0.016, p=0.000). For piece 1, which occurred between management layers 2 through 3, there was an increase of 0.085 in CLAS scores between successive management layers (SE=0.019, p=0.000). For piece 2, which occurred between management layers 3 through 4, there was a decrease of 0.002 in CLAS scores between successive management layers, but this was not significant (SE=0.062, p=0.978). The residual variance in this model was significant ($s^2=0.034$, SE=0.004, p=0.000). The mean CLAS score relative to the role
complexity range for each management layer is presented in Table N7. A Bonferroni correction was applied for 9 comparisons (0.05/9 z-tests = 0.006).

**Table N 7**

*Fit Statistics for the Linear, Quadratic, Piecewise and Freed Models (LSUA)*

<table>
<thead>
<tr>
<th>Model</th>
<th>Management Layer 2</th>
<th>Management Layer 3</th>
<th>Management Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>171</td>
<td>252</td>
<td>30</td>
</tr>
<tr>
<td>M</td>
<td>11.172</td>
<td>11.257</td>
<td>11.256</td>
</tr>
<tr>
<td>s^2</td>
<td>0.040</td>
<td>0.029</td>
<td>0.044</td>
</tr>
<tr>
<td>min</td>
<td>10.460</td>
<td>10.460</td>
<td>10.630</td>
</tr>
<tr>
<td>max</td>
<td>11.710</td>
<td>11.710</td>
<td>11.630</td>
</tr>
<tr>
<td>Z-test on difference between mean and lower boundary</td>
<td>z=0.172, SE=0.016, p=0.000</td>
<td>z=0.057, SE=0.015, p=0.000</td>
<td>z=0.144, SE=0.064, p=0.024</td>
</tr>
<tr>
<td>Z-test on difference between mean and midpoint</td>
<td>z=-0.058, SE=0.016, p=0.000</td>
<td>z=-0.173, SE=0.015, p=0.000</td>
<td>z=-0.374, SE=0.064, p=0.000</td>
</tr>
<tr>
<td>Z-test on difference between mean and upper boundary</td>
<td>z=-0.278, SE=0.016, p=0.000</td>
<td>z=-0.393, SE=0.015, p=0.000</td>
<td>z=-0.594, SE=0.064, p=0.000</td>
</tr>
</tbody>
</table>

For management layer 2, the mean CLAS score was located between the lower boundary and midpoint. For management layer 3, the mean CLAS score was also located between the lower boundary and the midpoint. For management layer 4, the mean CLAS score was located at the lower boundary. To better understand whether CLAS scores were distributed as expected, a series of chi-square analyses were undertaken. For each management layer, the actual and expected distributions of CLAS scores relative to role complexity were compared. Descriptive statistics are presented in Table N8.
One-sample chi-square analyses were conducted on each management layer. Analyses demonstrated significant differences between the expected and actual distributions for management layer 2 ($\chi^2(3, n=171)=99.679$, $p=0.00$), management layer 3 ($\chi^2(3, n=252)=151.394$, $p=0.00$), and management layer 4 ($\chi^2(3, n=30)=325.179$, $p=0.00$).

In summary, there is insufficient evidence to suggest the presence of a complexity gap at management layer 2. The mean CLAS score is located between the lower boundary and the midpoint, and the actual distribution appears to be somewhat higher than the expected distribution, particularly between the midpoint and upper boundary. There appears to be sufficient evidence to suggest a shift towards a complexity gap at management layer 3. The mean CLAS score is also

