


1-1-2011

Improving the Detection of Narcissistic Transformational Leaders with the Multifactor Leadership Questionnaire: An Item Response Theory Analysis

Dale Frederick Hosking Martin
Walden University

Follow this and additional works at: <https://scholarworks.waldenu.edu/dissertations>

 Part of the [Business Administration, Management, and Operations Commons](#), [Management Sciences and Quantitative Methods Commons](#), [Quantitative Psychology Commons](#), and the [Vocational Rehabilitation Counseling Commons](#)

This Dissertation is brought to you for free and open access by the Walden Dissertations and Doctoral Studies Collection at ScholarWorks. It has been accepted for inclusion in Walden Dissertations and Doctoral Studies by an authorized administrator of ScholarWorks. For more information, please contact ScholarWorks@waldenu.edu.

Walden University

COLLEGE OF SOCIAL AND BEHAVIORAL SCIENCES

This is to certify that the doctoral dissertation by

Dale Martin

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee

Dr. Tom Diebold, Committee Chairperson, Psychology Faculty

Dr. Scott Davies, Committee Member, Psychology Faculty

Dr. Gwynne Dawdy, Committee Member, Psychology Faculty

Dr. George Smeaton, University Reviewer, Psychology Faculty

Chief Academic Officer

David Clinefelter, Ph.D.

Walden University
2010

Abstract

Improving the Detection of Narcissistic Transformational Leaders with the Multifactor

Leadership Questionnaire: An Item Response Theory Analysis

by

Dale Frederick Hosking Martin

M.B.A., University of St. Thomas, 1986

B.S., Washington University, 1982

B.A., Gustavus Adolphus College, 1981

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Psychology

Walden University

November 2010

Abstract

Narcissistic transformation leaders have inflicted severe physical, psychological, and financial damage on individuals, institutions, and society. Multifactor Leadership Questionnaire (MLQ) has shown promise for early detection of narcissistic leadership tendencies, but selection criteria have not been established. The purpose of this quantitative research was to determine if item response theory (IRT) could advance the detection of narcissistic leadership tendencies using an item-level analysis of the 20 transformational leadership items of the MLQ. Three archival samples of subordinates from Israeli corporate and athletic organizations were combined ($N = 1,703$) to assess IRT data assumptions, comparative fit of competing IRT models, item discrimination and difficulty, and theta reliabilities within the trait range. Compared to the generalized graded unfolding model, the graded response model had slightly more category points within the 95% confidence interval and consistently lower X^2/df item fit indices. Items tended to be easier yet more discriminating than average, and five items were identified as candidates for modification. IRT item marginal reliability was .94 (slightly better than classical test theory reliability of .93), and IRT ability prediction had a .96 reliability within a trait range from -1.7 to 1.3 theta. Based on 8 invariant item parameters, selection criteria of category *fairly often* (3) or above on attributed idealized influence items and *sometimes* (2) or below on individual consideration items was suggested. A test case demonstrated how narcissistic tendencies could be detected with these criteria. The study can contribute to positive social change by informing improved selection processes that more effectively screen candidates for key leadership roles that directly impact the wellbeing of individuals and organizations.

Improving the Detection of Narcissistic Transformational Leaders with the Multifactor

Leadership Questionnaire: An Item Response Theory Analysis

by

Dale Frederick Hosking Martin

M.B.A., University of St. Thomas, 1986

B.S., Washington University, 1982

B.A., Gustavus Adolphus College, 1981

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Organizational Psychology

Walden University

November 2010

UMI Number: 3434575

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3434575

Copyright 2011 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Dedication

For my dad, Dr. Albert F. Martin, from crop farmer to renowned polymer chemist, he shared his enthusiastic global pursuit of knowledge. To my mom, Pat, she had a loving heart, full of grace. In memory of my brother, Paul and his family, who had a zest for living and befriending everyone they met. To my brother, Earl, he has always given selflessly so that others may prosper. For my friend and wife, Betsy, I am grateful for your encouragement, love, and steadfast support. Finally, to my son, Andy, and my daughter, Kelsey, for putting up with my long hours and short fuse.

Acknowledgments

I would like to thank several people that have substantially contributed to completion of this dissertation. I especially appreciate and value my long time friend, Dr. Carlos Mendes de Leon; he has been a source of inspiration and practical suggestions from refining the topic to critical statistical direction. His ability to quickly simplify complex concepts expedited my progress. My gratitude to Dr. Yair Berson, not only for providing the welcomed archival data used in this study but willingly offering encouragement, friendship, and guidance along the way. To Dr. Michael Edwards of Ohio State University, whose IRT instruction and personal support was critical in data analysis.

I wish to thank my chairperson, Dr. Charles “Tom” Diebold, for keeping me focused and encouraged. To committee member, Dr. Scott Davies, his IRT knowledge and dedication produced a stronger dissertation. It was committee member, Dr. Gwynne Dawdy, with her practitioner’s perspective that inspired my social change awareness. University reviewer, Dr. George Smeaton, provided invaluable improvement suggestions. Finally and gratefully, thanks to my wife, Betsy, for the countless hours of proof reading and refinement.

Table of Contents

List of Tables	vii
List of Figures	viii
List of Abbreviations	xi
Chapter 1: Introduction to the Study.....	1
Statement of the Problem.....	7
Nature of the Study	8
Research Questions	9
Research Objectives.....	10
Research Hypothesis.....	10
Purpose of the Study	11
Research Design.....	11
Theoretical Framework.....	15
Operational Definitions of Terms	17
Assumptions, Limitations, Scope, and Delimitations of the Study	22
Significance of the Study	27
Summary.....	29
Chapter 2: Literature Review	33
Introduction.....	33
Organization of Literature Review	34
Literature Search Strategy.....	34
Gap in Current Research.....	35

Classical Test Theory.....	36
Reliability Measures	37
Validity Measures	39
Item Response Theory	42
IRT Models	45
Conceptual Basis for GRM.....	47
Conceptual Basis for GGUM.....	51
IRT Parameters	53
Assumptions of IRT	55
Questionnaire Development and Refinement Using IRT	58
Summary Benefits and Limitations of IRT over Classical Test Theory.....	61
Leadership.....	62
Leadership Questionnaires.....	65
Multifactor Leadership Questionnaire	69
Transformational Leadership Theory	70
Transformational Leadership	73
Transactional Leadership	78
Laissez-Faire Leadership	83
The MLQ Development.....	84
The MLQ Psychometric Properties	89
The MLQ Summary.....	99
Methodology Considerations	100

Instrument Format and Contact Mode	101
Participant Characteristics	102
IRT Model Optimization Steps	107
Software	108
Literature Review Summary	109
Chapter 3: Research Method.....	112
Research Design and Approach	113
Justification of Design and Approach.....	115
Samples and Procedures	116
Instrument and Analytical Software	119
Data Preparation and Analysis.....	120
Data Preparation.....	120
Assumption Testing	121
Five Research Questions.....	124
Ethical Protections	127
Summary	128
Chapter 4: Results.....	130
Data Preparation.....	130
An Unanticipated Limitation	131
Assumption Testing	132
Assumption 1: Translation Accuracy.....	132
Assumption 2: Independent Observations	133

Assumption 3: Sufficient Category Responses.....	133
Assumption 4: Normal Distribution.....	134
Assumption 5: Unidimensionality	137
Assumption 6: Sample Homogeneity	145
Research Question 1: Observed Versus Expected IRT Model Responses	148
Research Question 2: Best IRT Model	152
Research Question 3: Discrimination and Difficulty Parameter Estimates	156
Research Question 4: Highest Trait Range Reliability Estimation.....	160
Research Question 5: IRT Versus Classical Test Theory Reliability	164
Summary	171
Chapter 5: Discussion, Conclusions, and Recommendations.....	176
Overview.....	176
Interpretation of Findings	178
Data Preparation.....	178
An Unanticipated Limitation	178
Assumption Testing	179
Research Question 1: Observed Versus Expected IRT Model Responses	181
Research Question 2: Best IRT Model	181
Research Question 3: Discrimination and Difficulty Parameter Estimates	183
Research Question 4: Highest Trait Range Reliability Estimation.....	185
Research Question 5: IRT Versus Classical Test Theory Reliability	186
Additional Finding Interpretation	187

Implications for Social Change.....	188
Recommendations for Action	192
Action Recommendation 1: Disseminate Results.....	192
Action Recommendation 2: Integrate IRT in the MLQ’s Research	192
Action Recommendation 3: Integrate IRT in Psychology Classes.....	193
Action Recommendation 4: Improve Transformational Leader Assessment	193
Action Recommendation 5: Improve Organizational Responsibility.....	194
Recommendations for Further Study	195
Future Study Recommendation 1: Extend Khoo and Burch (2007) Study.....	195
Future Study Recommendation 2: Modify Five MLQ Items.....	196
Future Study Recommendation 3: Extend This Study.....	197
Future Study Recommendation 4: Connect This Study to Derailment.....	198
Future Study Recommendation 5: Improve Leaders’ Response Rates.....	198
Conclusions.....	199
References.....	202
Appendix A: The GRM estimates without and with basketball players.....	230
Appendix B: The GRM and the GGUM fit metrics without basketball players.	231
Appendix C: The GGUM parameter estimates.....	232
Appendix D: The GRM graphs for idealized influence attributed items.....	233
Appendix E: The GRM graphs for idealized influence behavioral items.....	234
Appendix F: The GRM graphs for inspirational motivation items.....	235
Appendix G: The GRM graphs for intellectual stimulation items.....	236

Appendix H: The GRM graphs for individual consideration items.....	237
Appendix I: The GGUM graphs for idealized influence attributed items.	238
Appendix J: The GGUM graphs for idealized influence behavioral items.	239
Appendix K: The GGUM graphs for inspirational motivation items.	240
Appendix L: The GGUM graphs for intellectual stimulation items.	241
Appendix M: The GGUM graphs for individual consideration items.....	242
Appendix N: The GGUM test information function, standard error, and reliability.....	243
Curriculum Vitae	244

List of Tables

Table 1. Item distribution measures.....	136
Table 2. Item-item correlations of 20 study items	138
Table 3. Corrected item-total correlations of 20 study items.....	139
Table 4. Eigenvalues and percent of variance for exploratory factor analysis	140
Table 5. Factor loading for exploratory factor analysis	141
Table 6. Goodness of fit for exploratory factor analysis	142
Table 7. Factor correlation matrix for exploratory factor analysis	143
Table 8. Item and total score M and SD for each sample	146
Table 9. Theta M and SD for each sample	147
Table 10. The GRM and the GGUM item fit metrics.....	151
Table 11. The GRM and the GGUM fit metrics for item comparisons.....	154
Table 12. The GRM and the GGUM adjusted fit metrics.....	155
Table 13. The GRM parameter estimates	157
Table 14. Classical test theory item analysis	165
Table 15. Classical test theory and IRT person abilities.....	171
Table A. The GRM estimates without and with basketball players	230
Table B1. The GRM and the GGUM fit metrics for corporate samples.....	231
Table B2. The GRM and the GGUM adjusted fit metrics for corporate samples	231
Table C. The GGUM parameter estimates	232

List of Figures

Figure 1. Full range leadership model	72
Figure 2. Transformational leadership model	145
Figure 3. The GRM and the GGUM fit plots.....	149
Figure 4. The GRM and the GGUM expected response functions.....	150
Figure 5. The GRM item parameter estimates.....	159
Figure 6. The GRM item characteristic curves and IIF for mlq14	162
Figure 7. The GRM test information function	163
Figure 8. Classical test theory versus IRT person estimates.....	169
Figure 9. A possible narcissistic detection example	191
Figure D1. The GRM mlq10 characteristic curves and IIF	233
Figure D2. The GRM mlq18 characteristic curves and IIF	233
Figure D3. The GRM mlq21 characteristic curves and IIF	233
Figure D4. The GRM mlq25 characteristic curves and IIF	232
Figure E1. The GRM mlq6 characteristic curves and IIF.....	234
Figure E2. The GRM mlq14 characteristic curves and IIF.....	234
Figure E3. The GRM mlq23 characteristic curves and IIF.....	234
Figure E4. The GRM mlq34 characteristic curves and IIF.....	234
Figure F1. The GRM mlq9 characteristic curves and IIF.....	235
Figure F2. The GRM mlq13 characteristic curves and IIF	235
Figure F3. The GRM mlq26 characteristic curves and IIF.....	235
Figure F4. The GRM mlq36 characteristic curves and IIF	235

Figure G1. The GRM mlq2 characteristic curves and IIF	236
Figure G2. The GRM mlq8 characteristic curves and IIF	236
Figure G3. The GRM mlq30 characteristic curves and IIF	236
Figure G4. The GRM mlq32 characteristic curves and IIF	236
Figure H1. The GRM mlq15 characteristic curves and IIF	237
Figure H2. The GRM mlq19 characteristic curves and IIF	237
Figure H3. The GRM mlq29 characteristic curves and IIF	237
Figure H4. The GRM mlq31 characteristic curves and IIF	237
Figure I1. The GGUM mlq10 characteristic curves and IIF.....	238
Figure I2. The GGUM mlq18 characteristic curves and IIF.....	238
Figure I3. The GGUM mlq21 characteristic curves and IIF.....	238
Figure I4. The GGUM mlq25 characteristic curves and IIF.....	238
Figure J1. The GGUM mlq6 characteristic curves and IIF	239
Figure J2. The GGUM mlq14 characteristic curves and IIF	239
Figure J3. The GGUM mlq23 characteristic curves and IIF	239
Figure J4. The GGUM mlq34 characteristic curves and IIF	239
Figure K1. The GGUM mlq9 characteristic curves and IIF	240
Figure K2. The GGUM mlq13 characteristic curves and IIF	240
Figure K3. The GGUM mlq26 characteristic curves and IIF	240
Figure K4. The GGUM mlq36 characteristic curves and IIF	240
Figure L1. The GGUM mlq2 characteristic curves and IIF.....	241
Figure L2. The GGUM mlq8 characteristic curves and IIF.....	241

Figure L3. The GGUM mlq30 characteristic curves and IIF	241
Figure L4. The GGUM mlq32 characteristic curves and IIF	241
Figure M1. The GGUM mlq15 characteristic curves and IIF	242
Figure M2. The GGUM mlq19 characteristic curves and IIF	242
Figure M3. The GGUM mlq29 characteristic curves and IIF	242
Figure M4. The GGUM mlq31 characteristic curves and IIF	242
Figure N. The GGUM test information function, standard error, and reliability	243

List of Abbreviations

GGUM: Generalized graded unfolding model

GRM: Graded response model

ICCC: Interclass correlation coefficient

IIF: Item Information Function

IRT: Item response theory

MLQ: Multifactor Leadership Questionnaire

Chapter 1: Introduction to the Study

Narcissistic transformational leaders can be very destructive. Narcissistic leaders such as Adolph Hitler of Germany and Slobodan Milošević of Serbia are perceived as having been transformational and capable of abject cruelty (Rosenthal & Pittinsky, 2006; Volkan & Fowler, 2009). Post (2008) reviewed the transformational leadership of Kim Jong-II of North Korea and concluded that his narcissistic behaviors were a large component that society's deprivation. The scale of horrors inflicted by some transformational leaders with narcissistic tendencies is conveyed in acts of genocide, ethnic cleansing, holocaust, or rival purification (Volkan & Fowler). It is the combination of the extreme self-serving and self-preserving narcissism heightened by followers' idealization and zealous loyalty of truly transformative leaders that is often harmful (Resick, Whitman, Weingarden, & Hiller, 2009; Rosenthal & Pittinsky, 2006). Although some political and military leaders have been responsible for enormous suffering, narcissistic organizational leaders have also had catastrophic impact on subordinates and institutions.

The damage inflicted by narcissistic organizational leaders can have wide ranging consequences. Sex abuse by some clerical leaders left an estimated 10,000 victims emotionally traumatized (Ronan, 2008). A drug company was sued in a New York court for allegedly using third world countries to circumvent testing protections leading to brain damage and, ultimately, the deaths of many children (Pfizer, 2001). There was an epidemic in slave trade for prostitution in Asia and Eastern Europe targeting vulnerable children (Asia, 2000; Pilisuk, 1998). Sweatshops employed thousands of children and

illegal immigrants working long hours for low wages and in terrible conditions (Sullivan & Lee, 2008). These are examples of criminal exploitation of vulnerable populations by harmful organizational leaders reported to have engaged in narcissistic behaviors.

In the United States, narcissistic leadership behaviors have had a crippling impact on unsuspecting workers. Edid (2004), at the Cornell Institute for Workplace Studies, cited the cost to workers of bad corporate leadership. Using WorldCom, Incorporated, Enron Corporation, and MCI Incorporated as examples, Edid referenced the hundreds of thousands of job losses and billions of dollars lost to egocentric decisions. These economic losses were greater than the gross national products of many countries (Bureau of Economic Analysis, 2010). Companywide layoffs, leaders sent to jail, and finally, bankruptcy were some of the results of leaders acting in their own interests. Edid noted that frontline workers were largely unaware and paid for the self-serving decisions of the top leaders.

The size of egos at the tops of organizations may be related to total compensation. In 1982, Edid (2004) recorded that the average chief executive officers' compensations, as a ratio to production workers' compensations, was 42 to 1. By 2002, that ratio had changed to 400 to 1. Such excesses were not justified by similar corporate value increases. The ones who most often paid for these abusive practices were the production workers and shareholders (Edid, 2004). In seeking new jobs, the disgraced reputations of former employers can make rehiring difficult (Edid, 2004).

Corporate performance and employee morale can suffer when transformational leaders behave in narcissistic or neurotic ways. Hetland, Sandal, and Johnson (2007)

found leader neuroticism predicted exhaustion ($r = .48$) and cynicism ($r = .33$) among subordinates ($N = 298$). Hayward and Hambrick (1997) reported that hubris in senior leaders was related to excessive prices paid for corporate acquisitions, thereby harming stockholders and employees. Chatterjee and Hambrick (2007) found that narcissistic chief executive officers accounted for higher fluctuations in corporate performance, excessive salaries, and higher frequency and scale of acquisitions. There are psychological costs in an environment ruled by transformational leaders that exhibit narcissistic behaviors, defined in this study as harmful transformational leaders. Given the many personal and organizational problems caused by harmful transformational leaders, a screening process should be provided to identify, provide feedback, and develop or separate with potentially harmful leaders. However, accurately detecting potentially harmful transformational leaders is not straightforward.

Researchers suggest that beneficial transformational leadership and narcissistic leadership share some common traits that may be difficult to distinguish. Rosenthal and Pittinsky (2006) pointed out that extreme self-confidence and grandiose visions may be a positive sign of inspirational motivation given by a transformational leader. However, extreme self-confidence and grandiose visions may also be part of a narcissistic personality, compensating for deep insecurities (Resick et al., 2009). Judge, Piccolo, and Kosalka (2009) reviewed a large number of traits which can be found in transformational and well as narcissistic leadership including extraversion, intelligence, charisma, and openness to experience. Campbell and Campbell (2009) emphasized that narcissistic leaders look more like transformational leaders during emergence than several years into

their position, adding duration of tenure into the detection dynamics. Finally, Pullen and Rhodes (2008) studied roles assumed by narcissistic leaders such as servant or star performer that may initially seem indistinguishable from transformational leaders. It would be helpful to have a test that could differentiate harmful leaders from beneficial leaders, even with these similarities in traits.

The Multifactor Leadership Questionnaire (MLQ) may be able to detect subtle differences between harmful and beneficial transformational leaders at the item level. The MLQ is the most popular transformational leadership assessment. If a commonly used instrument can detect patterns of scores that distinguish amongst transformational leaders across the trait continuum, then early detection of potentially harmful transformational leaders can be followed by additional testing, feedback, development, or separation. The problem is that research results are reported as a composite, at a subscale or facet level (Hetland et al. 2007; Lim & Ployhart, 2004). Reliabilities are reduced because it is not known how discriminating or difficult each item of the MLQ is in detecting respondents' unique scoring patterns.

Information is lost when averaging the item scores to calculate the composite value. For instance, Khoo and Burch (2007) found that the MLQ's idealized influence attributed items were positively correlated with narcissism ($r = .27, p = .05$) and the MLQ's individual consideration items were negatively correlated with narcissism ($r = -.34, p = .01$). However, the reported results were at the facet level of the transformational leadership subscale and, therefore, there was no indication if this pattern of detection is viable across the entire range of the transformational trait. Nor was there an indication of

which item(s) within the facet were better at detection than others. It may have been that one item of the facet was particularly strong at discrimination while the composite facet score reduced the overall sensitivity through averaging with three less discriminating items. Scholars and leaders need a statistical procedure that is used to examine item reliability, including the discrimination and difficulty, of the MLQ's transformational subscale. The statistical process, known as item response theory (IRT), provides item level analysis that improves the detection reliability of potentially harmful transformational leaders.