---

**Table N 8**

*Comparison Between Expected and Actual Distribution of Scores (LSUA)*

<table>
<thead>
<tr>
<th></th>
<th>Management Layer 2</th>
<th>Management Layer 3</th>
<th>Management Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expected</td>
<td>Actual</td>
<td>Expected</td>
</tr>
<tr>
<td>Less than lower boundary</td>
<td>8.5 (5%)</td>
<td>34 (19.9%)</td>
<td>12.6 (5%)</td>
</tr>
<tr>
<td>Between lower boundary and midpoint</td>
<td>111.2 (65%)</td>
<td>68 (39.8%)</td>
<td>163.8 (65%)</td>
</tr>
<tr>
<td>Between midpoint and upper boundary</td>
<td>42.8 (25%)</td>
<td>59 (34.5%)</td>
<td>63.0 (25%)</td>
</tr>
<tr>
<td>Greater than upper boundary</td>
<td>8.5 (5%)</td>
<td>10 (5.8%)</td>
<td>12.6 (5%)</td>
</tr>
</tbody>
</table>
located between the lower boundary and midpoint, the percentage of CLAS scores below the lower boundary is somewhat greater than expected. There is sufficient evidence to suggest the beginning of a complexity gap at management layer 4. This is because the mean CLAS score is located at the lower boundary and the percentage of CLAS scores below the lower boundary is considerably greater than expected.

**Abbreviated results for Lectical Ethical Reasoning Assessment (LERA)**

The intraclass correlation (ICCs) exceeded the critical value of 0.05, but the design effect (Deffs) did not exceed the critical value of 2 (refer Table N9). An aggregated approach to multi-level modelling (MLM) was still applied with cohort as the clustering variable.

**Table N 9**

*Variability Attributable to Hierarchical Clustering in the Data Structure (LERA)*

<table>
<thead>
<tr>
<th>Clustering variable</th>
<th>ICCs for Mean CLAS score</th>
<th>Average cluster size (c)</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohort</td>
<td>0.080</td>
<td>11.118</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Fit statistics used to evaluate relative model fit were the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), sample-size adjusted BIC, and shared variance ($r^2$). Fit statistics for the linear, quadratic, piecewise (knot at management layer 3), and freed models are presented in Table N10.
Table N 10

*Fit Statistics for the Linear, Quadratic, Piecewise and Freed Models (LERA)*

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>AIC</th>
<th>BIC</th>
<th>Sample adjusted BIC</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>189</td>
<td>-131.761</td>
<td>-122.036</td>
<td>-131.538</td>
<td>0.066</td>
</tr>
<tr>
<td>Quadratic</td>
<td>189</td>
<td>-136.179</td>
<td>-123.212</td>
<td>-135.882</td>
<td>0.097</td>
</tr>
<tr>
<td>Piecewise</td>
<td>189</td>
<td>-136.179</td>
<td>-123.212</td>
<td>-135.882</td>
<td>0.097</td>
</tr>
<tr>
<td>Freed</td>
<td>189</td>
<td>-136.179</td>
<td>-123.212</td>
<td>-135.882</td>
<td>0.097</td>
</tr>
</tbody>
</table>

Statistical evidence could not distinguish between the quadratic, piecewise, or freed model.

The piecewise model appeared to be the best fit based on the confirmatory analyses which were undertaken with a Bonferroni correction (0.05/(2x2) = 0.0125). However, the smaller samples may have impacted these findings for more senior management layers. There was a significant increase of 0.118 from management layer 2 (n=61, M=11.276, $s^2=0.033$) to management layer 3 (n=111, M=11.394, $s^2=0.026$) according to the Wald Test (W=37.342, df=1, p=0.0000) and z-test (SE=0.019, p=0.000). There was a decrease of 0.023 from management layer 3 (n=111, M=11.394, $s^2=0.026$) to management layer 4 (n=17, M=11.372, $s^2=0.018$) but this was not significant according to the Wald Test (W=0.370, df=1, p=0.5432) and z-test (SE=0.037, p=0.543). As a result, the piecewise model was selected as the best fitting model.

For the piecewise model (n=189) with a knot at management layer 3, leaders at management layer 2 had a mean CLAS score of 11.276 (SE=0.023, p=0.000). For piece 1, which occurred between management layers 2 through 3, there was an increase in of 0.118 in CLAS scores between successive management layers (SE=0.019, p=0.000). For piece 2, which occurred between management layers 3 through 4, there was a decrease of 0.023 in CLAS scores between successive management layers, but this was not significant (SE=0.037, p=0.543). The residual variance in this model was significant ($s^2_r=0.027$, SE=0.003, p=0.000). The mean CLAS score relative to the role...
complexity range for each management layer is presented in Table N11. A Bonferroni correction was applied for 9 comparisons (0.05/9 z-tests = 0.006).

**Table N 11**

*Descriptive Statistics by Management Layer (LERA)*

<table>
<thead>
<tr>
<th>Model</th>
<th>Management layer 2</th>
<th>Management layer 3</th>
<th>Management layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>61</td>
<td>111</td>
<td>17</td>
</tr>
<tr>
<td>M</td>
<td>11.276</td>
<td>11.394</td>
<td>11.372</td>
</tr>
<tr>
<td>$s^2$</td>
<td>0.033</td>
<td>0.026</td>
<td>0.018</td>
</tr>
<tr>
<td>min</td>
<td>10.630</td>
<td>11.040</td>
<td>10.960</td>
</tr>
<tr>
<td>max</td>
<td>11.630</td>
<td>11.710</td>
<td>11.540</td>
</tr>
<tr>
<td>Z-test on</td>
<td>z=0.276, SE=0.023</td>
<td>z=0.194, SE=0.017</td>
<td>z=-0.028, SE=0.038</td>
</tr>
<tr>
<td>difference from lower boundary</td>
<td>p=0.000</td>
<td>p=0.000</td>
<td>p=0.460</td>
</tr>
<tr>
<td>Z-test on</td>
<td>z=0.046, SE=0.023</td>
<td>z=-0.036, SE=0.017</td>
<td>z=-0.258, SE=0.038</td>
</tr>
<tr>
<td>difference from midpoint</td>
<td>p=0.042</td>
<td>p=0.032</td>
<td>p=0.000</td>
</tr>
<tr>
<td>Z-test on</td>
<td>z=-0.174, SE=0.023</td>
<td>z=-0.256, SE=0.017</td>
<td>z=-0.478, SE=0.038</td>
</tr>
<tr>
<td>difference from upper boundary</td>
<td>p=0.000</td>
<td>p=0.000</td>
<td>p=0.000</td>
</tr>
</tbody>
</table>

For management layer 2, the mean CLAS score was located at the midpoint. For management layer 3, the mean CLAS score was also located at the midpoint. For management layer 4, the mean CLAS score was located at the lower boundary. To better understand whether LERA CLAS scores were distributed as expected, a series of chi-square analyses were undertaken. For each management layer, the actual and expected distributions of CLAS scores relative to role complexity were compared. Descriptive statistics are presented in Table N12.
Table N 12

Comparison Between Expected and Actual Distribution of Scores (LERA)

<table>
<thead>
<tr>
<th></th>
<th>Management Layer 2</th>
<th>Management Layer 3</th>
<th>Management Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=61</td>
<td>n=111</td>
<td>n=17</td>
</tr>
<tr>
<td>Less than lower</td>
<td>Expected</td>
<td>Actual</td>
<td>Expected</td>
</tr>
<tr>
<td>boundary</td>
<td>3.0 (5%)</td>
<td>5 (8.2%)</td>
<td>5.5 (5%)</td>
</tr>
<tr>
<td>Between lower</td>
<td>39.7 (65%)</td>
<td>19 (31.1%)</td>
<td>72.2 (65%)</td>
</tr>
<tr>
<td>boundary and midpoint</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between midpoint</td>
<td>15.3 (25%)</td>
<td>24 (39.3%)</td>
<td>27.8 (25%)</td>
</tr>
<tr>
<td>and upper boundary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greater than upper</td>
<td>3.0 (5%)</td>
<td>13 (21.3%)</td>
<td>5.5 (5%)</td>
</tr>
<tr>
<td>boundary</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One sample chi-square analyses were conducted on each management layer. Analyses demonstrated significant differences between the expected and actual distributions of CLAS scores for management layer 2 ($\chi^2(3, n=61)=50.407, p=0.00$), management layer 3 ($\chi^2(3, n=111)=16.268, p=0.00$), and management layer 4 ($\chi^2(3, n=17)=90.016, p=0.00$).