IRT is a collection of models for predicting response patterns to assessment items. The prediction of response behavior to an assessment item is dependent upon an estimate of an individual's latent trait (Embretson & Reise, 2000). IRT can also be used to estimate the individual's latent trait, such as transformational leadership, when that trait is measured by one or more items (De Ayala, 2009). The prediction of response patterns is specified by the selection of a mathematical model from a collection of models (Ostini & Nering, 2006). These models were designed to reflect the behaviors of respondents to different types of assessment items (Embretson & Reise, 2000). As such, De Ayala (2009) depicted IRT as a reliability analysis tool dependant on the assumptions underlying a particular model.

IRT has many applications and provides additional information that cannot be obtained through traditional techniques. In comparison to classical test theory, IRT is seen as a more complex set of techniques generally requiring larger sample sizes (Smith et al., 2007). Interest in applying IRT analysis to psychological assessments has grown

with the adoption of IRT in standardized testing (Smith et al., 2007). IRT is particularly good at describing item characteristics such as discrimination and difficulty (Zagorsek, Stough, & Jaklic, 2006). IRT can also estimate person abilities along a latent trait of interest, invariant from any sample (Samejima, 1977a). Further, IRT assumes that each item in an instrument is not equally reliable, that responses do not have to be normally distributed, and items in the same assessment are not required to be linearly related (Reeve, Hays, Chang, & Perfetto, 2007). These claims cannot be made for classical test theory techniques (Samejima, 1977b). Therefore, as an augmentation to classical test theory, Embretson and Reise (2000) depicted IRT as a set of models predicting specified response behaviors on assessments, thus comparing more precisely the expected responses to the observed responses for items measuring one or more latent traits.

The purpose of this exploratory study was to determine if sufficient reliability exists for detection and intervention of harmful transformational leaders. Reliability can be improved by examining the transformational leadership subscale of the MLQ using IRT analyses. Specifically, I sought to improve the detection reliability of respondents taking the transformational leadership subscale of the MLQ. No known study has been published that applies IRT to the MLQ, as is shown in the literature review. Justification for intervention is supported by sufficient reliability in assessment measures (Kleiman & Faley, 1978). Reliable detection of potentially harmful transformational leaders starts with a better understanding of the discrimination and difficulty parameters at the item level.

Statement of the Problem

Intervention is supportable only if detection is reliable. Since intervention can impact careers, detection reliability at .95 or above is recommended (Nunnally & Bernstein, 1994). The primary assessment for detecting transformational leaders is the MLQ, as is shown in chapter 2. Classical test theory results for the MLQ of transformational facet reliabilities were from .86 to .94 with an average of .90 (Tejeda, Scandura, & Pillai, 2001), insufficient for intervention. With reliabilities below .95, beneficial leaders and harmful leaders are less distinguishable and, therefore, corrective intervention is less justified. With no intervention, subordinates remain at the mercy of harmful leaders. Traditional methods do not support intervention.

Fortunately, IRT is an approach that shows promise in improving reliability estimates. Response information is retained at the most detailed level; the categories of an assessment item. When strict IRT assumptions are supported, the reliability of the perceived transformational abilities of the leaders may be sufficiently high (.95) to warrant intervention. With IRT of the MLQ, narcissistic leaders may be detected and their negative impact reduced (Avolio, Mhatre, Norman, & Lester, 2009). However, IRT is no panacea.

When IRT was applied to two popular leadership assessments, the Leadership Practices Inventory and the Multidimensional Leader-Member Exchange, researchers found that the assessments provided poor information or precision for significant portions of the latent traits being measured (Scherbaum, Finlinson, Barden, & Tamanini, 2006; Zagorsek et al., 2006). For the Leadership Practices Inventory, IRT analysis indicated

poor discrimination between leaders at the upper end of the transformational leadership trait (Zagorsek et al., 2006). For the Multidimensional Leader-Member Exchange, IRT analysis indicated that the extreme lower bands and upper bands of leader-member exchange had low information content (Scherbaum et al., 2006). IRT analyses of these two assessments served the professional community by increasing the precision of the middle portion of the trait range and an understanding of the assessments as a whole.

For the MLQ, detection of potentially harmful transformational leaders requires precisely known properties in discrimination and difficulty at the item level across the entire range of leadership trait. Without understanding the MLQ's ability to discriminate amongst leaders across the trait range, detection of potentially harmful transformational leaders may not be practical or enforceable (Kleiman & Faley, 1978). Therefore, as expressed in the study's problem statement, I explored the discrimination and difficulty for each of the 20 items across the transformational leadership trait range and determined the reliability of the subscale. Item parameterization was a necessary first step in the advancement of knowledge for the detection of potentially harmful transformational leaders. Future researchers may build upon this item parameterization study to augment, revise, or eliminate items to better detect potentially harmful transformational leaders to the relief of hundreds of thousands of subordinate workers.

Nature of the Study

This study was methodological in nature. Rather than examining how the MLQ items are scored differently based on changes in experimental conditions, this study was designed to apply the IRT methodology to a single administration of the MLQ. This

methodological design enabled me to investigate the raters' item response patterns to determine the item characteristics and subscale metrics. Although experimental studies use hypotheses to direct the investigation, methodological studies employ research questions and research objectives.

Research Questions

Research questions serve to guide this study's approach. As is shown in chapter 2, the MLQ suffers from lack of agreement in construct validity and enjoys substantial predictive validity (Antonakis, Avolio, & Sivasubramaniam, 2003). Although the reliability of the entire assessment was adequate (Judge & Piccolo, 2004), detailed study of item level reliability was minimal, as is shown in the literature review. There are a number of questions this study was designed to investigate:

1. How do the observed responses differ from IRT models' expected patterns for each of the five categories of the MLQ's 20 transformational leadership items?
2. Which of the selected IRT models best represents the response patterns observed in the sample?
3. What are the discrimination and difficulty parameters of each of the MLQ's 20 transformational leadership items?
4. What portion of the transformational leadership trait range has the highest reliability estimates?
5. What are the differences in reliability estimation of the MLQ's transformational leadership subscale using IRT versus classical test theory analysis?

These research questions can be stated in the form of research objectives.

Research Objectives

The objectives of this study were to: (a) test the fit of IRT models for the 20 item MLQ transformational leadership subscale, (b) estimate the IRT parameters for each of the 20 items, and (c) evaluate changes in the reliability estimation of scores from the subscale when using IRT versus classical test theory analysis. The results included item discrimination and difficulty parameters, item characteristic curves, and item information curves. Also included were item fit plots, total fit statistics, and an information function for the entire transformational leadership subscale. In addition, IRT marginal reliability estimates were provided for unidimensional items along the transformational leadership trait. These metrics facilitated a discussion about the appropriate utility of using the transformational leadership subscale for detection of potentially harmful leaders.

Research Hypothesis

A methodology study uses questions and objectives rather than hypotheses to guide the research process. Hypotheses can be tested with manipulation of well-designed experimental conditions. This study sought to understand more fully, the psychometric properties of the MLQ's transformational items without manipulating the conditions in which the responses were taken and, thus, use of archival data was appropriate. Instead of a hypothesis, there was a concern over possible findings and future research that should be stated. The IRT analysis of the Leadership Practices Inventory and the Multidimensional Leader-Member Exchange found inadequate discrimination along some of the range of latent traits being measured (Scherbaum et al., 2006; Zagorsek et al., 2006). If the same inadequate discrimination is prevalent with the MLQ, then detection of

potentially harmful transformational leaders along the entire trait range would not be possible without significant modification of instrument items. Although changes to the MLQ were not part of this study, estimating discrimination and difficulty parameters for the transformational leadership items was a necessary first step to possible future revisions of the MLQ for item level discrimination.

Purpose of the Study

In this study, I sought to improve the detection, thereby facilitating intervention of harmful transformational leaders. Detection improvements utilized the original response patterns of the MLQ to achieve the desired predictive reliability (Embretson & Reise, 2000). An IRT analysis examining the 20 items comprising the MLQ transformational leadership subscale retained category information of each item. By analyzing the item discrimination and difficulty parameters, the unique and combined contributions of items in the subscale can be determined (Baker, 2001). Leader's abilities can also be reliably estimated (De Ayala, 2009). I sought to improve the reliability estimates for the transformational leadership items and leaders' abilities by examining the IRT parameters of the responses. Reliable detection and intervention of harmful transformational leaders using the MLQ requires this study's research.

Research Design

A combination of classical and IRT techniques were used to analyze three combined samples ($n = 2,222$) of subordinates. Although the samples contain responses for all 36 leadership trait items of the MLQ, only the 20 items of the transformational leadership subscale were reviewed in this study. The three samples proposed for analysis

were provided by Yair Berson, who conducted research in Israel and gave permission for this review (Y. Berson, personal communication, October 14, 2009). Two of the samples were from Israeli corporations. The largest sample came from a large telecommunications company, in which employees rated their supervisors ($n = 1,600$). The smallest sample was from 26 Israeli companies, in which top management teams rated their direct leaders ($n = 282$). The middle sized sample was from professional Israeli basketball team players rating their coaches ($n = 357$). Combining these samples provided a larger calibration sample and therefore allows more stable parameter estimates (Edelen & Reeve, 2007).

Data screening and degree of dependency between raters' responses were established. The data were reviewed for missing values, indiscriminant responses, typing errors, and adequate item by category cell frequencies. The degree of independent observations must be known for adequate analysis. To examine the effect of correlated observations from raters within the same subordinate group, the intraclass correlation coefficient (ICCC), using a one way random effects model, was used to indicate the degree of within group variation of the combined samples (Shrout & Fleiss, 1979). If the ICCC was at or below .20, then the rater's individual responses was retained ($n = 2,222$). If ICCC was above .20, then a random rater would have been selected from each leader ($N = 357$) to achieve independent observations. For the MLQ, there was reason to believe the ICCC was below .20. Walumbwa, Avolio, and Zhu (2008) reported an ICCC of .10 on the MLQ rater version as part of their justification of analysis at the individual subordinate level.

Once the use of rater level responses had been investigated, classical psychometric item analysis was conducted. Internal consistency was examined for each item of the transformational leadership trait. Item discrimination and difficulty was evaluated using traditional methods. Scherbaum, Finlinson, Barden, and Tamanini (2006) calculated item discrimination using corrected item to total subscale correlations and item difficulty was calculated through item mean scores and standard deviations.

IRT assumptions were evaluated. Because single latent trait models were proposed, unidimensionality needed to be examined using maximum likelihood factor analysis. Details of this IRT assumption and of local independence are described in greater detail in the Literature Review chapter. If multiple dominant dimensions were discovered with the 20 transformational leadership items, either separate IRT analyses would have been conducted for each dimension using two IRT models, or the item(s) with low factor loadings would have been removed from the analysis based on examination of loadings and IRT discrimination parameters.

The two IRT models proposed were applicable for ranked polytomous items. Model selection and research methodology for this study was based on research by Scherbaum et al. (2006), who analyzed the Multidimensional Leader-Member Exchange scale. The first IRT model used for analysis was the generalized graded unfolding model (GGUM) by Roberts and colleagues (Koenig & Roberts, 2007; Roberts, 2008; Roberts, Fang, Cui, & Wang, 2006). The software for analyzing the GGUM was the GGUM2004 (Roberts et al., 2006). The second IRT model was based on Samejima's (1969) graded response model (GRM) for homogeneous ranked polytomous items. The IRT analysis for

the GRM utilized MULTILOG software version 7 (Thissen, Chen, & Bock, 2003). MULTILOG software had been shown to be robust to violations of various IRT assumptions (Kirisci, Hsu, & Yu, 2001). Detailed discussions concerning Samejima's (1969) GRM and Robert's (2008) GGUM models can be found in the Literature Review chapter, Methodology Considerations section.

The validation of choosing the right models was determined by the degree of fit between the observed sample data and the models' expected responses to item parameters. Data to model fit was indicated by the values of the chi-squared over degrees of freedom metric for item combinations in singles, doubles, and triples as described by Drasgow, Levine, Tsien, Williams, and Mead (1995) using MODFIT software version 1.1 (Stark, Chernyshenko, Chuah, Lee, & Wadlington, 2001). Observed versus expected responses can be partly influenced by the degree of local independence and, therefore, could have shown poor data to model fit indices above the three cutoff criteria (Drasgow et al., 1995, Careless, 1998). Graphical analysis of response functions for each category assisted in determining extent and possible impact of any problems with data to model fit.

After testing the combined calibration samples, each of the three samples was analyzed separately for mean person trait differences. To equate person mean differences of each of the three samples, the mean of the telecommunication sample was used as an anchor (Embretson & Reise, 2000). Item and transformational leadership subscale parameter estimations for all dimensions of the combined samples were analyzed. Results from all procedures were recorded, reviewed, and presented. Research questions and objectives were used to comment on the results. Further discussions of the Research

Design can be found in the Methodological Considerations section in Chapter 2 and in Chapter 3.

Theoretical Framework

Using every part of original data provides the most precise analytic results. IRT analysis of the MLQ is based on the responses of each individual in selecting only one of five categories for each of the 20 study items. The unique pattern of choosing 20 distinct categories sets that individual apart from those choosing differently. IRT retains this basic level of information throughout the analysis in estimating each individual's transformational leadership ability (De Ayala, 2009). The multiple patterns of all the respondents show the degree of difficulty respondents had with answering an item (Embretson & Reise, 2000). Examining the patterns of responses can also show which items are better at distinguishing those with lower transformational abilities from those with higher levels; called item discrimination (Samejima, 1977a). Therefore, IRT is used to determine an individual's ability level and an item's difficulty and discrimination levels with great precision (Embretson & Reise, 2000). Using every piece of original response data for each individual improves the reliability of ability and item estimates over the averaging approach of classical test theory (Samejima, 1977b).

IRT analysis is based on decision theory applied to assessments. Conceptually, IRT can be viewed as multiple logistic regressions, since respondents, conditional on their latent trait ability, are grouped by category difficulty and item discrimination across each item of a dimension (De Ayala, 2009). The MLQ is designed to be a ratio homogeneous polytomous assessment using a 5-point Likert scale based on observed

leader behaviors and attributes (Avolio & Bass, 2004a). There are many IRT models that may be viable for analyzing MLQ rater responses. Most of these models derive from the Rausch model, such as the partial credit model, the generalized partial credit model, the rating scale model, Robert's (2008) GGUM, and Samejima's (1969) GRM, to name a few. As indicated, the choice of IRT models for this study, the GRM and the GGUM, follows Scherbaum et al. (2006) methodology, which is discussed in the Literature Review chapter.

With the models chosen for this study, there was an item parameter procedure followed by a person parameter procedure required for every sample. In IRT analysis, an initial sample of responses to an instrument was used to calibrate item and person parameters using two sequential software analyses (Zagorsek et al., 2006). Once specified, the item discrimination and difficulty parameters are independent from the sample of responses (De Ayala, 2009). Person ability values were also estimated independent of any sample after calibration (Baker, 2001). With an assumption of local independence, the summation of item information produced total scale information and a standard error of measure (Samejima, 1977a). It was the lack of dependence on other persons in the sample, other items in the assessment, and the precision of the reliability estimates that provided the significant benefits of IRT over classical test theory (Samejima, 1977b). A comparison of IRT and classical test theory is provided in the Literature Review chapter. The improved reliability of detecting a harmful transformational leader, provided by IRT, is fundamental to supporting intervention.

Operational Definitions of Terms

Category boundaries – The interface between options or answers of an assessment item (Ostini & Nering, 2006). Typically these options appear as part of an assessment, questionnaire, instrument, or test. An example is the MLQ, which has five category choices for each item representing a behavioral statement. The choices range in score from zero to four with zero meaning *not at all* and four meaning *frequently, if not always*. For a 5-point Likert scoring system, as in the MLQ, there are four boundaries separating the five categories (De Ayala, 2009).

Contingent reward – A facet of transactional leadership, it is the leader's promise of a reward in exchange for the follower's efforts toward a goal (Avolio & Bass, 2004a). Contingent reward is considered a constructive leadership style (Avolio & Bass, 2004a).

Fit plots – Visual overlay between expected model responses and observed responses. This overlay allows graphical comparisons and a visual determination of model to data fit (De Ayala, 2009).

Full range leadership model – Also called transformational leadership theory. The term that represents three leadership styles: transformational, transactional, and laissez-faire (Avolio & Bass, 2004a). These three leadership styles have also been variously described as charismatic, constructive, corrective, coercive, and avoidant behaviors (Avolio & Bass, 2004a).

Harmful Transformational Leader – A narcissistic personality type with transformational leadership abilities has been a destructive historical combination (Post, 2008; Rosenthal & Pittinsky, 2006; Volkan & Fowler, 2009). For the purposes of this study, harmful

transformational leaders exhibit narcissistic behaviors and have an average or higher score on the MLQ's 20 item transformational leadership subscale.

Idealized influence attribute – Part of charismatic leadership, which in turn is part of transformational leadership, it is the emotional response of a follower who takes pride in being associated with the leader (Avolio & Bass, 2004a).

Idealized influence behavior – Part of charismatic leadership, which in turn is part of transformational leadership, it is the moral response of a follower to the leader's sense of purpose or mission (Avolio & Bass, 2004a).

Individual consideration – Part of transformational leadership, it is the attention paid by the leader as a mentor to the wants, needs, and ambitions of the follower (Avolio & Bass, 2004a).

Inspirational motivation - Part of charismatic leadership, which in turn is part of transformational leadership, it is the conveyance of meaning, optimism, and a compelling future vision the leader invites the follower to achieve (Avolio & Bass, 2004a).

Intellectual stimulation - Part of transformational leadership, it is the leader's efforts to increase the mindset of the follower to approach problems differently, increase innovation, and to question fundamental assumptions (Avolio & Bass, 2004a).

Intraclass correlation coefficient – A classical test theory method for determining the ratio of variation within a group as opposed to between groups as reflected in the grand mean (Tabachnick & Fidell, 2007).

Item – An item is a question or statement on an assessment, questionnaire, instrument, or test. For the purposes of IRT, an item has one or more choice of options, which can be

scored on a correct or incorrect basis, scored on a categorical choice basis, or scored on a continuous scale (Ostini & Nering, 2006).

Item characteristic curves – A visual graph of an item, in which each category of an item has separate curves that show discrimination and category difficulty (De Ayala, 2009).

Item information function – A visual graph showing the line traced by an item's information, which is derived by the underlying category discrimination and difficulty functions (De Ayala, 2009).

Item response theory – A set of models that characterize response patterns to various items of an instrument (De Ayala, 2009). Based on decision theory, item response theory models calculate the probability of a respondent choosing from the available options of an item conditional on the latent trait being measured by the instrument (De Ayala, 2009). Item difficulty and item discrimination are calculated in IRT to characterize an item (Samejima, 1977a). IRT analysis produces additional graphical and numerical indicators at the category, item, subscale, and assessment levels (Embretson & Reise, 2000).

Laissez-faire leadership – It is the avoidance of leadership responsibility (Avolio & Bass, 2004a). In the MLQ, laissez-faire is one of three subscales of the full range leadership model, which include transformational and transactional leadership (Avolio & Bass, 2004a). It is the only leadership style that is both a subscale and a lower order facet (Avolio & Bass, 2004a). Laissez-faire is considered an avoidant leadership style (Avolio & Bass, 2004a).

Local independence –For unidimensional IRT models, local independence occurs when the only relationship between responses, for any two items, is the underlying trait (Embretson & Reise, 2000).

Management by exception active - Part of transactional leadership, it is the leader's active control and correction of followers' mistakes in work performance (Avolio & Bass, 2004a). Management by exception active is considered a corrective leadership style (Avolio & Bass, 2004a).

Management by exception passive - Part of transactional leadership, it is the leader's coercive approach in disciplining followers for breaking a performance standard or expectation (Avolio & Bass, 2004a). Management by exception passive is considered a coercive leadership style (Avolio & Bass, 2004a).

Multifactor leadership questionnaire – Originally designed by Bass (1985) and subsequently jointly revised with Avolio, the MLQ is a popular transformational leadership assessment, in which 36 items test for three leadership styles: transformational, transactional, and laissez-faire (Avolio & Bass, 2004a). Nine additional items on the MLQ, test for subjective outcomes of leadership satisfaction, follower extra effort, and leadership effectiveness (Avolio & Bass, 2004a).