In summary, there is insufficient evidence to suggest the presence of a complexity gap at management layer 2. The mean CLAS score is located at the midpoint, and the actual distribution appears to be somewhat higher than the expected distribution, particularly between the midpoint and upper boundary, and above the upper boundary. There is insufficient evidence to suggest a shift towards a complexity gap at management layer 3. The mean CLAS score is also located midpoint, and the distribution of actual CLAS scores appears to represent relatively similar patterns to the expected distribution. There is sufficient evidence to suggest the beginning of a complexity gap at
management layer 4. This is because the mean CLAS score occurs at the lower boundary and the percentage of CLAS scores below the lower boundary is considerably greater than expected.

**Abbreviated results for Lectical Reflective Judgement Assessment (LRJA)**

The intraclass correlation (ICCs) did not exceed the critical value of 0.05, and the design effect (Deffs) did not exceed the critical value of 2 (refer Table N13). As a result, a multi-level approach to modelling (MLM) was not required.

**Table N 13**

*Variability Attributable to Hierarchical Clustering in the Data Structure (LRJA)*

<table>
<thead>
<tr>
<th>Clustering variable</th>
<th>ICCs for Mean CLAS score</th>
<th>Average cluster size (c)</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohort</td>
<td>0.021</td>
<td>12.867</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Fit statistics used to evaluate relative model fit were the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), sample-size adjusted BIC, and shared variance ($r^2$). Fit statistics for the linear, quadratic, piecewise (knot at management layer 3), and freed models are presented in Table N14.
Table N 14

Fit Statistics for the Linear, Quadratic, Piecewise and Freed Models (LRJA)

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>AIC</th>
<th>BIC</th>
<th>Sample adjusted BIC</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>193</td>
<td>-66.402</td>
<td>-56.614</td>
<td>-66.117</td>
<td>0.054</td>
</tr>
<tr>
<td>Quadratic</td>
<td>193</td>
<td>-65.986</td>
<td>-52.935</td>
<td>-65.606</td>
<td>0.061</td>
</tr>
<tr>
<td>Piecewise</td>
<td>193</td>
<td>-65.986</td>
<td>-52.935</td>
<td>-65.606</td>
<td>0.061</td>
</tr>
<tr>
<td>Freed</td>
<td>193</td>
<td>-65.986</td>
<td>-52.935</td>
<td>-65.606</td>
<td>0.040</td>
</tr>
</tbody>
</table>

Statistical evidence could not distinguish between the quadratic, piecewise, or freed model. The piecewise model appeared to be the best fit based on the confirmatory analyses which were undertaken with a Bonferroni correction (0.05/(2x2) = 0.0125). However, the smaller samples may have impacted these findings for more senior management layers. There was a significant increase of 0.105 from management layer 2 (n=72, M=11.156, $s^2=0.046$) to management layer 3 (n=110, M=11.261, $s^2=0.039$) according to the Wald Test ($W=11.140$, $df=1$, $p=0.0008$) and z-test ($SE=0.031$, $p=0.001$). There was an increase of 0.010 from management layer 3 (n=110, M=11.261, $s^2=0.039$) to management layer 4 (n=11, M=11.271, $s^2=0.017$) but this was not significant according to the Wald Test ($W=0.054$, $df=1$, $p=0.817$) and z-test ($SE=0.043$, $p=0.817$). As a result, the piecewise model was selected as the best fitting model.

For the piecewise model (n=193) with a knot at management layer 3, leaders at management layer 2 had a mean CLAS score of 11.156 ($SE=0.024$, $p=0.000$). For piece 1, which occurred between management layers 2 through 3, there was an increase in of 0.105 in CLAS scores between successive management layers ($SE=0.030$, $p=0.001$). For piece 2, which occurred between management layers 3 through 4, there was an increase of 0.010 in CLAS scores between successive management layers, but this was not significant ($SE=0.063$, $p=0.874$). The residual variance in this
model was significant ($s^2=0.040$, SE=0.004, $p=0.000$). The mean CLAS score relative to the role complexity range for each management layer is presented in Table N15. A Bonferroni correction was applied for 9 comparisons ($0.05/9$ z-tests = 0.006).

**Table N 15 Descriptive Statistics by Management Layer (LRJA)**

<table>
<thead>
<tr>
<th></th>
<th>Management layer 2</th>
<th>Management layer 3</th>
<th>Management layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>72</td>
<td>110</td>
<td>11</td>
</tr>
<tr>
<td>M</td>
<td>11.156</td>
<td>11.261</td>
<td>11.271</td>
</tr>
<tr>
<td>$s^2$</td>
<td>0.046</td>
<td>0.039</td>
<td>0.017</td>
</tr>
<tr>
<td>min</td>
<td>10.460</td>
<td>10.630</td>
<td>10.960</td>
</tr>
<tr>
<td>max</td>
<td>11.630</td>
<td>11.710</td>
<td>11.460</td>
</tr>
<tr>
<td>Z-test on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-test on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>difference from</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-test on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lower boundary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>difference from</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>upper boundary</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For management layer 2, the mean CLAS score occurred between the lower boundary and the midpoint. For management layer 3, the mean CLAS score also occurred between the lower boundary and the midpoint. For management layer 4, the mean CLAS score occurred below the lower boundary. To better understand whether CLAS scores were distributed as expected, a series of chi-square analyses were undertaken. For each management layer, the actual and expected distributions of CLAS scores were compared. Descriptive statistics are presented in Table N16.
Table N 16

Comparison Between Expected and Actual Distribution of Scores (LRJA)

<table>
<thead>
<tr>
<th>Management Layer 2</th>
<th>Management Layer 3</th>
<th>Management Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expected</td>
<td>Actual</td>
</tr>
<tr>
<td>n=72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than lower boundary</td>
<td>3.6 (5%)</td>
<td>17 (23.6%)</td>
</tr>
<tr>
<td>Between lower boundary and midpoint</td>
<td>46.8 (65%)</td>
<td>25 (34.7%)</td>
</tr>
<tr>
<td>Between midpoint and upper boundary</td>
<td>18 (25%)</td>
<td>27 (37.5%)</td>
</tr>
<tr>
<td>Greater than upper boundary</td>
<td>3.6 (5%)</td>
<td>3 (4.2%)</td>
</tr>
</tbody>
</table>

One sample chi-square analyses were conducted on each management layer. Analyses demonstrated significant differences between the expected and actual distributions of CLAS scores for management layer 2 ($\chi^2(3, n=72)=64.632, p=0.00$), management layer 3 ($\chi^2(3, n=110)=134.671, p=0.00$), and management layer 4 ($\chi^2(3, n=11)=189.139, p=0.00$).

In summary, there is some evidence to suggest a shift towards a complexity gap in management layer 2. The mean CLAS score is located between the lower boundary and the midpoint, and the percentage of CLAS scores below the lower boundary is somewhat greater than expected. There is also evidence to suggest a shift towards a complexity gap in management layer 3. The mean CLAS score is located between the lower boundary and the midpoint, and the percentage of CLAS scores below the lower boundary is somewhat greater than expected. There is sufficient
evidence to suggest the beginning of a complexity gap at management layer 4. The mean CLAS score occurs below the lower boundary, and the percentage of CLAS scores below the lower boundary is considerably greater than expected.
Appendix O: Syntax for all Analyses Presented in Chapter 5