Narcissism – A personality trait characterized by self-absorption and grandiosity on one side and hostility and self-preservation on the other (APA, 2000).

Option response function – A visual way of comparing expected individual category responses of an item to the observed responses (Drasgow et al., 1995). In MODFIT

software, option response functions are a graphical method of determining data to model fit (Zagorsek et al., 2006).

Ranked homogeneous polytomous items – This term applies to the type of options or categories available on an item of an instrument. Ranked items are similar to ordered items, in that the categories of an item are in increasing order of importance or value. Homogeneous items refer to the categories of an item being on the same scale of measure. Polytomous items refer to more than one correct or partially correct category choice for each item (Ostini & Nering, 2006).

Total information function – The sum of individual item information functions becomes the total information function of the entire instrument (De Ayala, 2009).

Transactional leadership - One of three higher order subscales that constitute the full range leadership model (Avolio & Bass, 2004a). The other two leadership styles are transformational and laissez-faire. Transactional leadership encompasses three lower order facets of contingent reward, management by exception active, and management by exception passive (Avolio & Bass, 2004a).

Transformational leadership theory – See full range leadership model. This theory describes the added performance possible from followers when their leader exhibits a combination of certain transactional and transformational behaviors and attributes (Avolio & Bass, 2004a).

Transformational leadership – One of three higher order subscales that constitute the full range leadership model (Avolio & Bass, 2004a). The other two leadership styles are transactional and laissez-faire. Transformational leadership encompasses five lower order

facets of idealized influence attribute, idealized influence behaviors, inspirational motivation, intellectual stimulation, and individual consideration (Avolio & Bass, 2004a).

Unidimensionality – An IRT assumption that requires the items of an IRT analysis to measure a single latent trait (De Ayala, 2009).

Assumptions, Limitations, Scope, and Delimitations of the Study

Some important facts were assumed but not necessarily verified for this study.

These assumed facts had to do with the collection of archival data used for analysis. The first such fact was that all three samples were based on three separate single administrations of the MLQ, using appropriate controls. The sample descriptions and procedures were reported in peer-reviewed journals (Berson & Linton, 2007; Berson, Oreg, & Dvir, 2008) and described in chapter 3. The second assumed fact was that adequate forward and backward translation techniques of the Hebrew paper version of the MLQ were used in the collection of archival data. The efficacy of the translation process was validated by Avolio et al. (1999) and was estimated in this study by comparing translated scores with untranslated scores reported in chapter 4. Conversations with the owner of the data (Y. Berson, personal communication, October 14, 2009), and published literature (Berson, 1999; Berson & Avolio, 2004; Berson & Linton, 2005; Berson & Sosik, 2007), suggested that these assumptions were appropriate.

The primary assumption for this study was that unidimensional models are useful in evaluating the transformational leadership subscale. Because the MLQ was designed to represent multiple dimensions, it was perhaps more appropriate to use multidimensional IRT models. Development of multidimensional models is an active area of research and

the software to run such models is in its infancy (De Ayala, 2009). With the lack of viable alternatives, unidimensional IRT models were used in this study with the expectation that factor analysis could adequately partition the transformational leadership subscale into useable item groupings. This dimensional partitioning may not have been viable and thus might have constituted a severe study limitation. As multidimensional models and associated software are further developed, research using the MLQ would undoubtedly be better served with these more complex and more appropriate model choices.

There were several important limitations to this study, which may indicate substantial weaknesses. The first is that only 20 items of the 36 leadership trait items in the MLQ assessment were examined. The transactional and laissez-faire subscales, therefore, were not evaluated. Although the transformational leadership subscale is the most heavily used subset of the MLQ (Judge & Piccolo, 2004), it does leave a significant gap in the item characteristics of the other 16 items and would be an area for further research.

The focus of this study was discriminating harmful from beneficial transformational leaders rather than examining nontransformational leaders. Including nontransformational leadership responses would have introduced additional data to model fit errors. One practical reason for excluding the transactional subscale from this study's IRT analysis was the inverted relationship between ranked categories and the latent trait. As described by the Avolio and Bass (2004a), the three transactional facets range from monotonically increasing to monotonically decreasing in relationship with the latent trait

of the facet. There appeared to be no clear point of reversal (Avolio & Bass, 2004a). A monotonically increasing item refers to item category choices, which increase as the latent trait increases (Scherbaum et al., 2006). For these items, a higher category choice implies higher latent trait ability (De Ayala, 2009). However, this relationship is reversed for some facets of the transactional leadership subscale and for all items of the laissez-faire subscale even though all 36 of the MLQ items use the same ranked category scale anchored by *not at all* to *frequently, if not always*.

A clear example of this reversal is that the *frequently, if not always* category response for a laissez-faire leadership item corresponds to lower transformational leadership ability. Because the GRM and the GGUM models assume a monotonically increasing response pattern, these models would be inappropriate for the transactional or laissez-faire subscales on transformational leadership ability. Reversing the scoring scale of transactional leadership subscale would not resolve this issue, because it is unclear where on the scale the reversal takes place (Judge & Piccolo, 2004). Besides focusing this study on discriminating amongst transformational leaders the other reason for considering only the 20 transformational leadership items, was that this subscale comprises a distinct theoretical construct related to charismatic leadership as is shown in the Literature Review chapter.

The second limitation was that the three samples combined for calibration were insufficient in size and would have introduced higher levels of standard error of measure. Unlike classical test theory, no agreed guidelines exist for determining appropriate sample size in IRT analysis (De Ayala, 2009; Embertson & Reise, 2000; Kirisci et al.,

2001; Reise & Yu, 1990). Samples sizes exceeding 3,000 are used in IRT analysis for three parameter logistic models due to the difficulty of estimating the guessing parameter (Drasgow et al., 1995). However, this study did not involve a guessing parameter and was able to use a smaller sample size. Although there appears to be no agreed minimum sample size in literature (De Ayala, 2009; Kirisci et al., 2001), the initial combined calibration samples for this study was 2,222 cases, expected to yield stable parameters. For leadership studies in general, obtaining sample sizes approaching 3,000 may be problematic (Avolio, Walumbwa, & Weber, 2009). Some authors suggest smaller sample sizes of 250 or 500 are usable for exploratory research (De Ayala, 2009; Russell, 2002). For instance, Scherbaum et al. (2006) used a sample size of 445 for IRT analysis of Multidimensional Leader-Member Exchange. However, as the number of items considered in a single analysis decreases, the number of individual responses must increase, to provide sufficient information. In this study, the number of items considered in one IRT analysis would have been impacted by the dimensional analysis. There was a possibility that as few as four items per analysis were used. The item parameters were reported with associated standard error of measure so that future researchers may improve on these estimates using larger sample sizes or multidimensional models (Embretson & Reise, 2000). Further discussion of sample sizes is found in the Methodological Considerations section of chapter 2.

The third limitation was insufficient responses for each category of each item. Although IRT analysis produces parameter estimates that are invariant of sample there needs to be at least five responses for each category of each item in the calibration sample

to provide stable and informative parameter estimates without collapsing categories. (De Ayala, 2009). Scholars suggest that sufficient responses exist without collapsing categories (Berson, Oreg, & Dvir, 2008; Berson & Sosik, 2007). The matrix of responses for item by category counts determined cell frequency sufficiency. The matrix cell frequencies are reported in chapter 4.

A fourth limitation was the use of categorical data to conduct item factor analysis. If category endorsement does not follow a normal distribution then item factor analysis may be influenced by item difficulty rather than true correlations between items (With & Edwards, 2007). Polychoric correlations are sometimes used in theoretical investigations to assign item difficulty to thresholds allowing for truer item correlations (Flora & Curran, 2004). However, the SPSS (2009) software used in this study did not support polychoric correlations. Instead, item level analyses was conducted and reported to show the extent the assumption of normal distribution was violated. Examination of the MLQ normative data suggested a slight negative skewness less than 1.0 but otherwise a fairly normal distribution of the category responses.

The fifth limitation dealt with the exploratory nature of this study. Because IRT analysis has not been previously established for the MLQ, there was no comparison to assess the viability of item parameter estimates. Additionally, if findings demonstrated that certain items did not add significantly to the measurement of transformational leadership trait or if there was poor discrimination along a certain portion of the latent trait continuum, it would not have been evident how to adjust the MLQ to accommodate these concerns since no alternative items were proposed. Replication of the findings can

determine the usefulness of parameter estimates and extensions to this study may provide insights to observed assessment limitations.

This study was bounded by examining 20 item characteristics comprising the MLQ transformational leadership subscale using item response theory. Transformational leadership subscale has been independently examined by peer-reviewed articles and constitutes a major focus of the transformational leadership theory (Judge & Piccolo, 2004). The specific description of what was in the scope and what was out of scope for this study is described in more detail in the Literature Review chapter.

Significance of the Study

The gap in the literature is that research using classical test theory is incomplete because reliable detection of harmful transformational leaders has not been investigated using the MLQ. Detection revolves around differentiation of responses at the item level across the trait continuum as is shown in chapter 2. The MLQ has not had item level research performed in the manner proposed by this study in the 25-year history of the assessment, as will also be shown in chapter 2. A second gap in the literature is that IRT has not been applied to the MLQ, as is shown in chapter 2. This study explicitly fills the two gaps in the literature by using IRT analysis to support detection of harmful transformational leaders.

Professional application of the results can lead to improved identification of harmful transformational leaders. A test case in chapter 5 illustrates one detection method. Discovering all the combinations of responses to the MLQ, which indicate a potentially harmful transformational leader, requires additional research and testing using

the invariant item characteristics from this study. Professional application of screening for harmful transformational leaders is improved by this study's reliability research.

Professional application of improved detection must be accompanied by intervention for positive social change. Fortunately, professional application of leadership intervention has worked. Avolio, Mhatre, Norman, and Lester (2009) conducted a meta-analysis of 57 different types of leadership interventions to determine impact through effect size. Results showed moderate ($d = .43$, $SD = .31$) effect size for mostly male environments and large ($d = .53$, $SD = .53$) for mostly female environments. Of these 57 studies, the seven transformational leader interventions had intervention effect sizes of $d = .47$ for mostly males and $d = .60$ for mostly females. The moderate to large effect sizes provided evidence that, after detection, intervention made a substantial impact on the way leaders related to subordinates. If professional application of intervention can lead to positive social change then detection must accurately identify those in need of intervention.

Positive social change is upholding the worth, dignity, and positive development of those persecuted by harmful leaders and those falsely accused of being harmful leaders. To bring relief to suffering subordinates and correctly identify the perpetrators, detection must be accurate and reliable. However, detection methods using classical test theory are inadequate. Cronbach's alpha guidelines for selection and intervention are .95 (Nunnally & Bernstein, 1994). Transformational leadership facets of the MLQ were shown to range from a Cronbach's alpha of .86 to .94 with an average of .90 (Tejeda et al., 2001). Left to classical test theory, interventions would not be readily supported. Two

other leadership assessments were found that have applied IRT techniques, demonstrating increased precision; however, neither study was used to detect harmful transformational leaders (Scherbaum et al., 2006; Zagorsek et al., 2006). Positive social change designed to bring improvement of human and social conditions is the basis of this study by increasing the reliability of item and person ability detection using IRT, giving hope to hundreds of thousands subjected to harmful transformational leaders.

Positive social change of widespread detection and intervention of harmful transformational leaders can restore dignity and worth to more than individuals. Organizations can benefit by having their leaders screened, possibly preventing situations like Enron, MCI, and WorldCom (Edid, 2004). Institutions may retain their reputations and promote human welfare such as the Catholic Church through detection and intervention of those narcissists capable of sex abuse (Ronan, 2008). Cultures and societies like those in Sri Lanka may feel that their children are safer by screening out candidates who seek adoption as a means of sexual exploitation (Cook, 2005). Finally, countries may reduce mass murder by detecting harmful transformational leaders before granting control of their armed forces. Positive social change can come by denying access to vulnerable populations based on reliable detection and professional intervention. This study is a critical step in identifying harmful transformational leaders thereby promoting positive social change for individuals of all societies.

Summary

Harmful transformational leader have some traits in common with beneficial transformational leaders (Khoo & Burch, 2007). Discerning amongst transformational

leaders is not straightforward (Judge, Piccolo, & Kosalka, 2009). The MLQ is the primary research vehicle for transformational leadership studies, as is shown in chapter 2. Item level analysis was needed to lay the foundation for detection of potentially harmful transformational leaders (Hetland et al., 2007). IRT item level analysis had not been performed for the MLQ during its 25-year history, as demonstrated in the Literature Review chapter. This study analyzed 20 items comprising the MLQ's transformational leadership subscale using IRT.

IRT has many advantages over traditional item analysis techniques. IRT is a reliability analysis method conceptually similar to logistic regression based on modeling response patterns of various instrument items (Ostini & Nering, 2006). The advantages provided by IRT analysis over classical test theory include sample independence for people and item characteristics (Samejima, 1977a). However, the use of IRT for the MLQ is not without severe limitations as described in chapter 1. Because the MLQ factor structure is less stable for heterogeneous samples, unidimensionality of the transformational leadership subscale was not assured (Antonakis et al., 2003). With separate item loadings by factors and the robustness of IRT software packages to some violations of unidimensionality, estimates of item characteristics were expected to be viable. IRT parameters included item discrimination and item difficulty values of the 20 transformational leadership items using two unidimensional IRT models (Kirisci et al., 2001; Tabachnick & Fidell, 2007).

A combination of three archival samples was proposed that yielded viable item and person parameter estimates. Three samples were combined to provide a larger

calibration sample ($n = 2,222$) for more stable IRT parameter estimates (Edelen & Reeve, 2007). This sample size limitation is discussed further the Literature Review chapter and the Research Method chapter. These samples included Israeli business employees and sports team players rating their direct supervisors and coaches, respectively. It was expected that this study would be the first published IRT analysis of the MLQ and would assist researchers and practitioners increase precision in detecting potentially harmful transformational leaders while providing greater information on the MLQ's psychometric properties.

The purpose of the study was to provide greater detail of item parameterization needed to differentiate harmful from beneficial transformational leaders. Many of the advantages and limitations of the MLQ and IRT analysis are described in the Literature Review chapter. Besides comparing IRT to classical test theory, a detailed account of the history, underlying theory, and research results of the MLQ are presented. The Literature Review chapter concludes with the Methodological Considerations section from past research findings. These research approaches are then detailed in the Research Method chapter of this study. The Research Method chapter includes a description of the samples and the analysis techniques that were applied, including any significant criteria. The Results chapter will describe the factor analysis output and results from IRT analyses describing the 20 transformational leadership item parameters in detail. Any modifications of the proposed methodology that were necessary are explained in the Results chapter. The last chapter includes the discussion of results along with conclusions and recommendations. Study limitations and future research suggestions conclude this

study. The expectation was that greater discrimination precision of the MLQ items could be achieved with the results from this study. With greater assessment precision, detecting potentially harmful transformational leaders and adopting appropriate intervention strategies may be possible at an earlier stage.

Chapter 2: Literature Review

Introduction

The psychometric properties of the MLQ's transformational leadership subscale were explored through IRT for the first time. The MLQ is the most widely used research instrument for transformational leadership (Hinkin & Schriesheim, 2008a). Over 25-year history of the MLQ, no IRT analysis was found, as is demonstrated in the Gap in Current Research section of this chapter. For instance, the degree of difficulty and discrimination posed by each assessment question in predicting transformational leadership traits was not explored in detail. However, according to De Ayala (2009), IRT is a body of knowledge stemming from decision theory and logistic regression that facilitates examination of individual instrument items to determined reliability with a level of precision not available using classical test theory.

New information is available using IRT that can benefit leaders in assessing transformational trait abilities through item parameter estimates. By applying IRT to the MLQ, the scale for person traits is the same as the scale for item parameters, so that comparisons and score predictions are possible at an individual participant level (Emberson & Reise, 2000). In addition, rater responses were examined for multiple leader sources, including leaders of a large Israeli telecommunications company, top business professionals of various companies, and professional basketball coaches. It was expected that this study would add new information on the reliability of the transformational leadership subscale of the MLQ. Conclusions can improve the detection

of potentially harmful transformational leaders across the trait continuum and adoption of intervention strategies at earlier stages.

Organization of Literature Review

The literature review provides confidence in the assertions that IRT can provide new, relevant, and practical information about the MLQ's transformational leadership subscale. The search strategies and the gap in literature are followed by a detailed discussion comparing and contrasting classical test theory and IRT. IRT benefits and limitations are examined to provide knowledge about how item parameters are calculated. Leadership is introduced as it pertains to leadership assessments. The MLQ is reviewed in terms of the instrument's development, underlying theory, and findings in research literature. The implications of bounding this study with the transformational leadership items are discussed. Pertinent methods of research are examined including contextual variables, assessment form and language, participant characteristics, model specification, and software usage from past research. Finally, the study is summarized with anticipated benefits described.

Literature Search Strategy

Sources examined were extensive, and provided a practical foundation upon which to construct this study's design. The majority of information in this review came from EBSCO electronic databases of peer-reviewed journals. Specifically, Academic Search Premier, Business Source Premier, and PsycINFO were used. Sage electronic databases were also searched to provide additional peer-reviewed content of a methodological or statistical nature. Because of the importance of historical development

for the MLQ and IRT, seminal books and articles were used starting in 1978 and 1927, respectively. A limited number of reference books from 2000 to 2009 were used as primary sources in technical descriptions of statistical concepts and processes. For reference sources, peer-reviewed articles cited these same or similar sources. Key assumptions, upon which this research was based, were from peer-reviewed articles whose publication dates ranged from 2004 to 2009.

Gap in Current Research

No published report can be found applying IRT to the MLQ. Numerous electronic databases have been searched with keywords *multifactor leadership questionnaire* or *MLQ* and *item response theory* or *IRT*. These electronic databases included EBSCO, Gale, Ovid, Proquest, and Sage. In addition to electronic searches, the copyright holder of the MLQ, Mind Garden Incorporated, had no knowledge of any study of this nature (R. Most, personal communication, May 7, 2009). Inquiries with the remaining author of the MLQ (B. Avolio, personal communication, May 12, 2009) and additional discussions with the author's research colleague (Y. Berson, personal communication, May 14, 2009) also confirmed that no such published report existed.

Because a gap existed in terms of evaluating the MLQ using IRT, this was an area of research that provided additional insights into the psychometric properties of this heavily used instrument. De Ayala (2009) showed that IRT analysis can increase the precision of certain reliability parameters at the item, test, and participant levels unavailable using current classical test theory methods. Because psychometric characteristics of the MLQ have not been completely resolved (Embretson & Reise,

2000), this study provides added analytical clarity and details for further investigations to enhance understanding of item level reliability parameters.

Classical Test Theory

Traditional psychometric techniques sacrifice item and respondent granularity to achieve important assessment level details. Analysis techniques for instruments, in classical test theory, seek to discover latent trait measures at the entire assessment level (De Ayala, 2009). This approach optimizes the information available from the instrument while sacrificing details about information at the item level and individual participant level (Samejima, 1977a). The focus of classical test theory tends to be on an entire test rather than at the item or participant level and this focus is reflected in the formulation of reliability indices.

To achieve useful assessment wide metrics, traditional techniques rely on important underlying assumptions. For testing multiple latent traits and the effects of independent variables on assessment items, such as in the MLQ, classical test theory works under the assumption that latent and measured variables of interest are quantitative and continuous and that the study variables are normally distributed (Tabachnick & Fidel, 2007). The relationship between any two or more variables is also assumed to be linear. Although some violations of these assumptions can be accommodated, according to Tabachnick and Fidel (2007), the precision of predicting relationship outcomes can degrade quickly if these assumptions are violated.

The MLQ's design does not conform to classical technique assumptions, which can reduce result precision. For instance, the MLQ's items are not measured on a

continuous metric (Avolio & Bass, 2004a). The ordered categorical scale of the MLQ items is not the same as a single continuous variable and does not have even distributions (Avolio & Bass, 2004a). Nonlinearity is typified by unique cumulative distribution functions per category of an item (De Ayala, 2009). Because of these multiple violations to classical test theory assumptions, any conclusions must be cautiously applied to the test as a whole.

Reliability Measures

Classical test theory offers several methods of calculating reliability measures. Questionnaire reliability can be determined through test-retest reliability, parallel forms, and internal consistency (Cohen & Swerdlik, 2005). In the MLQ research, internal consistency was reported as the primary means of determining reliability using Cronbach's alpha (Kanste et al., 2007). However, there are fundamental limitations with classical test theory when it comes to reliability measures.