TITLE: Preliminary analysis on ICCs

DATA: FILE IS Long.dat;

VARIABLE: NAMES ARE
userkey !unique ID for each leader/test-taker
birth !birthyear
lang !first language code
geo !geography code
grade !number of years of education
sector !sector code
sex !sex code
ethnic !ethnicity code
manlo !original management layer
manlc !collapsed management layer
manla !adjusted management layer
provider !leadership provider
programc !leadership program code
cohortc !leadership cohort code
instru !type of assessment e.g., LDMA
dtkey1 !unique ID for assessment at time 1
lakey1 !another unique ID for assessment at time 1
clas1 !clas score at time 1
wc1 !word count at time 1
uwc1 !unique word count at time 1
days1 !number of days since time 1 assessment (this is 0 at time 1)
weeks1 !number of weeks since time 1 assessment (this is 0 at time 1)
months1 !number of months since time 1 assessment (this is 0 at time 1)
dtkey2 !unique ID for assessment at time 2
lakey2 !another unique ID for assessment at time 2
clas2 !clas score at time 2
wc2 !word count at time 2
uwc2 !unique word count at time 2
days2 !number of days since time 1 assessment
w2i !number of weeks since time 1 assessment (integer)
m2i !number of months since time 1 assessment (integer)
w2d !number of weeks since time 1 assessment (decimal)
m2d !number of months since time 1 assessment (decimal)
dtkey3 !unique ID for assessment at time 3
lakey3 !another unique ID for assessment at time 3
clas3 !clas score at time 3
wc3 !word count at time 3
uwc3 !unique word count at time 3
days3 !number of days since time 1 assessment
w3i !number of weeks since time 1 assessment (integer)
m3i !number of months since time 1 assessment (integer)
w3d !number of weeks since time 1 assessment (decimal)
m3d !number of months since time 1 assessment (decimal)
dtkey4 !unique ID for assessment at time 4
lakey4 !another unique ID for assessment at time 4
clas4 !clas score at time 4
wc4 !word count at time 4
uwc4 !unique word count at time 4
days4 !number of days since time 1 assessment
w4i !number of weeks since time 1 assessment (integer)
m4i !number of months since time 1 assessment (integer)
w4d !number of weeks since time 1 assessment (decimal)
m4d !number of months since time 1 assessment (decimal)
umass !total assessments taken by each leader/test-taker
months !total duration of leadership intervention in months
hours !total effort of leadership intervention in months
hpm !average number of hours per month for leadership intervention
fs !Lectical Assessments used formative or summative manner

USEVARIABLES ARE
    clas1
    clas2
    clas3
    clas4

MISSING ARE all (-999);

USEOBSERVATIONS ARE
    provider /= 2 AND
    cohortc /= 36 AND
    cohortc /= 37;

    CLUSTER IS
        provider;
        programc;
        cohortc;

ANALYSIS: TYPE IS TWOLEVEL BASIC;

OUTPUT: SAMPSTAT
    MOD(3.84)
    STDYX
    TECH1;

TITLE: Longitudinal growth modelling - test fit of variety of models that successively deviate from linearity

DATA: FILE IS Long.dat;

VARIABLE: NAMES ARE !All variables defined above
userkey
birth
lang
go
grade
sector
sex
ethnic
manlo
manlc
manla
provider
program
cohort
programc
cohortc
instru
dtkkey1
lakey1
clas1
wc1
uwc1
days1
weeks1
months1
dtkkey2
lakey2
clas2
wc2
uwc2
days2
w2i
m2i
w2d
m2d
dtkkey3
lakey3
clas3
wc3
uwc3
days3
w3i
m3i
w3d
m3d
dtkkey4
lakey4
clas4
wc4
uwc4
days4
USEVARIABLES ARE
clas1
clas2
clas3
clas4
days1
weeks1
months1
days2
w2i
m2i
w2d
m2d
days3
w3i
m3i
w3d
m3d
days4
w4i
m4i
w4d
m4d
numass
hpm
fs
provider
programc
cohortc
clas1_s
clas2_s
clas3_s
clas4_s
years1
years2
years3
years4;