Reliability is dependent on the sample used to derive the measure (Tabachnick & Fidell, 2007). This dependency can be seen in the formulation of the coefficient alpha $\alpha = [k/(k-1)] [1 - (\sum \sigma_i^2 / \sigma^2)]$, where k is the number of items, $\sum \sigma_i^2$ is the sum of the variance of all items, and σ^2 is the total score variance (Cohen & Swerdlik, 2005). The total score variance σ^2 is in turn made up of σ_{tr}^2 true variance plus an error variance σ_e^2 such that $\sigma^2 = \sigma_{tr}^2 + \sigma_e^2$ (Cohen & Swerdlik, 2005). The test score variance σ^2 is also equivalent to the standard deviation of the observed test scores squared. The observed scores are dependent on the sample used to gather the scores for the test. Therefore, coefficient alpha is dependent on a variance that in turn is directly dependent on the sample of

respondents marking their answers on a questionnaire (De Ayala, 2009). Internal consistency in the form of coefficient alpha is, as described by De Ayala (2009), dependent on the sample used in describing a single administration of an instrument.

Besides sample dependence, classical test theory incorporates nonsystematic errors variances in reliability measures (Zagorsek et al., 2006). Errors of random differences that affect the participants can be such conditions as: level of anxiety, conditions of administering the test such as time of day, and the test itself, such as web based or paper and pencil versions of the test (Cole, Bedeian, & Field, 2006). Classical test theory assumes that these random variances, with enough repetition of test administrations to the same participants, eventually cancel out (Grimm & Yarnold, 2000). This approach is problematic, according to Cohen and Swerdlik (2005), when only one test is administered to a specific sample of participants.

Classical test theory assumes linearity across all items, which is rarely found in practice. Because published metrics aggregate the assessment items, an instrument is generally assumed to be equally difficult and discriminating across all test items (Zagorsek et al., 2006). This linear assumption poses a problem of precision loss as aggregation subsumes item difficulty and discrimination differences (Samejima, 1969). Even with items that are equally difficult, some items are differentially discriminating (Samejima, 1969). The information showing that some questions are better for certain cutoff criteria can be lost in aggregation. Kleiman and Faley (1978) showed that item information lost in aggregation reduced the assessment's face validity, which became problematic when justifying employment decisions.

Validity Measures

The MLQ has one or more continuous latent traits as independent variables for multiple items. Ordered categorical in nature, the MLQ's items are the dependent variables. For this study's samples, there was a single occasion to collect responses from the participants. In classical test theory, there are no multivariate analysis methods available that exactly meet these conditions (Tabachnick & Fidell, 2007).

By relaxing these constraints, however, classical test theory can provide meaningful insights into assessments. Instead of a continuous scale, the Likert scale can approximate an interval scale (Tabachnick & Fidell, 2007). A further relaxation of assumptions can allow all dependant variables to appear normally distributed across those intervals (De Ayala, 2009). Factor analysis is then a potential technique available for these conditions. Tabachnick and Fidell (2007) indicated that, after some relaxation of critical assumptions, traditional techniques can be used to examine unique, shared, and error variances of the dependant variables.

Confirmatory factor analysis is an important traditional technique, which can be useful as assumptions of the model are relaxed. Factor analysis is one of the primary methods to validate that observed scores on an instrument fit the explicitly hypothesized constructs of the underlying theory (Cohen & Swerdlik, 2005). In confirmatory factor analysis, factors are specified a priori to affect particular items. It is a more stringent approach than exploratory factor analysis, which does not constrain factor loading a priori (Grimm & Yarnold, 2000). Unlike structural equation modeling, however, confirmatory factor analysis does not investigate causal relationships (Grimm & Yarnold,

2000). Antonakis, Avolio, and Sivasubramaniam (2003) and Kanste, Miettunen, and Kyngäs (2007), found that the MLQ has had a history of inconsistent results when using exploratory and confirmatory factor analysis.

There are a number of metrics to determine if the assessment measures what it is supposed to measure. Validity measures such as construct, content, criterion, face, predictive, concurrent, differential, internal, and external validities all help to establish that the latent trait underlying the responses is the objective of the measure (Grimm & Yarnold, 2000). Although some work has been done on construct validity comparing the MLQ with charismatic assessments (Rowold & Heinitz, 2007) and personality tests (Lim & Ployhart, 2004), construct validity measures comparing the MLQ with other transformational tests were not reported (Avolio & Bass, 2004a). Bono and Judge (2004) emphasized that the MLQ's predictive validity is the assessment's primary benefit, despite what Antonakis et al. (2003) described as the construct validity inconsistencies.

Item analysis is not just the domain of IRT. Classical test theory has techniques for calculating the difficulty and discrimination of individual items. One of those techniques is calculating the item difficulty index and item discrimination index (Cohen & Swerdlik, 2005). Item difficulty index is the number of participants scoring correctly on an item divided by the total number of participants. A high item difficulty index indicates an easier item. Item discrimination index, on the other hand, uses the assumed normal distribution of total test scores to compare a correct score on an item with those participants in the top part and the bottom part, of the total score range. A high discrimination index value means that the item was more likely to have been answered

correctly by high total test score participants than low total test score participants. Negative discrimination index values are problematic indicating that lower scoring participants did better on this item than high scoring participants (Cohen & Swerdlik, 2005). The item difficulty and item discrimination indices are crude ratio measures, which according to De Ayala (2009), are heavily sample dependent and provide no error estimates or measurement precision.

A more robust classical technique can also be used on item level metrics. An additional classical test theory approach to calculating item discrimination and difficulty metrics is available. Scherbaum et al. (2006) calculated the item difficulty as the mean score per item reported together with the standard deviation, and the item discrimination was the corrected item to total correlation. This study used this classical technique to provide a comparison between traditional metrics and IRT.

IRT is an augmentation to classical test theory. Typically, a combination of classical test theory and IRT techniques might be required to more fully understand an assessment at an item, facet, subscale, or entire instrument level (Scherbaum et al., 2006). IRT is an alternative theory or set of models that supplements many of the limitations of classical test theory. IRT can truly separate the participants from the item parameters, provides greater invariant information about participants' latent traits, and can provide precise descriptions of the function of each item over the range of ability being tested (Zagorsek et al., 2006). However, IRT depends on the construct validity of the underlying test with regards to the latent traits and, therefore, augments rather than replaces classical test theory (Ostini & Nering, 2006). Because the factor structure of the MLQ cannot be

assumed, classical test theory was employed in this study to determine dimensionality of the MLQ before proceeding to IRT analysis.

Item Response Theory

IRT predicts responses to items on an assessment based on specific models. IRT comes out of decision theory and logistic regression and is a set of models that estimate the amount of a latent trait possessed by respondents on an assessment (De Ayala, 2009). Because latent traits cannot be directly observed and measured, assessments are an indirect method of determining the amount of latent trait the examinees might possess (Smith et al., 2007). The degree of precision in predicting the amount of a trait, such as mathematical ability, intelligence, or leadership, possessed by an individual is assumed to be a direct reflection of the responses to assessment items, together with the model's predictability of matching those responses to an ability level (Emberson & Reise, 2000). It is these models of a person's trait prediction that is the subject of IRT. Samejima (1969) showed that IRT models response behaviors to an item or series of items to predict the amount of latent ability respondents possess.

IRT models are based on an observed phenomenon in testing. In developing a common intelligence trait scale for multiple Binet tests across separate age ranges, Thurstone (1925) documented a repeating pattern of sigmoidal responses. In Thurstone's Binet test analysis, when the x axis was scaled to chronological age and the y axis was scaled to the proportion of children with a correct answer to a test item, the response pattern was ogive in shape. Generalizing to other trait tests, the same sigmoidal pattern was evident when replacing the chronological age on the x axis for standardized test

scores (Thurstone, 1925). Ogive is an s shape, in which one end is concave and the other end is convex but the ends of the sigmoid shape finish parallel to each other. This shape is similar to a cumulative distribution curve (De Ayala, 2009). The normal ogive pattern forms the nonlinear basis for IRT models depicting item responses to latent trait assessments (De Ayala, 2009). IRT models may use a log metric or normal ogive metric. The scale multiplier value, signified by D , is 1.0 for normal metric scale and 1.702 for logistic scale (Embretson & Reise, 2000). This early work by Thurstone later evolved into the law of comparative judgment and was foundational to decision theory (Thurstone, 1927). Thurstone's early work on choice probabilities was further developed by Lazarsfeld and Robinson (1940), Rasch (1966), Lord (1968), and Samejima (1969), into what now is known as a set of models in IRT for predicting responses to various latent trait assessments.

IRT and logistic regression both estimate respondent's relation to dependant variables. From a logistic regression standpoint, IRT is similar in that logistic regression is concerned with predicting the probability of grouping respondents into categories (Tabachnick & Fidell, 2007). The extrapolation is that IRT predicts the grouping of respondents into categories of an item on an assessment (De Ayala, 2009). In multiple choice assessments, such as for math ability, in which there is a single right category choice and several wrong choices, logistic regression is employed to predict grouping respondents into right and wrong groups. Tabachnick and Fidell (2007), showed the general logistic regression formula is $\hat{Y}_i = e^u / (1 + e^u)$, where \hat{Y}_i is the estimated probability of one category as opposed to other categories of the i th case. This estimated probability

is a nonlinear function. The term e is a constant with the approximate value of 2.72. The term u is the typical linear regression equation such that $u = B_0 + B_1X_1 + B_2X_2 + \dots + B_kX_k$, where B_0 is a constant and B_0, B_1, \dots, B_k are coefficients, and X_1, X_2, \dots, X_k are predictor variables.

IRT and logistic regression are related mathematically. There is a striking similarity with all IRT model formulations to the general logistic regression equation. For instance, in Samejima's (1969) GRM used in this study, the probability of cumulative attraction to a category boundary of an item is described by the equation $P_{ig}^* = \frac{e^{Dai(\theta - \delta)}}{(1 + e^{Dai(\theta - \delta)})}$, where the linear regression equation is substituted by a nonlinear category boundaries equation (Ostini & Nering, 2006). P_{ig}^* is the probability of responding affirmatively at a category boundary with all lower ranked categories conditional on the latent trait θ for item i at category g . In this model e is the same constant, D is the metric scale constant, a_i is the discrimination parameter for item i , θ is the person latent trait ability, and δ is the item difficulty parameter (De Ayala, 2009). The logistic regression association to IRT is shown primarily as a conceptual link with techniques that may be more familiar. Both IRT and logistic regression classify respondents into groups. Tabachnick and Fidell (2007), stated that logistic regression, like other classical test theory techniques, rested on the assumptions of linear combinations of normally distributed variables and De Ayala (2009), showed that IRT was not linearly dependent.

Misspecification of IRT models is the norm. Before describing IRT further, a significant limitation must be noted. IRT models are more tightly constrained or specified than equivalent instrument representations in confirmatory factor analytic models (De

Ayala, 2009). Although the assumptions in IRT are more realistic, as they model specific response behaviors, the tighter constraints mean that misspecification is the norm, resulting in degraded parameter estimates (Kirisci et al., 2001). The degree of misspecification as it relates to estimate precision can be seen in the standard error of measure (De Ayala, 2009). These standard errors are additive across the items in a test and can significantly degrade the test information function of the instrument (Reckase, 1979). The efforts in correctly specifying the model are thus rewarded with greater reliability precision (Russell, 2002). This study will detail the model specification steps in the Methods section.

IRT Models

IRT models were designed to predict response behavior on specific types of assessment items. As the choice of model is critical to item and person parameter estimation it may be useful to explore IRT model taxonomy (Ostini & Nering, 2006). The various IRT models can be grouped by the type of item. For instance, a dichotomous item requires either a no or yes answer or is scored on an incorrect or correct basis. These dichotomous items can be modeled in IRT using the Rasch model or a modified Rasch model using one, two, or three parameters (Ostini & Nering, 2006). The Rasch one parameter model allows the difficulty of items to vary while holding the discrimination of all items to the value of 1.0. A two parameter model allow variations in both difficulty and discrimination across items. A three parameter model adds a guessing parameter to account for this behavior (Baker, 2001). According to Ostini and Nering (2006), multiple choice items are examples of dichotomous items since only one choice is correct.

Response behavior changes markedly when confronted with multiple correct or partially correct categories of an item. In addition to dichotomous items, some assessments employ polytomous items having two or more correct responses. Polytomous items have either continuous scales or categorical scales. Polytomous continuous models represent graphic rating scales, in which an examinee marks a response on a labeled continuum (Noel & Dauvier, 2007; Samejima, 1973). Woods (2008) stated that the polytomous categorical models apply to items, which have multiple correct or partially correct choices, of which only one choice is selectable.

Some scales increase in strength as choices are considered. Polytomous categorical models have been derived for discrete ordered responses, in which the response categories are ranked in increasing order of score value. Discrete ordered responses are further subdivided into heterogeneous and homogeneous models. Heterogeneous discrete ordered models are often applied to generalized partial credit scales, in which scores are based on incorrect, partially correct, and correct responses (Penfield & Bergeron, 2005).

An example of a generalized partial credit item is a geographic knowledge quiz with an item asking where London is located. The choices of this item may be Europe, England, Russia, and Dublin. In a partial credit score, Dublin and Russia would be wrong, Europe would be partially right, and England would be right. This ranked order may not represent how the items are displayed on the actual test; instead it refers to how the item is scored (Ostini & Nering, 2006). The heterogeneous nature of this example is in the mix of continent, country, and city concepts.

The homogeneous discrete ordered model applies to Likert type scales. An example of an IRT model for Likert type responses is known as the GRM by Samejima (1969). The MLQ is an example of a ranked homogeneous categorical polytomous instrument. The MLQ employs a consistent behavioral and attribution observation scale from zero to four. The scale for all items is anchored with zero being *not at all* to four being *frequently, if not always* (Avolio & Bass, 2004a). If the transformational leadership subscale involves a monotonically increasing relationship between the ordered categories and the latent trait, Samejima's (1969) GRM is one of the IRT models that can be used (Ostini & Nering, 2006). Another IRT model is Robert's (2008) GGUM. This model assumes the responses are not monotonic in nature (Scherbaum et al., 2006). Scherbaum et al. (2006) described this model as an ideal point response, in which the category chosen is the closest subjective match between a respondent's belief and the latent trait.

New models continue to appear in literature. There are many other models that can apply to ranked homogeneous categorical polytomous assessments (Ostini & Nering, 2006). A definitive selection criterion does not exist for choosing the appropriate model for a particular analysis (Embretson & Reise, 2000). Therefore, researchers sometimes compare at two different models for best data to model fit (Scherbaum et al., 2006). The degree of data to model fit, an issue of functional form, is discussed further in the Methods section.

Conceptual Basis for GRM

In the GRM, respondents are attracted to incrementally stronger stated categories. A short review of the underlying concepts and associated mathematical equations

describing the GRM may be useful to appreciate the model's utility. The first concept in Samejima's (1969) model is that participants answering a questionnaire, with a Likert type scale, are unequally attracted to the offered scale categories of each item (Samejima, 1969). In an ordered response pattern, such as a Likert scale, participants become increasingly attracted to higher categories until a category is selected (Ostini & Nering, 2006). Selection of categories is then a cumulative probability for the selected category and those categories higher in the latent trait. The MLQ presents numerous leadership behavioral items asking respondents to rate the frequency of observed behaviors on a 5-point Likert scale. A response in the lowest category, *not at all*, indicates that the rated leader did not exhibit sufficient leadership behaviors for a response to be recorded in that item category. The highest category represents that the observed leadership behaviors happened *frequently, if not always* (Avolio & Bass, 2004a). Rating leaders who possess high leadership abilities will typically attract the raters to choose higher categories on the Likert scale for the items that correspond with the observed leadership behaviors. In this way, Ostini and Nering (2006) showed that each item's ordered category, in the increasing Likert scale, differentially attracts a rater's response.

It is the point between any two sequential categories of an item that a response decision is made. A concept that is important in the GRM is that of category boundaries (Ostini & Nering, 2006). In a Likert scale, a category boundary exists between any two adjacent choices of an item. For instance, there exists a boundary condition between the lowest two Likert responses on the MLQ. The lowest anchor is *not at all*, and the second

lowest is, *once in a while*. There are also four category boundary conditions in any 5-point Likert scale.

There are some mathematical simplifications that accompany some IRT models. In the graded response probability equation, the probability of selecting the lowest or higher categories equals 1.0 and the probability of selecting higher than the highest category is 0.0. The 0.0 value occurs, for instance, because the lower boundary of the lowest category in a Likert scale is theoretically negative infinity. Therefore, the probability of choosing the lowest category or all higher categories is $P_{i_0}(\theta) = 1$. The probability then decreases from this 100% probability value as the assessment participant examines categories higher than the extreme lowest boundary condition. Ostini and Nering (2006) showed that this provides the distinctive monotonically decreasing probability curve from one to zero of the first category of a Likert scale.

The mathematical simplification applies at both extremes of response curves. In a similar fashion, choosing higher than the highest category in the Likert scale is $P_{i_{g+1}}(\theta) = 0$ where $g+1$ is one higher than the total g categories (Ostini & Nering, 2006). The probability of selecting the highest category equals the probability of selecting the highest category boundary $P_{i_g}(\theta) = P^*_{i_g}(\theta) - 0$ (Ostini & Nering, 2006). In practical terms, the probability of selecting the highest category increases as all previous categories have been rejected, therefore, the probability of the highest category being selected approaches 1 (Samejima, 1969). This probability function provides the distinctive monotonically increasing curve from zero to one for the last category of a Likert scale (Samejima,

1969). The categories in between the first and last choice appear as probability curves somewhat analogous to normal distribution curves.

An item's discrimination acts as a differentiator of respondents based on the latent trait being measured. In ranked homogeneous polytomous cases, such as with Likert scales, the item discrimination parameter a_i is held constant for each category of item i . Discrimination parameters, however, can vary across items. This restriction should make intuitive sense as the MLQ scale's lowest category, *not at all*, to the highest category *frequently if not always* reflects one response to the same behavioral leadership statement. There is no additional information a participant is asked to evaluate as there was in the heterogeneous example of geography. In that example, a participant not knowing that Dublin was a city would be differentially tested on that category choice and therefore the item would be modeled allowing discrimination parameter to be free between categories.

Mathematically, categories are assumed to be equally discriminating for the GRM. Slopes of all categories boundaries within an item are held constant for GRM, (Samejima, 1973). This constant slope is the same as keeping the discrimination parameter constant for an item across all category boundaries. However, the discrimination parameter is allowed to vary from one item to another (Samejima, 1973). Further, the boundary locations of each category within an item δi_g are at the point on the latent trait scale, in which the probability is $P = .50$ (Samejima, 1973). No term in the mathematical IRT formulas for probabilities included more than one participant's latent ability θ . These probability functions allow precise prediction of person, category, item,

and test response probabilities without dependence on other participants or an entire sample (Ostini & Nering, 2006). According to Thissen and Steinberg (1988), independence of sample is a major advantage of IRT over classical test theory.

Information about an assessment is the simple sum of the item information. Category information $Ii_g(\Theta)$ for an item of a latent trait Θ equals the negative second derivative of the log of the probability function of category g (Samejima, 1977a). Further, item information function (IIF) is the sum of successive categories of the square of the first derivative of the probability function over the probability function or $Ii(\Theta) = \sum$ from $g = 0$ to m of $P'i_g(\Theta)^2 / Pi_g(\Theta)$ (Samejima, 1977a). Test information function is the simple sum of each IIF (Ostini & Nering, 2006). Therefore, all IIF's are independent of participants who provided independent reliability information on the items and overall questionnaire (Thissen, Steinberg, & Gerrard, 1986). Because each of these probability and IIFs is conditional on the underlying latent trait Θ , Zagorsek, Stough, and Jaklic (2006), used IRT to explore the degree of difficulty each item presents in terms of the latent trait, such as transformational leadership.

Conceptual Basis for GGUM

The GGUM model assumes only one point on the scale is optimally attractive to those whose ability is below or above that point. Robert's (2008) GGUM incorporates both subjective and objective responses to items (Scherbaum et al., 2006). The objective response is the category selected. The subjective responses come from two different respondent groups representing a bias from below and above the selected category (Roberts & Sim, 2008). Typically, the GGUM model is used with a Likert scale anchored

between “*strongly disagree*” and “*strongly agree*” where the bias toward an item may be more directly excited (Roberts, 2008). In this study, a subordinate might have been negatively inclined towards their leader. Suppose the MLQ behavioral statement for a transformational leadership item should elicit a *sometimes* response due to the leader’s objective behavioral frequency. However, the unfavorable subordinate may approach the determination of response from the lower, *once in a while* category choice, since this subordinate subjectively wishes to rate the leader more critically. On the other hand, a subordinate positively disposed towards the same leader may approach the *sometimes* category selection from the higher, *fairly often* category choice, due to a favorable bias. According to Scherbaum et al., (2006), Robert’s (2008) GGUM accommodates these differences in selection approach and are called ideal point response models.

The GGUM mathematical equations are more complex than with the GRM. Roberts and Shim (2008) should be referenced to explore the mathematical equations for the GGUM. Both Samejima’s (1969) GRM and Robert’s (2008) GGUM produce a discrimination parameter per item. The key difference, from the GRM output results, is that a number of subjective category threshold parameters, $\check{\tau}_{ij}$, are produced. The number of threshold parameters equals one minus the number of categories of an item (Roberts & Shim, 2008). $\check{\tau}_{ij}$ are the subjective boundary locations for item i , category j , between response choices associated with that item’s difficulty location. These subjective category thresholds are defined as $\Theta - \delta = 0$ at the threshold location. Roberts and Shim showed that knowing the threshold values $\check{\tau}_{ij}$ and difficulty location value for an item allows the

calculation of the person ability level; below which the lower category is selected and above which the next highest category is chosen.

GGUM is a newer model but based on the GRM and Muraki's (1992) generalized partial credit model (Scherbaum et al., 2006). The interest in using this model for the rater's version of the MLQ is that raters may approach their leader's evaluation nonmonotonically. Scherbaum et al. (2006) found that the GGUM described the Leader-Member Exchange scale responses better than the GRM, indicating that self-reporting leadership assessments involved this unfolding subjective response behavior. This model has had limited use in rating leadership behaviors (Scherbaum et al., 2006). Using Robert's (2008) GGUM in this study could further research on unfolding IRT models.

IRT Parameters

At the heart of IRT is an estimate of item parameters. For the MLQ, the IRT analysis provides an estimate of each item's difficulty parameter along the trait scale, once item calibration has been completed. Measures of item characteristics include a standard error of measure, so a metric of precision is retained. In the case of Leadership Practice Inventory, Zagorsek et al. (2006) found that most items were easy to moderate in difficulty and therefore did not adequately test participants with higher leadership abilities. Because IRT analysis is not available for the MLQ, it is not known the degree of item difficulty for every item within the questionnaire. This study provided item difficulty and discrimination estimates for each of the 20 transformational leadership trait items.

Item discrimination parameters can also be calculated for each item. High discrimination, in the context of an instrument, is an item that precisely separates respondents of lower trait ability from upper trait ability (Samejima, 1977a). Once difficulty and discrimination estimates are known for an item, they uniquely identify the item; invariant from the influence of sample characteristics, administration, or other items on the instrument (Thissen & Steinberg, 1988). Due to this sufficient statistical property, the characterized item can then be eliminated if redundant or combined with other items from different tests, whose parameters are known, to form a new instrument testing the same latent trait (Action, Kunz, Wilson, & Hall, 2005). Using item characteristics is a common IRT process, according to Smith et al. (2007), for new test construction or existing test revision.

An advantage of IRT analysis is the ability to predict the latent trait of a participant, independent of other participants (Orlando & Marshall, 2002). Once item calibration has been completed, IRT analysis can estimate a person's location on a latent trait (De Ayala, 2009). This ability to use IRT to predict a person's latent trait means that, relative to others with this trait, the person being assessed is likely to respond in a particular way due to their trait value. In the MLQ, this means that transformational leadership ability can be reduced to a single value for each person being evaluated. In the rater's version of the MLQ, this means the rater's perception of their leader's transformational behaviors. On the IRT trait scale, a higher θ value indicates greater underlying trait ability (Zagorsek et al., 2006). Ranking and selection of transformational leaders would be possible with appropriate IRT analysis and established cutoff values

(Zagorsek et al., 2006). Although no transformational leadership cutoff levels have been reported for selection using the MLQ (Avolio & Bass, 2004a), IRT provides the analytical basis for such an approach.

Finally, IRT can be used to detect differential item functioning or item bias (Thissen et al., 1986). When two or more groups answer the same item differently there is often a concern that more than one trait is confounding the measure (Teresi & Fleishman, 2007). For instance, gender may influence the responses to transformational leadership behavior statements on the MLQ for females differently than males, therefore are reported differently. IRT can detect these disparities and determine whether the difference is uniform or nonuniform across the latent trait range (De Ayala, 2009). A uniform bias means that a group is affected in a consistent, negative or positive, manner. A nonuniform bias means that a group may be positively affected in part of the trait range and neutral to negatively impact in other parts of the range (De Ayala, 2009). Orlando and Marshall (2002) showed how IRT can aid in detecting these differential responses, once the groups are identified and analyzed.

Assumptions of IRT

As in classical test theory, important assumptions underlie IRT that are not always met. There are four main assumptions underlying IRT: unidimensionality, local independence, functional form, and testability (De Ayala, 2009). The first assumption is that the latent trait being examined is unidimensional, which means that only one continuous ability or latent trait is measured for a set of items within an IRT analysis (De Ayala, 2009). In the MLQ, using the entire instrument in a single IRT analysis would

likely violate the unidimensionality assumption since there are potentially nine factors or dimensions to the assessment. One way to overcome this issue is to separate the various factors or dimensions and perform an IRT analysis on each factor separately (De Ayala, 2009). For a full nine factors, this would mean performing nine IRT analyses, each containing four facet items. As noted, however, IRT analysis on transactional and laissez-faire subscales would require the use of IRT models that did not monotonically increase (Embretson & Reise, 2000). The Method sections of this study will explore in more detail this issue of satisfying the IRT assumption of unidimensionality.

The second assumption for IRT is that of local independence (Grimm & Yarnold, 2000). There must be sufficient statistical independence in responses to any two or more items of an assessment. More specifically, local independence is fully specified by the IRT model so that the latent trait is the only relationship between any two items or any two responses to an item (Scherbaum et al., 2006). In the case of the combined calibration samples for this study, this local independence assumption was violated by including responses from subordinates of the same leader. Responses for members within the same subordinate group introduced a relationship other than the transformational leadership trait being measured by the IRT model. In the same way, taking a sample from the same organization with a strong culture might violate local independence as the relationship with the organization might influence the responses to transformational leadership items. For this study, local independence was related to unidimensionality in the sense that factor analysis detected and assigned variation of item responses to one or more factors. Therefore, to the degree the subscale measured a single dimension; the

local independence assumption was satisfied. However, unidimensionality is not, in itself, sufficient to satisfy local independence (Embretson & Reise, 2000). As noted, MULTILOG software was robust to some violations of unidimensionality (Kirisici et al., 2001); however, local independence conditions were not specifically tested.

The third major assumption for IRT is that of functional form (De Ayala, 2009). Essentially, the data must conform to a specific model fit within a sampling error. Often this assumption is implied rather than stated because one of the steps in any IRT analysis is to perform a data to model fit analysis (Dragow et al., 1995). However, this assumption should be made explicit, as IRT is model dependent (De Ayala, 2009). As noted, the MLQ's Likert scale suggests the use of Samejima's (1969) GRM and Robert's (2008) GGUM. With so many context variables affecting participants' responses it was expected that functional form determination would show poor data to model fit. This issue is discussed more completely in the Research Method chapter.

The fourth and last major IRT assumption is of testability. Testability assumes that there are sufficient responses across all categories of all items. Sufficient responses are a minimum of five per category, for meaningful estimates of item and person parameters (Edelen & Reeve, 2007). For an extreme example, response patterns in which all answers to all items for all participants were wrong or equally, all answers to all items for all participants were right, is rather useless. It means the instrument was too hard or too easy, respectively. Another way of stating this is that the matrix of item by categories must have sufficient cell frequencies for useful analysis. Likewise, there needs to be sufficient items and participants to arrive at a calibration with reasonable sampling errors

(De Ayala, 2009). The appropriate size of the calibration sample is discussed in detail as part of the Methodological Considerations within this chapter.

Questionnaire Development and Refinement Using IRT

Using IRT in the construction of new instruments or reanalyzing existing instruments with IRT analysis seems to be on the increase. From 1925 to 1979, EBSCO databases showed 29 articles that the term *item response theory* was incorporated. That number had increased to 884 by 1989, 3,231 by 1999, and 9,101 by the summer of 2009. Of these IRT articles, over 20% described psychometric development of instruments.

Although the movement to use IRT with instrument analyses is currently fairly broad, this was not always the case (Edelen & Reeve, 2007). Due to the complexities of IRT and a historical limitation in available software, IRT was used in specific domains using simple IRT models such as: early formulation in the educational field by Thurstone (1925), personality studies by McArthur (1956), and applied psychology by Rosen and Rosen (1955). Thissen and colleagues (Thissen et al., 2003; Thissen & Steinberg, 1988; Thissen et al., 1986; Thissen, Steinberg, Pyszczynski, & Greenberg, 1983), did much to disseminate the use and applicability of IRT to additional disciplines with the introduction of software and useful articles, which supported multiple uses for IRT.

New instruments are published using IRT analysis. IRT has been used for new instrument constructions in the areas of leadership (Craig & Gustafson, 1998), general psychology (Cox & Sergejew, 2003; Mayers, Khoo, & Svartberg, 2002; Rauch, Schweizer, & Moosbrugger, 2008), legislation (Clinton & Lapinski, 2006), and the health care field (Reeve et al., 2007; Smith et al., 2007). But IRT analysis is not confined to new

instruments. A body of literature is devoted to IRT analysis on existing instruments. For instance leadership practices inventory (Zagorsek et al., 2006), 16PF (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001), NEO-PI-R (Reise, Smith, & Furr, 2001), PTSD checklist (Orlando & Marshall, 2002), HPI (Davies & Wadlington, 2006) and TAT (Blankenship et al., 2006) were all relatively well known instruments that were retrofitted with IRT analyses. There were even some new instruments that were created from the items of several tests with the desired item characteristics (Acton, Kunz, Wilson, & Hall, 2005; Chernyshenko et al., 2001). Finally, Bjorner, Chang, Thissen, and Reeve (2007) developed computer adaptive instruments that were based on a pool of questions with selected item characteristics.

IRT is used to evaluate existing assessments. As various applications of IRT demonstrate, there are two primary uses for IRT is questionnaire development and refinement. The first purpose is simply identifying the distinguishing characteristics of each item, such as in this study. This type of exploratory study is depicted using item parameters of discrimination and difficulty, item characteristic curves, and IIFs. From this characterization, the researcher can comment of the potential applicability of the items for development, evaluation, and selection. This type of IRT evaluation still requires large sample sizes, because item parameter information is dependent on sufficient response vectors (Orlando & Marshall, 2002; Wright, 1977). The current study characterized item and person parameters rather than altering the MLQ, so that items outside of the current MLQ were not be introduced. The adequacy of sample sizes is discussed in the Methodology Considerations section of this chapter.

IRT is also used to revise assessments. A second type of IRT application with assessments is one of elimination, substitution, or adaptation of items. With elimination of items, an IRT analysis is performed on the existing items in a questionnaire, characterized, and those that are redundant or do not add significantly to the latent trait information are removed (Zagorsek et al., 2006). Substitution of IRT characterized items is common practice in standardized testing (Reise & Waller, 2003). A large pool of questions is developed, in which the characteristics of each item are known with great precision (Smith et al., 2007). This pool of items allows substitution of equivalently difficult and discriminating items for any two or more persons taking the same administration of the test. The equivalent item substitution produces different tests that measure the same latent trait (Samejima, 1977b).

Construction of computer adaptive tests is an additional example of IRT item pool usage (Bjorner et al., 2007). This style of test uses a computer to select the difficulty and discrimination of the next question based on a correct or incorrect score on the current question. With this method a person's location on the latent trait scale can be quickly determined based on answers to precisely know characteristics of items, especially the discrimination parameters. A larger sample size is needed in calibration of each item in the pool for this type of IRT analysis due to the greater precision required in estimation of item location θ values and discrimination characteristics (Bjorner et al., 2007). Although this second usage of IRT analysis is not part of the current study, future researchers may seek to increase the information content along the transformational leadership subscale by using IRT to eliminate, substitute, or adapt specific items in the MLQ.

Summary Benefits and Limitations of IRT over Classical Test Theory

IRT analysis of the MLQ's transformational subscale can provide new psychometric insights of practical usefulness. IRT may furnish a number of benefits that have not been achieved using classical test theory alone. IRT analyses can furnish an estimate of a leader's transformational leadership ability independent of other leaders (Zagorsek et al., 2006). The degree of precision of a leader's transformational ability estimate is also available (Zagorsek et al., 2006). Further, the item difficulty and discrimination can be calculated independent of any specific sample characteristics or any other item (Blankenship et al., 2006). Therefore, an IRT analysis provides item reliability statistics that are free from influence of other items and respondents (Acton et al., 2005). Finally, information across all response categories is available for any item and for the entire test with precision measured by a standard error metric (Vidotto, Carone, Jones, Salini, & Bertolotti, 2007). According to Clinton and Lapinski (2006) and Samejima (1977a), these benefits are not available using classical test theory.

There are necessary cautions in using IRT. Limitations with the use of IRT stem primarily from models that are highly constrained therefore are subject to misspecification over and above that of classical test theory factor modeling (Drasgow et al., 1995). Unidimensional models of IRT are sensitive to dimensionality violations, which can reduce parameter precision (Kirisci et al., 2001). The dimensionality in turn is dependent on enough items within a dimension to make comparisons of scores meaningful (Reckase, 1979). Sample sizes required for calibration are typically in excess of those used in classical test theory (Edelen & Reeve, 2007; Fan, Thompson, & Wang,

1999). Conceptual and computational complexities of IRT have impeded broad use in psychological research (Smith et al., 2007). These limitations can be partially overcome through careful design and instituting recommendations from past research. This study reviews the recommendations from past research in the Methodology Considerations section of this chapter.

It was appropriate to use IRT for the MLQ's transformational subscale. The need for an IRT review of the MLQ over and above classical test theory rested upon the MLQ's prominent use in leadership research (Kanste et al., 2007). With widely varying factor structures and sample dependant internal reliability measures, the MLQ was not without criticism and modification attempts (Heinitz, Liepmann, & Felfe, (2005). However, IRT offered greater precision in item analysis and could lead to deeper understanding of psychometric issues with the MLQ and how best to resolve them. With greater assessment precision, detecting potentially harmful transformational leaders and adopting appropriate intervention strategies may be possible with earlier intercessions.

Leadership

Global and local leadership has gained high levels of interest and recognition. The selection of leadership books at retail stores are about various subjects: practical self-help volumes, examinations of individual leaders, and company performance under various leadership styles, to name a few. Due to the interest in the topic of leadership, scientists have been recognized for their contributions. For instance, author James Burns, renowned for his early leadership work, received the Pulitzer Prize and National Book Award for contributions to the field (University of Maryland, 2009). Avolio (2008) wrote a touching

eulogy for the MLQ co-author, Bernard Bass, a celebrated leadership authority, who received the Distinguished Scientific Contributions Award.

Leadership interest is increasing. Further evidence of research and application interest in leadership can be seen in the number of peer-reviewed publications dedicated to the leadership subject. For instance, *Leadership*, *Leadership Quarterly*, *Journal of Leadership and Organizational Studies*, *Leadership & Management in Engineering*, and *Nonprofit Management and Leadership* are journals that exist to convey leadership research knowledge and application to those interested in the field. Even in journals with broader interests, special issues about leadership can be found. For example, *Consulting Psychology: Practice and Research* ran a special issue in the winter of 2003 titled, "Leadership Development: New Perspectives." In January of 2007, *American Psychologist* ran a special issue titled, "Leadership." In November of 2007, *Applied Psychology: An International Review* published a special issue on the "Romance of Leadership." In the educational field, *Cambridge Journal of Education* completed a November 2003 special issue on "Changing the World of Leadership." Of the 24 issues of *Harvard Business Review* examined from January 2007 until April 2009, over 90% of the publications contained articles dealing specifically with leaders or leadership. Clearly, the topic of leadership continues to have relevance in business, education, and psychological communities.

Leadership research has had a personality emphasis. Psychological research on leadership has predominately focused on individual personality or style differences of leaders and their affect on followers (Hogan & Kaiser, 2005). Kaiser, Hogan and Craig

(2008) concluded that only a fraction of psychological studies dealt with objective measures of leadership performance outcomes. Economic literature, however, has focused predominantly on objective outcomes (Bloom & Van Reenen, 2007). One conclusion from economic journals was that management and leaders from first line supervisors through chief operating officers were acknowledged to have a significant effect on organizational culture, policies, practices, and performance (Bertrand & Schoar, 2003). Typical of outcome based economic literature studies, Bloom and Van Reenen (2007) investigated management disciplines across Western cultures including Germany, France, UK, and the U.S. at 732 medium sized manufacturing firms. The conclusion was that US companies were, on average, better managed than the European counterparts. According to Bloom and Van Reenen (2007), U.S. superior management practices accounted for half of the variation in performance on a number of objective metrics.

Leadership concepts are often contrasted with management functions. One of the distinctions that appear significant in the definition of leadership ability is in contrast to management ability. Avolio and Bass (2004a) defined leadership as getting subordinates to internalize a higher purpose than self-interest. Hogan and Tett (2003) defined management as directly supporting subordinates' self-interest by proffering rewards in exchange for specified performance. In psychological literature, the leadership ability has been termed charismatic or transformational and management ability was termed transactional (Avolio, Walumbwa et al., 2009). Terpstra, Mohamed, and Kethley (1999) believed that the desire to predict leadership potential or simply to separate leaders from

managers drove the leadership assessment industry to create multiple methods of testing potential leaders.

Leadership Questionnaires

Leadership assessments have enjoyed an established history. Standardized testing to differentiate leadership trait and ability levels has had a long history in the U.S. (Murray & MacKinnon, 1946). Typically, multiple methods are used to test for anticipated leadership performance levels (Thornton & Gibbons, 2009). Because some instruments are required by employers or have employment related consequences, the judicial courts have issued rulings on appropriate criteria for validity and reliability (Kleiman & Faley, 1978; Terpstra et al., 1999). Of the instrument types, questionnaires are frequently used in leadership assessments and vary widely in terms of content, length, approach, and intended purpose (Cole et al., 2006; Hogan & Tett, 2003). Broadly speaking, Yukl (2006) grouped leadership instruments by the types of traits or abilities that form the theoretical basis of the questionnaires.

The MLQ's transformational subscale is related to personality influenced behaviors. In the charismatic tradition, leadership has often been ascribed personality dimensions (Hogan & Tett, 2003). These personality traits can be tested using popular questionnaires such as NEO-PI-R by Costa and McCrae (1992) and the 16PF by Walter (2000). Other questionnaires exploring personality constructs are numerous and include Hogan Personality Inventory (1995) and Hogan Development Survey (1997) by Hogan and Hogan, California Psychological Inventory by Gough (1987), Eysenck Personality Questionnaire (1975), and the Guilford-Zimmerman Temperament Survey (1976). In

terms of research examining links between leadership and personality, extraversion was positively associated with leadership effectiveness and neuroticism was strongly and negatively associated with leadership (Judge & Piccolo, 2004). Neuroticism could be differentially detected by those interacting with leaders and was more strongly associated with subordinate ratings (Lim & Ployhart, 2004). Bono and Judge (2004) found leadership emergence had a higher association with personality constructs than effectiveness or charisma.

Some leadership assessments blend personality with cognition. Emotional stability and cognitive ability were combined in tests for Emotional Intelligence by Goleman (1995). Although the psychometric properties have been debated, the business sector seemed to utilize this questionnaire extensively due to face validity (Zeidner, Roberts, & Matthews, 2008). Conceptually, there seemed to be general acceptance that personality traits and cognitive abilities contributed to leadership emergence and subsequent performance making tests such as emotional intelligence readily accepted in business communities (Zeidner et al., 2008). Zeidner, Roberts, and Matthews (2008), also found that the debate over the use of Emotional Intelligence test for leadership performance predictability reinforced the need for multiple measures, which included cognitive and affective facets.