TSCORES ARE
MISSING ARE all (-999);

SUBPOPULATION ARE
instru==6 AND
instru==1 AND
provider /=2 AND
cohortc /= 36 AND
cohortc /= 37;

provider==5;

provider /=8

CLUSTER IS
provider;
programc;
cohortc;

ANALYSIS: TYPE IS COMPLEX; !for non time-varying analyses
TYPE IS COMPLEX RANDOM; !for time-varying analysis only
TYPE IS RANDOM;
PROCESSORS = 2;
STARTS=800;

MODEL:        !i | clas1 clas2 clas3 clas4; !intercept model
        !i s | clas1@0 clas2@1 clas3@2 clas4@3; !linear model
        !i s q| clas1@0 clas2@1 clas3@2 clas4@3; !quadratic model
is s1 | clas1@0 clas2@1 clas3@1 clas4@1; !piecewise model piece 1
is s2 | clas1@0 clas2@0 clas3@1 clas4@2; !piecewise model piece 2
i s | clas1 clas2 clas3 clas4 AT months1 m2d m3d m4d; !time varying model
is s | clas1 clas2 AT months1 m2d; !time1 vs time2
is s | clas2 clas3 AT m2d m3d; !time2 vs time3
is s | clas3 clas4 AT m3d m4d; !time3 vs time4
is ON numass hpm fs; !covariates of rate of growth

OUTPUT: SAMPSTAT
MOD(3.84)
STDYX
TECH1;

TITLE: Longitudinal growth modelling – testing covariates of rate of growth in CLAS scores

DATA: FILE IS Long.dat;

VARIABLE: NAMES ARE !All variables are defined above
    userkey
    birth
    lang
    geo
    grade
    sector
    sex
    ethnic
    manlo
    manlc
    manla
    provider
    program
    cohort
    programc
    cohortc
    instru
    dtkey1
    lakey1
    clas1
    wc1
    uwc1
    days1
    weeks1
    months1
dtkey2
    lakey2
    clas2
wc2
uwc2
days2
w2i
m2i
w2d
m2d
dtkey3
lakey3
clas3
wc3
uwc3
days3
w3i
m3i
w3d
m3d
dtkey4
lakey4
clas4
wc4
uwc4
days4
w4i
m4i
w4d
m4d
numass
months
hours
hpm
fs
;

USEVARIABLES ARE
clas1
clas2
clas3
clas4
days1
weeks1
months1
days2
w2i
m2i
w2d
m2d
days3
w3i
m3i
w3d
m3d
days4
w4i
m4i
w4d
m4d
numass
hpm
fs
manlc
provider
programc
cohortc
clas1_s
clas2_s
clas3_s
clas4_s
years1
years2
years3
years4
;

TSCORES ARE
days1
weeks1
months1
days2
w2i
m2i
w2d
m2d
days3
w3i
m3i
w3d
m3d
days4
w4i
m4i
w4d
m4d
years1
years2
years3
years4
;

MISSING ARE all (-999);

SUBPOPULATION ARE
INSTRU==1; 
PROVIDER==5; 
PROVIDER /=8; 
PROVIDER /=2 AND 
COHORTC /= 36 AND 
COHORTC /= 37 
;

CLUSTER IS 
PROVIDER; 
PROGRAMC; 
COHORTC;

!DEFINE: CENTER numass hpm fs (GRANDMEAN);

ANALYSIS: TYPE IS COMPLEX RANDOM; 
STARTS=50;

MODEL: 
  i s | clas1 clas2 clas3 clas4 AT months1 m2d m3d m4d; 
  s ON numass hpm fs; 
  [i s];