Lists of correlated items are sometimes provided to determine leadership. Inventory type questionnaires for specific knowledge, skills, and abilities in work situations are another approach to detecting and predicting leadership outcomes. Myers-Briggs type indicator (Myers, McCaulley, Quenk, & Hammer, 1998) and Campbell

interest and skills survey (2002) are two better known general work inventory questionnaires. A fundamental distinction examined by some of these inventories is the affinity for people and relational aspects of the work environment in support of task or end result orientations (Campbell, Hyne, & Nilsen, 1992; Pittenger, 2005). Researchers sometimes combine these inventories to predict leadership performance (Culp & Smith, 2005). Pittenger (2005) found that leadership inventories do not always hold the highest reliabilities and validity levels and therefore may require other measures to achieve appropriately supported predictions.

Other leadership assessments measure subjective determination factors. A class of questionnaires deals with motivational impetus to lead. For instance, Motivation to Lead by Chan and Drasgow (2001), considers cognitive ability, personality, and values as inputs to leadership motivation. The questionnaire examines affective, social normative, and noncalculative basis for assuming leadership responsibilities (Chan & Drasgow, 2001). An older and more general motivation questionnaire is the Thematic Apperception Test by Morgan and Murray (1938). The three motivational constructs measured are achievement, affiliation, and power (Langan-Fox & Grant, 2006). Van Iddekinge, Ferris, and Heffner (2009) showed that motivation to lead developed from personality attributes, especially conscientiousness, and the knowledge, skills, and ability to lead.

There are many forms of leadership behaviors and preferences around the world. Cross cultural leadership questionnaires may be of interest to researchers due to globalization of the workforce (Javidan & Dastmalchian, 2009). One such effort involved a coordinated testing regime in 62 cultures conducted to determine similarities and

differences among managers in various geographic regions, called the GLOBE project. A questionnaire was constructed along nine cultural dimensions and six leadership behaviors and attributes. The GLOBE project questionnaire was developed for many languages (Javidan & Dastmalchian, 2009). Over 17,000 participant managers were asked to complete the questionnaire. Country comparisons and dynamic intersections between cultures and leadership were reported. Javidan and Dastmalchian (2009) found that cross country investigations can aid in understanding how leadership varies with situational contexts and cultures.

Sometimes, specific working conditions or tasks require specialized testing. There are questionnaires that specifically target unique supervisory, managerial, or leadership behaviors and attributes. For the supervisory level there are questionnaires, such as supervisor behavior description questionnaire by Fleishman (1953), which came from early behavioral research at Ohio State University (Schriesheim, 1982). Managerial questionnaires include managerial practices survey by Yukl and Lepsinger (1990).

Specific leadership questionnaires are numerous and derive from different theoretical backgrounds. One such questionnaire is the leadership behavior description questionnaire from Stodgill (1963), which also came from early behavioral work at Ohio State University. Charismatic leadership has been examined using Conger and Kanungo scales (Rowold & Heinitz, 2007). Another leadership questionnaire, which has interested researchers, is the Leadership Practices Inventory by Kouzes and Posner (1988) based on neocharismatic or more commonly called transformational leadership theory (Carless, 2001; Posner & Kouzes, 1988; Zagorsek et al., 2006). Comparisons of charismatic and

transformational leadership questionnaires have shown high convergent validity (Rowold & Heinitz, 2007). Indeed, the most commonly used questionnaire specific to leadership is the MLQ by Avolio and Bass (2004a), from the neocharismatic or transformational research and the subject of the current study.

Multifactor Leadership Questionnaire

The MLQ is relatively easy to administer. Several researchers (Antonakis et al., 2003; Hinkin & Schriesheim, 2008a; Kanste et al., 2007) stated that the most widely used transformational leadership questionnaire for research was the MLQ (Avolio & Bass, 2004a). Authors Avolio and Bass suggested the 45 questions can be completed in 15 minutes and recommended for leader feedback, development, and selection (p. 2). Benefits of transformational leadership include a host of positive psychological and financial performance outcomes (Avolio, Walumbwa et al., 2009). Research using the MLQ has spanned over 25 years. According to Hunt (1999), the MLQ reenergized the leadership research area.

The MLQ is the preeminent assessment for leadership. Commercially available from Mind Garden, the MLQ in its various forms and translations has garnered unprecedented leadership research interest (Hunt, 1999). The claim of the MLQ as the most researched leadership instrument was confirmed using summer, 2009 searches of publication databases. In EBSCO databases, 297 articles involved the MLQ compared to 229 articles for all other leadership questionnaires combined. Of the 297 articles using the MLQ, PsycINFO contains 188 of those articles. These 188 articles on the MLQ compares

to 170 articles on all other leadership questionnaires. Therefore, the MLQ is the most studied leadership assessment as represented by articles in EBSCO databases.

The MLQ usage continues to grow. The pace of using the MLQ in research had consistently increased over the 29 years since Bass originally explored the concepts in a 1980 pilot study (Bass 1985; Bass 1997). The database searches revealed that, on average, the number of published articles more than doubled each decade. In the last five years, 92 articles were published involving the MLQ. In contrast, a competing transformational leadership assessment, the Leadership Practices Inventory, totaled 58 in the last five years.

Another competing transformational leadership instrument, the Transformational Leadership Questionnaire, is not much used. Developed in 2000 by Alimo-Metcalfe and Alban-Metcalfe, there were seven articles devoted to Transformational Leadership Questionnaire in PsycINFO through summer of 2009. The reason some researchers (Hinkin & Schriesheim, 2008b; Hunt, 1999; Judge & Piccolo 2004; Kanste et al., 2007) insisted that the MLQ was the most widely used leadership instrument for research, above all other leadership instruments, was due to the dominance of transformational leadership theory in the leadership field of study and the MLQ's predictive validity.

Transformational Leadership Theory

The constructs of transformational leadership are many and interwoven. Although leadership theories in general and transformational leadership theory in particular are not the focus of this research, having been studied extensively for over 25 years, it is useful to state the underlying assumptions of the MLQ. Detailed discussions of the MLQ's

constructs will assist in reviewing psychometric issues and this study's approach. Therefore, transformational leadership theory is reviewed from the perspective of operationalization of the transformational theory in the MLQ and how far the theoretical development has progressed. It is shown that the MLQ represents nine distinct facets of leadership style. As outlined by Heinitz, Liepmann, and Felfe (2005), all nine facets of the MLQ are rarely found as distinct factors in practice.

Explored in Burns' book on leadership (1978) and expanded by Bass (1985), the MLQ was constructed using the transformational leadership theory; sometimes called full range leadership theory (Yukl, 2006). This theory expanded to encompass three higher order distinct conceptualizations along a performance continuum, with transformational at the top, transactional at the midpoint, and laissez-faire at the bottom. The three conceptualizations are associated with nine distinct underlying behaviors and attributes or facets (Avolio & Bass, 2004a). The MLQ supports these three higher order concepts by operationalizing them into nine behaviors and attributes.

As shown in Figure 1, the full range leadership model is composed of three higher order leadership subscales comprising nine facets. Each of the nine facets are associated 4 items. Of the nine continuum facets, transformational leadership trait is operationalized in the MLQ as five facets: four behaviors and one attribute (Avolio & Bass, 2004a). It is these five transformational leadership facets embodied in 20 items that are marked in Figure 1 that are the focus of this study. Peterson, Walumbwa, Byron, and Myrowitz (2009) found that depending on the context, transformational leadership style can be the

most effective and satisfying style and may promote the greatest effort from followers at the top of the performance continuum from a dyadic relationship standpoint.

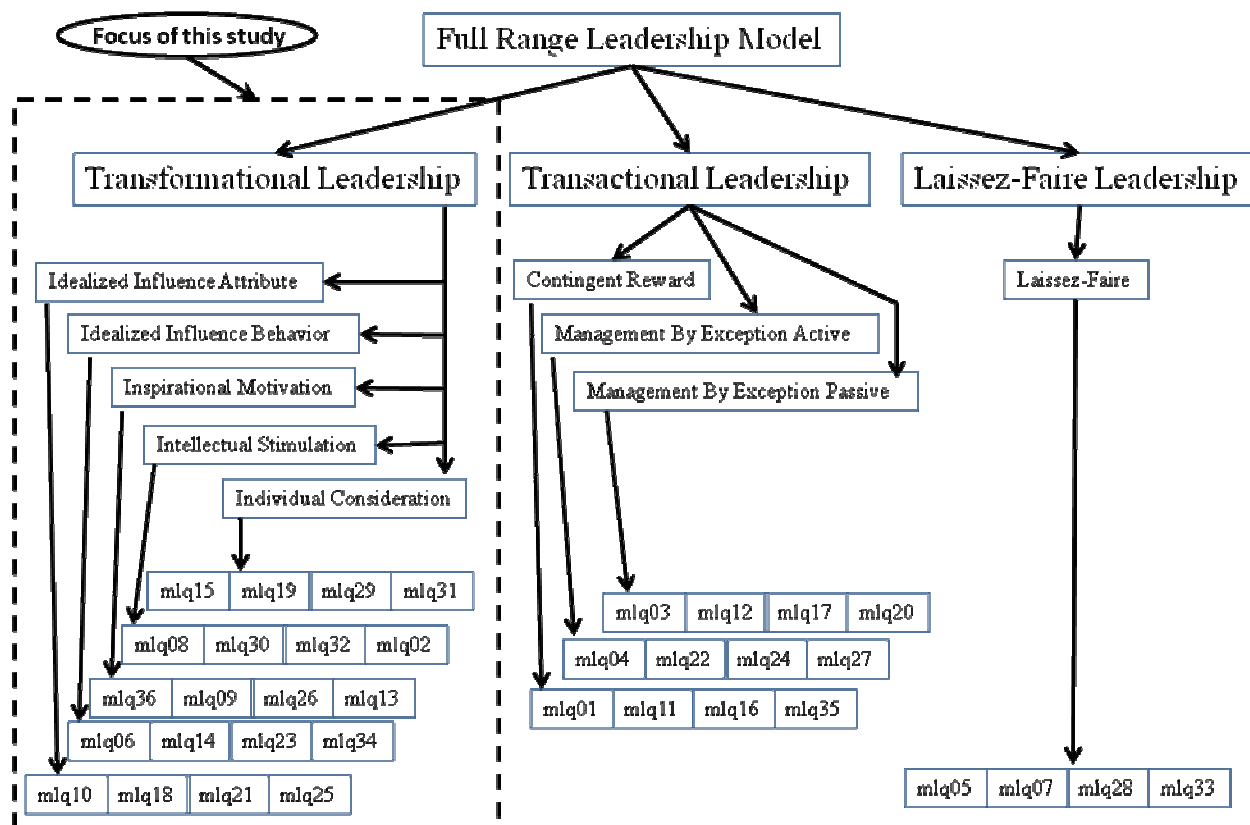


Figure 1. Full range leadership model: Showing the three higher order subscales, nine lower order facets, and 36 associated items of the MLQ.

Although initial expressions of full range leadership model focused on dyadic relationships, later modifications extended this concept to teams, groups, and to entire organizations (Avolio & Bass, 1995). In addition, there was evidence that transformational leadership was present across cultures, across organizational levels, across industries, and sectors (Bass, 1997). Full range leadership model appeared to be a prominent approach to considering leader and follower relations. With charisma

subsumed by transformational leadership, Lowe (2000) found that research articles encompassing full range leadership model have produced more leadership research than all other leadership theories combined.

Before proceeding with a detailed examination of each of the three higher order subscales, it may be useful to define some terms used in this study. The term *full range leadership model* and *transformational leadership theory* are often used interchangeably in literature to refer to all nine lower order facets and the term *transformational leadership* or simply *transformational* applies to the five transformational facets, represented by 20 items, examined in this study (Avolio & Bass, 2004a). To prevent confusion, full range leadership model is used to describe the 36 item set. Restricting the study to transformational items was discussed in the Limitations section of Chapter 1.

Facets refer to the factor structure at the lowest level of conceptualization. Factors and IRT dimensions represent conceptual and measurable latent traits, respectively (De Ayala, 2009). From an IRT perspective, dimensions represent the least number of factors emerging from an exploratory or confirmatory factor analysis above some criterion. Using the defined terms going forward, it may be helpful to discuss each of the higher order concepts and associated facets.

Transformational Leadership

At the top of the potential performance continuum is transformational leadership (Avolio & Bass, 2004a). In the dyadic relationship between leader and follower, it is the follower that is transformed by the leader. The leader through idealized influence, inspirational motivation, intellectual stimulation, and individual consideration, creates the

conditions in which the follower supersedes purely short term and self-interested goals for broader, higher, and nobler purposes through individual, group, or organizational objectives.

Avolio and Bass (2004a) made a distinction between socialized and personalized transformational leaders. Socialized transformational leadership benefited others as demonstrated by the leaders' self-sacrifice. Followers were transformed by emulating and internalizing the leaders' moral values, goals, and sacrifices toward a shared vision. The end result was followers who developed into leaders (Bass, 1985). Personalized transformational leadership was focused on the ego and power of the leader for personal gain (Bass, 1985). Followers quickly discovered the nature of this self-enhancing leadership and separated themselves from the consequences (Rowold & Heinitz, 2007). It was socialized transformational leadership behaviors, not personalized or self-glorification, that Bass (1985) designed into the MLQ.

Good in times of change or crisis, transformational leadership is about examining the current situation with new perspectives and different approaches. Peterson et al. (2009) found that transformational leadership was equally useful during steady state periods for developing new methods that radically reduced the cost or significantly increased the efficiency through implementing new processes. This type of leadership also increased the satisfaction of workers with their leaders and in turn, commitment to the organization (Barling, Weber, & Kelloway, 1996). Avolio and Bass (2004a) demonstrated through their research that transformational leadership was a source of human energy, driving growth, and change.

The MLQ was designed to test for idealized influence as two facets: as an attribute and a behavior (Avolio & Bass, 2004a). These two of the five transformational facets plus four other nontransformational facets, made for nine facets in total. Initially part of charisma (Bass, 1985), idealized influence was used to distinguish positive leadership from the negative side of charisma, which served only the leader in self-gratification but keeping the followers in a subservient role (Schyns, Felfe, & Blank, 2007). The MLQ separately tests for idealized influence attributes and idealized influence behaviors. These attributes include awareness by the follower of the leaders' self-confidence, self-sacrifice, and include the follower's desire to be associated with the leader. There is a heightened level of respect from followers who idealize their leaders. According to Avolio and Bass (2004a), idealized influence attributes were the follower's perceptions of the leader's ability to draw the follower into a heightened sense of collective contribution.

Idealized influence, as an attribute is the emotional facet of charismatic impact (Avolio & Bass, 2004a). Rowold and Heinitz (2007) found that idealized influence was the psychological attachment that a follower experienced when decision making power was transferred to the leader. This transference of authority, depended on the degree to which the leader clarified, developed, and promoted a higher sense of mission, was considered a supporting behavior (Schyns et al., 2007). Integrated into the mission were moral and ethical considerations. The followers were drawn into this higher sense of purpose, predicated upon the leader's well articulated sense of values and beliefs (Avolio & Bass, 2004a). These observed idealized influence behaviors provided the follower with

a perspective and invitation to be part of a significant undertaking, one that would benefit others more than self (Rowold & Heinitz, 2007). Idealized influence behaviors are the second facet of the transformational subscale out of a total of nine facets for the full range leadership model. Avolio and Bass (2004a) constructed idealized influence attributes and behavior statements to represent engendered trust, respect, and a desire for followers to emulate the leader.

The third facet associated with transformational leadership, is inspirational motivation (Avoilo & Bass, 2004a). Inspirational motivation along with idealized influence was once termed charisma (Bass, 1985). Casting a compelling vision, the leader approached the future with optimism and enthusiastically invited followers to participate in its completion (Bass, 1985). The confidence of the leader influenced the followers to believe that the vision could be successfully achieved (Berson, Shamir, Avolio, & Popper, 2001). These behaviors created in the follower, the motivation to put self-interest aside; to sacrifice with greater effort for the benefit of the articulated vision (Berson & Linton, 2005). What made this vision compelling was that the leader and the followers work to the betterment of others rather than themselves. Inspirational motivation could be likened to a noble cause. It was the affiliation of the follower with the leader and other peers who were inspired by the same cause (Rowold & Heinitz, 2007). Inspirational motivation is the third of nine facets, of the MLQ. If the first two behaviors of transformational leadership are about the influence the leader had to inspire the follower than Bass (1985) constructed the last two behavioral facets to represent how the leader developed the follower to succeed.

The fourth facet associated with transformational leadership, and tested through the MLQ, is intellectual stimulation (Avoilo & Bass, 2004a). Because problem solving is a key process in completing an objective, intellectual stimulation focused on how the follower conceptualized, analyzed, and approached complex problems (Rowold & Heinitz, 2007). If the follower had preconceived notions about how problems should be solved, often the most optimal solution stayed elusive. Instead, the follower needed to develop a broader perspective of the problem definition and multiple ways of approaching possible solutions. It was the leader's responsibility to develop the intellectual stimulation of the followers (Avoilo & Bass, 2004a). According to Bass (1985), the test of the degree to which a leader developed a follower in intellectual stimulation, was how well the follower performed on new situations in the absence of the leader.

The fifth and final facet associated with transformational leadership is individual consideration (Avoilo & Bass, 2004a). The concern by the leader, for the follower is expressed by an interest in all aspects of the follower. All of the abilities, hopes, aspirations, and fears of the follower are relevant to the leader, towards the management and development the follower. Rowold and Heinitz (2007) found that it was this genuine concern for the followers, which returned the respect and trust for the leader. Only through understanding the followers at such a deep level, could the leader influence the follower's perspective and elevate aspirations (Schyns et al., 2007). Individual consideration was designed by Avolio and Bass (1995), to be the mechanism that turned

the follower aside from self-interest to embrace a higher purpose, thus developing as a future transformational leader.

These five transformational facets of idealized influence attributes and behaviors, inspirational motivation, intellectual stimulation, and individualized consideration are part of the nine facets of the full range leadership model (Avolio & Bass, 2004a). It is only idealized influence that has an attribution facet along with the behavioral facet. Bass (1990) explained retaining the idealized influence attribution as the followers' emotive response that accompanied behavioral observation of the leader. Similar to charisma, idealized influence was not only how the leader behaved but also the effect on the emotions of the follower (Lim & Ployhart, 2004). Inspirational motivation, then, provided the direction to impel the followers into action. With intellectual stimulation and individualized consideration, the leader developed the follower through exercising greater autonomy and by achieving the vision (Osborn & Marion, 2009). The five facets of transformational leadership and the associated 20 subscale items are the focus of this study. All five facets may be used by a transformational leader, whereas in the transactional subscale, a leader may exhibit three separate and distinct facets (Avolio & Bass, 2004a). Avolio and Bass (2004a) described numerous differences between transformational and transactional facets.

Transactional Leadership

Perhaps the most familiar leadership style from early industrial psychology research is transactional leadership at the center of the performance continuum.

Transactional leadership, as the name implies, is the exchange of something beneficial

from the leader for compliance with expectations of performance from the subordinate, team, or organization (Avolio & Bass, 1995, 2004a). Organizational culture typifies this type of behavior with an expectation by the employee that regular payments is forthcoming in exchange for specific work activities (Bass, 1997). A leader might suggest that a particular reward such as a bonus is paid if a specific business goal is achieved by the subordinate. This inducement to perform is an example of transactional behavior initiated by the leader targeting an individual performer or group. In order for this constructive approach to work, the leader needs to clearly communicate the expected outcomes (Avolio & Bass, 2004a). The reinforcement of this transaction is then dependent on the followers' belief that the organization or leader is able to deliver the promised reward (Bass, 1990). In addition, the reward must be perceived as beneficial and desirable to the follower (Bass, 1985). The contingent reward facet has garnered a large amount of research by Bass, Jung, Avolio, and Berson (2003), Bycio, Hackett, and Allen (1995), and Judge and Piccolo (2004) in relation to transformational subscale to determine any additive effect.

Full range leadership model predicts that transformational leadership is an augmentation of transactional leadership (Bass, 1985). Transformational leadership does not occur without some level of transactional leadership behaviors. Contingent reward is required to build some level of trust and relationship between the leader and the follower (Judge & Piccolo, 2004). Upon this initial trust is built the individual consideration, the intellectual stimulation, the inspired motivation, and finally the idealized influence. The transformational theory is not clear whether other transactional facets are required before

transformational leadership behaviors could be observed (Purvanova & Bono, 2009). However, contingent reward behaviors are highly correlated with all transformational leadership behaviors (Avolio & Bass, 2004a). Judge and Piccolo (2004) found that transformational leadership created a more effective unit, was highly satisfying to the followers, and created greater effort towards achieving collective objectives.

Transactional leadership styles impacted those who responded to contingent reward reinforcement by putting in enough effort to gain the promised reward or avoid punishment for insufficient performance, according to Hinkin and Schriesheim (2008a). Once the reward was achieved or withheld, motivation to produce diminished until a new reinforcement is introduced. Contingent reward is the first facet of three transactional leadership behaviors.

The second facet behavior associated with transactional leadership is active management by exception (Avolio & Bass, 2004a). In this behavior, leaders actively seek to correct follower mistakes (Hinkin & Schriesheim, 2008b). With the objective of maintaining the standard operating procedures, leaders find fault with followers or publically comment on mistakes by followers and focus exclusively on these deviations. This type of behavior is less about improving performance than maintaining existing standards of performance. The role of the leader is seen as upholding the status quo. In high risk situations, such as leaders in underground mining operations, active management by exception might be seen by followers as critical and necessary for the followers' survival. Unlike contingent reward, Hinkin and Schriesheim (2008b) showed

that active management by exception behavior was focused on correction of mistakes made by followers.

The third facet behavior of transactional leadership is passive management by exception (Avolio & Bass, 2004a). If the leader stepped in and corrected a follower under this type of leadership, it is only because the follower's behavior was so egregious that the leader has no choice but to intercede. A leader's coercive action under this scenario might have been due to a concern that the leader would get punished if there was no intervention. A leader exhibiting passive management by exception only initiated correction if the actions by the follower were chronic and had significant consequences to the leader (Avolio & Bass, 2004a).

These three facet behaviors, contingent reward, active management by exception, and passive management by exception constitute transactional leadership in the full range leadership model. The goal of transactional leadership is to manage within the organizational bounds and ensure that the followers were conforming to clearly specified expectations (Hinkin & Schriesheim, 2008b). As such, transactional leaders' obligation to the organization is to enforce work and safety standards and performance goals (Avolio & Bass, 2004a). These three transactional facets add to the five transformational facets, which operationalized eight of the nine underlying constructs of full range leadership model. Of the three facets, Purvanova and Bono (2009) found management by exception active and passive amassed less research literature.

Interestingly, it was primarily with transactional rather than transformational leadership behaviors that were the source of research debate (Avolio & Bass, 2004a).

Although contingent reward was consistently and positively correlated with transformational leadership and improved outcomes, the two forms of management by exception have not (Bass et al., 2003). With all three are combined in the higher order transactional subscale, inconsistent higher order results were produced. Management by exception has been linked with both contingent reward and with laissez-faire leadership throughout research (Bono & Judge, 2004; Lowe, Kroeck & Sivasubramaniam, 1996, Tejeda, Scandura, & Pillai, 2001). This inconsistency makes including transactional items in this study problematic. However, the lack of agreement on appropriate factor structure cannot be resolved through IRT analysis, a predominantly item level approach.

To give relative positions of the three facets of transactional leadership, Judge and Piccolo (2004) provided a meta-analysis estimating true score correlations. Of the transactional facets, contingent reward had the highest positive effect (.39) followed by active management by exception (.15). Passive management by exception had a similar, though negative, effect (-.18). Although this general pattern found by Judge and Piccolo was similar to other researchers, contextual variations may have influenced active management by exception to be low to negative in strength, such as in broader organizational measures (Hinkin & Schriesheim, 2008a; Judge & Piccolo, 2004). Therefore, mindful of contextual differences generally, the transactional subscale was from strongly positive for contingent reward to weakly negative with passive management by exception (Judge & Piccolo, 2004). The most negative leadership style is considered a separate higher order construct, laissez-faire leadership.

Laissez-Faire Leadership

Laissez-faire leadership, or nonleadership behaviors, is at the bottom of the performance continuum. Unlike transactional leadership, laissez-faire leadership consistently reduced the output of the individual, group, or organization (Hinkin & Schriesheim, 2008a). In the MLQ, laissez-faire leadership is represented by avoidant behaviors. Laissez-faire is the most negatively producing leadership style within the full range leadership model (Avolio & Bass, 2004a). It is marked by the absence of all decision making or corrective action. It is as if the leader occupied the position with the associated power and privileges, but never acted (Avolio & Bass, 2004a).

Under most circumstances laissez-faire leadership reduced the motivation of employees to perform expected tasks (Hinkin & Schriesheim, 2008a). Unlike virtual teams or remote management, in which the length of time before a manager interceded, may be longer, a laissez-faire leader simply did not respond to their environment and withdrew from all responsibility (Howell, Neufeld, & Avolio, 2005). Further, a laissez-faire styled leader blocked followers from assuming the absent leaders' power and privileges and thus reduced the ability of the team to achieve desired goals (Howell et al., 2005). This type of individual was one who had no desire and no motivation to make a decision on behalf of their employees. Laissez-faire is the ninth and final facet in the full range leadership model.

With these nine conceptual facets, Bass (1985) and Avolio (2004a, 2004b), designed, constructed, and revised the MLQ over a 25-year period. The theoretical approach, research, and subsequent incorporation of findings into the MLQ's designs

were necessarily, an iterative process. As is demonstrated, Avolio and Bass (2004a) further developed the MLQ to reflect the research effort in stabilizing the MLQ factor structure.

The MLQ Development

Burns in his 1978 book on leadership, contrasted transformational leadership with the cause and effect approach that tended to dominate research literature called transactional leadership. Burns' original formulation of transformational leadership was conceptually opposite to transactional leadership. Embodied in archetypal political leaders, Burns envisioned transformational leadership as one that inspired followers to transcend their own self-interest to achieve higher goals. Although charismatic leadership had significant overlap with transformational leadership style (Schyns et al., 2007), Burns emphasized the societal good that could be achieved by devoting oneself to moral imperatives beyond self-interest.

Bass (1985) adapted this conceptualization of transformational leadership as an augmentation rather than in opposition to the transactional leadership style. In a further expansion to the political sphere used by Burns (1978), Bass envisioned transformational leadership as fundamental to all forms of leadership, regardless of organization or affiliation. Finally, Burns was writing in response to a perceived overemphasis on transactional leadership in research and thus wanted to juxtapose the transformational concept but Bass seemed less bound by this concern. Transactional leadership was therefore, something Bass (1985) expanded to include constructive, corrective, coercive, and absent leadership behaviors designated as active, passive, and avoidant in his model

(Avolio & Bass, 2004a; Bass, 1985). It was this conceptualization by Avolio and Bass (2004a), of the transformational leadership theory that was embodied in the construction and subsequent refinement of the MLQ.

The initial development of the MLQ prior to its publication in 1985, involved a 1980 pilot study using 70 male South African senior executives (Bass, 1985; Bass, 1997). The 1980, open ended survey led to construction of a pool of 142 items describing transformational leadership. Of these items, 73 were selected by consensus from 11 graduate masters of business administration and social science students. In turn, the 73 items were evaluated by 104 officers, primarily from the U.S. Army. Instead of an intensity scale, a frequency scale was used. In this scale, the 5-point score ran from A to E decreasing in behavioral frequency: *frequently, if not always, fairly often, sometimes, once in a while, and not at all*, respectively. This frequency scale had the ratio of 4:3:2:1:0, meaning that A., *frequently, if not always*, implied an observed frequency of behaviors four times that of D., *once in a while*. In later versions the scale would be reversed, increasing rather than decreasing, using a 5-point Likert scale from zero to four. The 73 items were supplemented by five demographic items and six additional result indicator items, totaling 84 items in the first published the MLQ. The result indicator items included tests of the leaders' effectiveness and follower satisfaction. The 84 item version of the MLQ was published in 1985 by Bass in his book, *Leadership and Performance beyond Expectations*.

Bass' book (1985) also contained correlational analysis and factor analysis applied to the 73 leadership items. Originally seven factors were found with eigenvalues

above one representing 89.5% of common variance. However, Bass also reported on a later study with a larger sample size, which when analyzed retained only the first five factors. These five factors were labeled, in order from highest loadings to lowest: charismatic leadership, contingent reward, individualized consideration, management by exception, and intellectual stimulation. Bass mentioned that inspirational leadership was a cluster found within the charismatic leadership factor. In later versions of the MLQ, the term *charisma* was replaced by idealized influence attributed, idealized influence behavioral, and inspirational motivation (Avolio & Bass, 2004a). However, during this early process, management by exception included some elements of nonleadership such as laissez-faire. In addition; Bass conducted a higher order factor analysis revealing two factors that were called, *active-proactive* and *passive-reactive* leadership. The passive reactive leadership included laissez-faire behaviors. Bass' published his baseline of the MLQ in 1985. After the initial publication, the MLQ has had a number of revisions based on continued research.

In the early stages of using the MLQ, the version naming convention seemed to follow a somewhat sequential nature with versions termed, *Form*, plus a number. For instance, Form 1, contained the 73 items and was published in the 1985 book. Form 2, contained 31 of the 73 items, also mentioned in the 1985 book by Bass. Form 4 contained 50 items, 10 from each of the five factors. Finally, there was mentioned an unidentified Form containing 37 items. These versions were all described in Bass' 1985 book.

As noted, by 1995 Bass had made the substitution of idealized influence attribution, idealized influence behaviors, and inspirational motivation for what in earlier

versions was called charismatic leadership. Management by exception had also been subdivided into separate active and passive facets (Bass 1990). Bass has termed the entire continuum, from transformational leadership to laissez-faire leadership, the *full range leadership model*. Avolio insisted that full range applied only to transformational behaviors rather than encompassing every conceivable leadership construct (Antonakis et al., 2003; Bass, 1997; Yukl, 2006). Careless (1998), Kanste et al. (2007), McAlearney (2005), and Tejeda, Scandura, and Pillai (2001) noted that the contextual variables such as environmental factors, organizational factors, participant variables, and personality traits that were assumed to exist by Bass (1985) as antecedents to transformational leadership were often ignored in research experiments.

Complicating the picture of inadequate experimental design (Judge & Piccolo, 2004), was the frequency and number of early version changes. Following the first publication of the MLQ (Bass, 1985), the research literature suggested that letter designations, whose meaning is unclear, followed some of the form numbers. For instance, Form 5 was often followed by an R or later by an X (Bass, 1997). There is also an 8Y version of the Form used in a Dutch translation of the instrument (Den Hartog, Van Muijen, & Koopman, 1997). Other versions noted by Antonakis et al. (2003) were Form X, Form 5S, and a 1990 and a 1993 version of Form 5X. Even Form 5X had multiple versions, in which items were rewritten or amended (Hinkin & Schriesheim, 2008a). Form 5X released in 1993 had 90 items; 78 items for full range leadership plus 12 outcome items. In the third edition manual (Avolio & Bass, 2004a), two versions of Form 5X were listed: Form 5X short with 45 items and Form 5X long with 63 items.

However, Form 5X long was not recommended for research purposes and was not included in the manual and sampler set (Avolio & Bass, 2004a). Instead, Form 5X long was to be used only for training purposes and development of those wishing to increase their transformational leadership behaviors. Avolio and Bass (2004a) designed Form 5X short to be used for testing the extent of transformational behaviors in organizations, individual leader feedback, evaluation, selection, and for general research.

The current Form 5X short includes four items for each of the nine facets of full range leadership plus nine items on leader efficacy, satisfaction, and extra effort making 45 total items (Avolio & Bass, 2004a). For Form 5X short, there is a leader version and a rater version. The leader version is a self-rating Form and the rater version asks for responses to named leaders. The rater evaluating a designated leader could be a subordinate, peer, supervisor, or someone the rater was sufficiently familiar with to indicate the observed frequency of certain behavioral responses. According to the manual by Avolio and Bass (2004a), these leader and rater versions of Form 5X had no separate designation.

Another practice that was common by researchers (Carless, 2001; Cole et al., 2006; Hinkin & Schriesheim, 2008a; Ling, Simsek, Lubatkin, & Veiga, 2008; Peterson et al., 2009; Schyns et al., 2007) was the study individual items or subscales rather than the entire MLQ. Additionally, other researchers (Den Hartog et al., 1997; Heinitz et al., 2005) altered the questionnaire by removing items to improve the factor structure. These modifications to the assessment were not adopted by other researchers, who retained the published MLQ (Avolio & Bass, 2004a; Judge & Piccolo, 2004). In addition,

transformational leadership facets were often combined and reported as one measure in research results (Peterson et al., 2009; Snodgrass, Douthill, Ellis, Wade, & Plemons, 2008; Walumbwa et al., 2008). Judge and Piccolo (2004) suggested that these assessment versions and inconsistent reporting of results impeded the advance of theory and practice.

In 2003, a new normative sample was introduced (Avolio & Bass, 2004a). Avolio and Bass described the data base as consisting of samples collected through 2000 and additional samples through 2003. A nine factor model, or full range model, was shown as the best fit despite rater differences or geographic differences. The goodness of fit of .92 and root mean squared error of approximation of .05 for the nine factor model was short of acceptable limits (Hu & Bentler, 1999). Other than rater type and geographic region, Avolio and Bass (2004b) did not provide demographic data to determine the extent of sample diversity. Further, Avolio and Bass (2004b) did not analyze moderating variables such as organizational type or leadership level, using the 2003 normative data.

The design, construction, and revision of the MLQ occurred over a 25-year period (2004a). Research findings have been extensive and varied (Judge & Piccolo, 2004). For heterogeneous samples, the MLQ has been inconsistent psychometrically. However, use of the MLQ has been increasing due to predictive validity (Antonakis et al., 2003; Wylie & Gallagher, 2009). From an IRT analysis perspective, examining these psychometric properties is useful.

The MLQ Psychometric Properties

The response to Bass' 1985 seminal work, proposing transformational leadership, reinvigorated the field of leadership research and provided over 25 years of extensive

research using the MLQ (Avolio, Walumbwa et al., 2009; Hunt, 1999). From the beginning, this research was international in scope, with the MLQ having been translated into at least 24 languages (Avolio & Bass, 2004a; Cole et al., 2006) conducted on every continent except Antarctica (Bass, 1997). The richness of research results has provided ample evidence of the weaknesses and strengths of the full range leadership model, as operationalized in the MLQ's three main higher order leadership subscales and nine lower order facets and outcomes. From an IRT analysis perspective, it may be useful to examine the reliability, construct validity, external validity, and predictive validity of the MLQ research results. Finally, these findings is summarized before proceeding to the Methodological Considerations section.

Reliability. Reliability of the MLQ is considered relatively stable (Kanste et al., 2007). Multiple versions were the source of some early reliability discrepancies (Eagly, Johannesen-Schmidt, & Engen, 2003). Kanste et al. (2007) reported internal consistency using Cronbach's alpha coefficient generally above an alpha level of .70 for all subscales Tejada et al. (2001) found, across the four samples, coefficient alpha levels for transformational facets averaged .90 ranging from .86 for idealized influence to .94 for inspirational leadership. Nunnally and Bernstein (1994) recommended minimum coefficient alpha levels above .90, preferable above .95, for decisions based on test scores. Studies of item total correlations were generally above .30 and inter item correlations ranged from .30 to .70 (Kanste et al., 2007). Cronbach's alpha has been the primary means of evaluating reliability (Judge & Piccolo, 2004). As noted, Cronbach's alpha is sample dependant as is not a precise or invariant measure of reliability. IRT analysis can provide

precise reliability parameters for items and persons. It was expected that this study would aid in determining the reliability of the 20 MLQ's transformational items. It was the MLQ's construct validity issues that posed practical difficulties for IRT analyses.

Construct validity. As has been noted, the MLQ has had issues with construct validity due to lack of clear convergent and discriminant evidence at the lower, nine facet level. Before 2003, researchers using exploratory or confirmatory factor analysis could converge upon no more than six of the factors (Avolio & Bass, 2004a). Neither could other researchers isolate all nine factors at appropriate statistical levels to validate the conceptual structure of the MLQ (Judge & Piccolo, 2004). Lack of agreement on factor structure had not been remedied by Antonakis et al. (2003), even with very large sample sizes.

The MLQ authors, (Avolio & Bass, 2004a) using 1999 normative data had come to conclude a six factor structure using a large aggregation of samples ($N = 56,479$). However, the model fit did not meet appropriate criteria (AGFI = .91, CFI = .91, RMSEA = .05) provided by literature (Hu & Bentler, 1999). With additional data representing 2003 normative samples, the nine factor structure was found (Avolio & Bass, 2004b), however, below acceptable model fit guidelines (AGFI = .92, CFI = .91, RMSEA = .05). Another attempt to find all nine theorized factors occurred when one of the MLQ authors, Avolio, joined Antonakis and Sivasubramaniam (2003) with a data driven approach to search for the conditions that favored a nine factor answer. They partially accomplished that task by isolating several moderating variables that resulted in a nine factor confirmatory analytic solution using years of archival data from Mind Garden, the MLQ

copyright holders. However, Antonakis et al. did not fully meet current model fit standards (CFI = .90, RMSEA = .04) leaving the nine factor solution in doubt. No other known researcher has attempted replication of the work by Antonakis et al.

Moderating variables. Each of the full range leadership model's nine facets is theorized to represent an independent latent trait with a distinct factor loading structure (Avolio & Bass, 2004a). However, analysis of the MLQ factor structure produced widely different results when moderators had not been taken into account. For instance, factor structures examined using exploratory and confirmatory analysis had been found to vary from one higher order factor (Carless, 2001) to nine lower order factors (Antonakis et al., 2003; Avolio & Bass, 2004a; Hater & Bass, 1988; Heinitz et al., 2005). Cole, Bedeian, and Field (2006) and Kanste et al. (2007) found additional factor structures.

Each of the studies used either an older version of the MLQ, before 2004 Form 5X short, or did not incorporate sufficient moderator variables to separate confounding influences (Antonakis et al., 2003; Avolio & Bass, 2004a). Lack of incorporation in the design of all the moderators is understandable due to the large number of variables that would need to be incorporated. It would also mean a sample size sufficient to separate variances of each main and interaction effect (Tabachnick & Fidell, 2007). Very large sample sizes and moderating variable analysis of the design envisioned is not always practicable and rarely been done, with the possible exception Antonakis et al. (2003) accessing years of Mind Garden's information. However, Hogan and Kaiser (2005), Hunt (1999), and Kaiser et al. (2008) stated that understanding situational influences was critical in leadership analysis as leadership remains a contextually based construct.

Bass (1985), Antonakis et al. (2003), and Osborn and Marion (2009) suggested that contextual variables moderate these distinct factor structures and thus must be designed into studies using the MLQ. These contextual variables investigated by Antonakis et al. (2003) and others were critically important (Osborn & Marion, 2009). Early emphasis by Bass (1985) that the MLQ results by themselves were limited without taking note of the context have led to numerous studies of possible moderating variables that influence the MLQ results (Antonakis et al., 2003; Hogan & Kaiser, 2005; Kaiser et al., 2008; Lim & Ployhart, 2004). The variables studied have been extensive including environmental risk factors (Antonakis et al., 2003) and geographic region (Avolio & Bass, 2004b; Bass, 1997; Cole et al., 2006) at the macro level, plus many internal organizational characteristics such as firm size (Eagly et al., 2003; Ling et al., 2008), founder status (Ling et al., 2008), organizational type (Antonakis et al., 2003; Bass, 1997; Hetland & Sandal, 2003; Lowe et al., 1996), stability (Antonakis et al., 2003; Felfe & Schyns, 2002), and cascading leadership (Avolio & Bass, 1995).

The greatest moderator research concentration, however, was on the numerous participant variables such as team cohesion and collective goal commitment (Cole et al., 2006), leader gender (Antonakis et al., 2003; Eagly et al., 2003; Hetland & Sandal, 2003), leader distance (Howell et al., 2005; Purvanova & Bono, 2009), leader personality (Bono & Judge, 2004; Hetland & Sandal, 2003; Judge & Piccolo, 2004; Lim & Ployhart, 2004), and level of leader (Antonakis et al., 2003; Avolio, Bass, & Jung, 1999; Lowe et al., 1996). In addition, leader nationality (Schyns et al., 2007), leader age (Eagly et al., 2003), length of leader and follower relationships (Avolio, Bass, Jung, 1999; Avolio & Bass,

1999; Howell et al., 2005), and leadership training (Barling et al., 1996; McAlearney, 2005; Wylie & Gallagher, 2009) have been examined.