OUTPUT: SAMPSTAT;

TITLE: Longitudinal growth modelling – comparing rates of growth in CLAS scores between different leadership providers

DATA: FILE IS Long.dat;

VARIABLE: NAMES ARE !All variables are defined above 
  userkey 
  birth 
  lang 
  geo 
  grade 
  sector 
  sex 
  ethnic 
  manlo 
  manlc 
  manla 
  provider 
  program 
  cohort 
  programc 
  cohortc 
  instru 
  dtkey1 
  lakey1
USEVARIABLES ARE
clas1 clas2 clas3 clas4 days1 weeks1 months1

TSCORES ARE
years2
years3
years4;

MISSING ARE all (-999);

USEOBSERVATIONS ARE
cohortc /= 36 AND
cohortc /= 37;

GROUPING IS
  provider (  
  1=Provider 1  
  2=Provider 2  
  3=Provider 3  
  4=Provider 4  
  5=Provider 5  
  6=Provider 6  
  7=Provider 7  
  8=Provider 8  
  9=Provider 9  
  10=Provider 10  
  11=Provider 11  
  12=Provider 12  
);

CLUSTER IS
  provider;
  programc;
  cohortc;

MODEL Provider 5:
  [s] (s1g1);
  [i] (i1g1);

MODEL Provider 12:
  [s] (s2g2);
  [i] (i2g2);

MODEL CONSTRAINT:
  NEW(a);
  a=i1g1-i2g2;

NEW(b);
  b=s1g1-s2g2;

ANALYSIS: TYPE IS COMPLEX RANDOM;

MODEL:
  i s  | clas1 clas2 AT months1 m2d;
  s@0;
OUTPUT: SAMPSTAT;

TITLE: Longitudinal growth modelling – testing for moderators of rates of growth in CLAS scores

DATA: FILE IS Long.dat;

VARIABLE: NAMES ARE !All variables are defined above 
  userkey
  birth
  lang
  geo
  grade
  sector
  sex
  ethnic
  manlo
  manlc
  manla
  provider
  program
  cohort
  programc
  cohortc
  instru
  dtkey1
  lakey1
  clas1
  wc1
  uwc1
  days1
  weeks1
  months1
  dtkey2
  lakey2
  clas2
  wc2
  uwc2
  days2
  w2i
  m2i
  w2d
  m2d
  dtkey3
  lakey3
  clas3
  wc3
  uwc3
  days3
  w3i
USEVARIABLES ARE
clas1 clas2 clas3 clas4 days1 weeks1 months1
days2 w2i m2i w2d m2d days3 w3i m3i
w3d m3d days4 w4i m4i w4d m4d numass
hpm fs provider programc cohortc
clas1_s clas2_s
clas3_s
clas4_s
years1
years2
years3
years4
sexN !define new variable to test for moderating effect refer below
sectorN !define new variable to test for moderating effect refer below
langN !define new variable to test for moderating effect refer below
;
TSCORES ARE
days1
weeks1
months1
days2
w2i
m2i
w2d
m2d
days3
w3i
m3i
w3d
m3d
days4
w4i
m4i
w4d
m4d
years1
years2
years3
years4
;
MISSING ARE all (-999);

SUBPOPULATION ARE
provider /=2 AND
cohortc /= 36 AND
cohortc /= 37;

CLUSTER IS
provider;
programc;
cohortc;

DEFINE: sexN=(sex-1); !0=male 1=female
sectorN=(sector-1); !0=public 1=private
langN=(lang-1); !0=english 1=nonenglish
ANALYSIS: TYPE IS COMPLEX RANDOM; !for time-varying analysis only
STARTS=800;

MODEL:
   i s | clas1 clas2 clas3 clas4 AT months1 m2d m3d m4d;
   i s ON sexN;
   i s ON sectorN;
   i s ON langN;

OUTPUT: SAMPSTAT;