Other variables have included rater tenure (Felfe & Schyns, 2002; Howell et al., 2005; Wylie & Gallagher, 2009), rater job function (Felfe & Schyns, 2002; Wylie & Gallagher, 2009), raters' relationship to leader (Hetland & Sandal, 2003), and follower identification and self-efficacy (Walumbwa et al., 2008). Understanding of transformational leadership was improved through the study of so many moderating variables (Hetland & Sandal, 2003). Although the main moderating variables have been mentioned, significant interactions effects have been found with many of these variables thus complicating the covariance matrix (Tabachnick & Fidell, 2007). With so many possible moderating variables, it was not surprising that research findings varied (Heinitz et al., 2005). This variation in results meant that no literature consensus was developed on a minimal design recommendation necessary to incorporate moderating variables. Each study continued to choose various moderators to include without any apparent consistency (Judge & Piccolo, 2004). However, Judge and Piccolo (2004) and Wylie and Gallagher (2009) noted that the design rigor of the MLQ based research had slowly improved.

Gender as a moderating variable, was an example of the problems of reaching consensus on recommendations for the research design. Gender seemed to be studied in some depth (Eagly et al., 2003; Wylie & Gallagher, 2009). However, even with gender, the incorporation in research studies varied widely. From the beginning of Bass's (1985) publication of the MLQ, gender was thought to be a significant moderator. Eagly et al.

(2003) found small but significant effects ranging from 0.02 to 0.12 with homogeneous samples for gender as a moderating variable. In general, women showed greater transformational behaviors than men. Also, men were more transactional and exhibited more laissez-faire leadership behaviors. Results using large number of samples ($N = 6,525$) by Antonakis et al. (2003) found that when the gender of the leader was the same as the rater, the model fit was better for the nine factor solution than other models. Further meta-analyses of 18 additional studies by Antonakis et al. indicated gender was a significant moderator in high risk or stable organizations such as in military combat units or public educational organizations and for low level leaders.

These homogeneous conditions in research by Antonakis et al. (2003) using large sample sizes may be difficult to replicate and therefore results may not be readily validated. Unlike Eagly et al. and Antonakis et al., Wylie and Gallagher (2009) did not find gender was a significant moderator. Hetland and Sandal (2003) found varied influence of gender on outcome measures. It may be that interaction with other moderating variables that were excluded from research designs may have influenced results.

Finally, lack of discriminate validity evidence for nine lower order constructs may also have been exacerbated by high correlations between contingent reward, a transactional construct, and many of the transformational constructs (Tejeda et al., 2001). However, Bass (1985) predicted this relationship. Transformational leaders use a combination transactional and transformational means to motivate followers, which then are reflected in the MLQ's results (Avolio & Bass, 2004a). As noted, the full range

leadership model terms this combination of behaviors, *augmentation*, with transformational behaviors adding to and building upon the contingent reward facet of transactional behaviors (Avolio, Walumbwa et al., 2009; Heinitz et al., 2005). This augmentation effect was supported in most studies (Heinitz et al., 2005; Judge & Piccolo, 2004). Wylie and Gallagher (2009) along with Judge and Piccolo (2004) noted that the number of moderators, interactions, and high shared variance through inter correlations contributed to the MLQ's reputation as a psychometrically difficult assessment with limited external validity and strong predictive validity.

External validity. A number of external construct validity studies comparing the MLQ to other nonleadership instruments were generally supportive of transformational leadership and in the anticipated direction (Avolio & Bass, 2004a; Bono & Judge, 2004; Lim & Ployhart, 2004; Rowold & Heinitz, 2007). Although no direct comparisons to other transformational questionnaires or leadership tests were described by Avolio and Bass (2004a), correlations with personality and cognitive tests were performed. These personality and cognitive test comparisons included Gordon personal profile, Myers-Briggs type indicator, Gough and Heilbrun adjective check list, 16PF intelligence scales, Constructive Thinking Inventory, Defining Issues Test, and Personality Orientation Inventory (Avolio & Bass, 2004a). Summary results were reported of correlations with self-confidence, self-efficacy, internal locus of control, and dominance as some of the moderators to inspirational motivation and idealized influence (Avolio & Bass, 2004a). Further, individual consideration correlations were noted with attributes such as tenacity, honesty, and persistence. However, no comparisons were reported for transactional or

laissez-faire leadership constructs (Avolio & Bass, 2004a). Although certainly not definitive in construct validity comparisons, these personality and cognitive correlations were theoretically explainable and seemed generally supportive of full range leadership model (Bass, Jung, Avolio, & Berson, 2003). However, in examples like Heinitz et al. (2005), in which critical correlations were not reporting, external validity conclusions were not supported.

Predictive validity. In terms of outcomes and therefore predictive validity, there were nine items not shown in Figure 1 that dealt with subjective outcomes. Three subjective outcomes of subordinates extra effort with three items, effectiveness of the leader with four items, and satisfaction with the leader with two items were explicitly measured in the MLQ (Avolio & Bass, 2004a). Outcomes of transformational leadership have generally been what practitioners were interested in, when applying new theories and strategies (Kaiser et al., 2008). The first subjective outcome of extra effort would be evident when the transformational leader succeeded in enlisting the entire person of the follower in the vision and goal achievement. Specifically, Hetland et al. (2007) reported a sustained subordinate effort level above that asked for by the leader; received due to contingent reward.

The second subjective outcome, leader effectiveness, was conceived as the performance output of the leader as perceived by the followers (Rowold & Heinitz, 2007). As such it was a subjective measure, as opposed to an objective financial or sales indicator, which was more easily captured in self-report surveys such as the MLQ (Snodgrass et al., 2008). As expected, there was some initial links between individual

consideration and perceptions of leadership effectiveness (Avolio & Bass, 1995). Avolio and Bass (1995) found that effective leaders seemed to use individual consideration to cascade organizational values and mission.

The final subjective outcome, satisfaction with the leader, incorporated the environment created by the leader at work (Hinkin & Schriesheim, 2008b). Satisfaction with the leader also included how well the leader encouraged colleagues to integrate their efforts for a common purpose (Avolio & Bass, 2004a). The sum total of over 25 years of studies using the MLQ showed that transformational leadership behaviors are positively associated with desirable additional subjective outcomes such as job satisfaction (Berson & Linton, 2005), organizational commitment (Barling et al., 1996), goal commitment (Cole et al., 2006), and innovation (Osborn & Marion, 2009). Bass, et al. (2003), and Heland, Sandal, and Johnsen (2007) found these subjective outcomes, along with additional objective sales and financial performance measures, formed the basis for the MLQ's predictive validity.

From a performance standpoint, the MLQ's predictive validity comes from results of strong relationships between transformational leadership behaviors on the MLQ and higher performance outcomes measured from sources external to the MLQ's nine outcome items (Avolio, Walumbwa et al., 2009). The effect size for transformational leadership was large from .44 to .73 (Judge & Piccolo, 2004; Lowe et al., 1996) as shown by many meta-analyses (Antonakis et al., 2003; Bono & Judge, 2004; Eagly et al., 2003; Judge & Piccolo, 2004; Lowe et al., 1996; Schyns et al., 2007) that were performed on the MLQ data. Objective measures were also strongly linked to transformational

leadership behaviors including productivity (Bass et al., 2003), financial performance (Howell et al., 2005; Peterson et al., 2009), team performance (Howell et al., 2005; Purvanova & Bono, 2009), and sales growth (Ling et al., 2008). Howell, Neufeld, and Avolio (2005) and Walumbwa et al., (2008) found that these subjective and objective results have encouraged researchers to increasingly use the MLQ for leadership assessments and to examine of mechanisms for this positive performance association.

The MLQ Summary

The MLQ's validity and reliability measures have received much attention and are being slowly refined (Antonakis et al., 2003). There seems to be a concerted effort to minimize main effects and interactions of so many moderating variables through increasingly rigorous research designs using homogeneous environments (Antonakis et al., 2003; Judge & Piccolo, 2004; Kanste et al., 2007). Regardless of the varying construct validity characteristics, the fundamental conceptual basis for transformational leadership, one of the three higher order leadership constructs, has shown sufficient predictive validity, with a range of performance outcomes, to warrant increasing use of the MLQ (Walumbwa et al., 2008). The bounding of this study on the transformational leadership items is a reflection of the psychometric differences with transactional and laissez-faire leadership items (Eagly et al., 2003; Hinkin & Schriesheim, 2008a, 2008b; Rowold & Heinitz, 2007). Although transformational leadership items also differ in construct validity results, they appear to be a function of the assessment version used in analysis (Avolio & Bass, 2004a; Heinitz et al., 2005; Judge & Piccolo, 2004). Further, many researchers (Barling et al., 1996; Carless, 1998; Hunt, 1999, Lim & Ployhart, 2004;

Purvanova & Bono, 2009) found transformational leadership, upon which the theory was named, to have greater utility in predicting individual and group behaviors.

Ideally, the MLQ would not have been designed with such a complex set of bi-level factor structures (Edelen & Reeve, 2007; Heinitz et al., 2005; Kanste et al., 2007; Tejada et al., 2001). Classical test theory and IRT procedures could have been combined to engineer an updated transformational instrument that had a simplified factor structure with high model fit items designed to test the latent ability evenly along the trait continuum with high discrimination supporting possible leadership selection criterion (Hu & Bentler, 1999; Reckase, 1979; Reeve et al., 2007; Russell, 2002; Samejima, 1977a; Zagorsek et al., 2006). With greater discrimination precision, detecting potentially harmful transformational leaders and adopting intervention strategies may be possible.

Methodology Considerations

Using the proposed archival samples, there were a number of design and analysis considerations that could have impacted the results (Kirisci et al., 2001; Reckase, 1979; Russell, 2002). Literature recommendations include a consideration for the assessment format and method of data collection, plus sample size and participant characteristics. Psychometrically, an inconsistent factor structure, IRT dimensional issues, and the use of appropriate analytical software were reviewed (Drasgow et al., 1995; Kirisci et al., 2001; Wilkinson, 1999). These areas are described in terms of literature suggestions. Further details on how these recommendations were employed in this study can be found in the Research Method chapter.

Instrument Format and Contact Mode

Instrument format. Wright (2005) expressed concern over potential differences in response between electronic-based versions versus paper-based versions of questionnaires. Mind Garden, holder of the copyright on the MLQ, offers both and does not distinguish between electronic or paper versions in reporting research results (Avolio & Bass, 2004a). Researchers have conducted equivalency experiments to determine if these concerns over instrument format affect responses (Donovan, Drasgow, & Probst, 2000). Davies and Wadlington (2007) found interactions between personality scores and administration type. Cole et al. (2006) conducted an extensive investigation to determine measurement equivalence of Internet and paper versions of the MLQ. The context of Cole et al. study was a single multinational power generation equipment manufacturer with employees in 50 countries, using 16 language translations of the MLQ's 20 transformational items. Participants included over 4,900 employees. Cole et al. found a similar factor structure and scalar invariance of the electronic version of the MLQ, using the same 20 transformational items as in this study, versus the paper version of the MLQ. Cole et al. further reported that the coefficient alpha for the two formats of the MLQ were identical at .96. The MLQ has been translated into multiple languages and evidence suggests that language does not, by itself, influence the mean ratings (Avolio & Bass, 2004a, 2004b; Cole et al., 2006). For this study all samples used paper-based versions of the same Hebrew or Russian translation of the MLQ, Form 5X short. The efficacy of the translation process was validated by Avolio et al. (1999) and was estimated by comparing translated scores with untranslated scores which is reported in Chapter 4.

Contact mode. Cole et al. (2006) used electronic mail and paper mail contacts for invitations and discovered no significant differences in the response data for the MLQ's transformational items. Porter and Whitcomb (2007) further investigated whether the contact type and relationship to the requestor had an influence on response rates. Neither the strength of the relationship to the requestor nor the invitation type had a significant impact on response rates (Porter & Whitcomb, 2007). The contact mode of the three proposed samples is not known, however, Porter and Whitcomb (2007) found that contact mode differences were not significant as outcome moderators.

Participant Characteristics

Sample size. Unlike classical test theory, there are no agreed guidelines for sample size in IRT analysis (De Ayal, 2009; Embretson & Reise, 2000; Kirisci et al., 2001; Reise & Yu, 1990). For tightly constrained modeling techniques such as confirmatory factor analysis, structural equation modeling, and IRT analysis sample size has a direct bearing on results (Russell, 2002). Many estimation techniques and fit indexes are sensitive to sample size variations (Hu & Bentler, 1998, 1999; Wilkinson, 1999). Further, differential item analyses may reduce samples sizes of some subgroups. Sample sizes above 3,000 are preferable when estimating a guessing parameter (Dragow et al., 1995). This study used models without guessing parameters; therefore, smaller samples sizes were used. Missing data, uneven distribution of data, large number of parameters, degrees of freedom, and factor loadings were considered in matching model estimation and fit techniques to available data characteristics (Russell, 2002). If partitioning data into a few items per dimension was required, De Ayala (2009) suggested

that the sample size be increased to accommodate shorter response vectors for IRT analysis.

As previously noted, smaller sample sizes together with fewer items are used in IRT analysis on an exploratory basis. Measurement errors would consequently increase, providing less stable parameter estimates. However, the exact size of the recommended sample for different IRT analyses has not been established in literature (De Ayala, 2009; Edelen & Reeve, 2007; Orlando & Marshall, 2002; Thissen et al., 1988; Wright, 1977). Minimum sample size recommendations range from 100 participants for dichotomous exploratory purposes with a one parameter model (Wright, 1977) to over 3,000 participants for item pool construction with a three parameter model (Drasgow et al., 1995) and is also driven by the number of items in the assessment. However, there are a number of variables that enter into calibration sample size considerations. For instance, De Ayala (2009) suggests such issues such as generalizability, amount of missing data, number of items, intersection between items and people locations on the latent trait scale, and data to model fit all influence the decision of appropriate sample size. Also, the distribution of the latent trait in the population is of concern when determining adequate sample size according to Birnbaum, in Lord and Novich (1968). Adding to these variables is the ability of estimating equations in the IRT software, which Kirisci, Hsu, and Yu (2001) found to deal robustly to some violations of unidimensionality.

Although there is no consensus on minimum sample size, there are sample size investigations that pertain to GRM using MULTILOG relevant to this study. Kirisci et al. (2001) concluded that using 40 items and 1000 sample size, was excessive. Instead,

Kirisci et al. recommended 20 or more items should be used with at least 250 cases to minimize the effects of most violations to IRT unidimensionality and normality assumptions. Reise and Yu (1990) found marginal maximum likelihood estimates correlated to true estimates ($r = .85$) when sample sizes were at least 500. Given the findings from Kirisci et al. (2001) and Reise and Yu (1990) using GRM with MULTILOG, even with 20 items, as in this study, a calibration sample size above 500 should have lead to relatively stable parameter estimates. An additional source of guidance on sample size comes from IRT research on leadership assessments using the GRM or the GGUM models. Scherbaum et al. (2006) used a sample size of 445 and Zagorsek et al. (2006) used a combined 801 sample size. Combining the three samples used for calibration in this study is expected to be about 2,200 cases. The proposed sample size for this study should have resulted in relatively stable parameter estimates. Limitations of using a small sample size were noted in Chapter 1. Samples and procedures is discussed further in the Research Method chapter.

Participant homogeneity. Including participant leaders from diverse settlings in this study could reduce the homogeneity of the sample and may lead to finding fewer factors (Antonakis et al., 2003). However, using a singular setting may influence results by restricting generalization (Peterson, 2001). The telecommunications company sample and the professional basketball sample were thought to be fairly homogeneous and therefore produce a larger number of significant factors than the sample of 26 companies (Antonakis et al., 2003).

The three combined samples used for calibration were Israeli companies and professional sports teams rating direct supervisors. IRT software used in this study defined the midpoint of the trait axis by the mean of the person abilities (Embretson & Reise, 2000). The comparison of sample means was used to equate the relative scales for comparison purposes.

Rater type. The final participant variable to consider was self-rating versus rating someone else. The authors of the MLQ discouraged the use of self-ratings as being too subjective and inflated by a full scale point above ratings from subordinates (Avolio & Bass, 2004a). There were also psychometric problems with self-rating factor structures. Avolio and Bass (2004a) found that, for a nine factor structure, only self-ratings had six items with factor loadings below .40. These six items were from six different lower order facets: both types of idealized influence, individual consideration, contingent reward, passive management by exception, and laissez-faire. Whereas, for all other rater types, only item 17 had a factor loading below .40 from passive management by exception (Avolio Bass, 2004a), which is not included in this study. Therefore, self-rating data should be treated with caution and combining self-rating with other types of rating data may result in less precision in predicting item location parameters. Fortunately, all three archival samples proposed were screened to include only employees rating their direct supervisors and therefore contained no self-rating responses. However, this introduced the issue of correlated observations within the leader's group of subordinates.

Therefore, one of the issues examined was the use of the raters' version of the MLQ. A group of subordinates rating the same leader on observed behaviors does not

constitute independent observations. However, literature on multisource feedback (Allen, Barnard, Rush, & Russell, 2000; Atkins & Wood, 2002; Atwater, Roush, & Fischthal, 1995; Atwater, Waldman, Ostroff, Robie, & Johnson, 2005; Berson & Sosik, 2007; Gentry, Hannum, Ekelund, & de Jong, 2007; Sala 2003; Sala, & Dwight, 2002; Schaefer 2008) suggests that behavioral leadership assessments seemed to record perceptual rather than actual responses of the selected behaviors. For instance, self-ratings were known to be inflated and this inflation increased with managerial level (Gentry et al. 2007; Sala, 2003). This rater perception was thought to introduce significant individual response variations within a leader's group.

Two IRT articles concluded that the trait being measured by subordinates of the same leader seemed to be individual perceptions of the leadership behaviors rather than measuring the leadership behaviors directly (Barr & Raju, 2003; Craig & Kaiser, 2003). As such, the high amount of variance within each subordinate group, compared to between leader's variations, would have justified retention of individual rater's responses. The trait reported in the IRT analysis would be the rater's perception of leaders' transformational traits rather than the trait itself.

Thus, the ICCC, using a one way random effects model, was used to determine if there was sufficient within group variation to justify retaining each rater's individual response (McGraw & Wong, 1996a, 1996b). An ICCC value at or below .20 would have indicated that subordinate responses had sufficient within-group variation to use all subordinate ratings. Walumbwa et al., (2008) reported an ICCC of .10 on the MLQ rater version. Therefore, there was reason to believe this study would find similar results.

IRT Model Optimization Steps

The MLQ's factor structure has been found to vary by sample source precisely because of the large number of moderating variables interacting to form even more complex relationships (Antonakis et al., 2003; Teresi & Fleishman, 2007). The factor structure for any given analysis was stable (Antonakis et al., 2003). Therefore, the IRT analysis for this study was based on the dominant dimensions of the reported maximum likelihood factor analysis (Drasgow et al., 1995). Drasgow et al. (1995) found that all IRT models were misspecified to some degree and that data to model fit analysis would indicate the extent and impact of the misspecification.

Data screening. Screening involves reviewing raw data to determine obvious issues before more detailed analysis is performed. Data for this study needed to be examined for typographical errors, indiscriminant responses, and adequate category responses. Integer responses other than one to four would have constituted typographical errors and be treated as missing values. Selecting the same category for all MLQ's 36 leadership trait items would have indicated possible indiscriminant responses. Five responses per category for all 20 items were the minimal requirement per analysis (Tabachnick & Fidell, 2007). Any data that needed to be removed was treated as missing data. MULTILOG treats missing data as mean theta values (Thissen et al., 2003). GGUM2004 treats missing data as random theta values (Roberts & Shim, 2008). The Research Method chapter will further discuss data screening provisions for this study.

IRT model and dimensionality tests. Chi-squared over degrees of freedom for single item analysis, doublet item analysis, and triplet item analysis was the primary

criteria recommended to validate appropriate IRT models and also to detect dimensionality violations (Drasgow et al., 1995; Kirisci et al., 2001). The models chosen are Samejima's (1969) GRM and Robert's (2008) GGUM. The probability of observed responses was compared to probability of expected responses along with marginal reliability. Category responses, item parameters, and person abilities were also calculated using the two models. MODFIT is the software program that was used to analyze the degree of data to model misspecification, which was used as an indication of unidimensional violations (Stark et al., 2001). The MODFIT software is discussed further in this chapter.

Software

Classical test theory software. PASW and AMOS software version 18 was used to examine the data, perform traditional item analysis, and utilized in maximum likelihood estimation for factor analysis (SPSS, 2009). Software programs are continually improving their offerings (Russell, 2002). DiStefano and Hess (2005) found 16% of peer-reviewed journal psychological assessment articles used the SPSS software suites for factor analysis. Reckase (1979) suggested that for factor analysis, the first factor should be above 10 or the total variance above 20%.

IRT analysis software. Because of the difficulties of IRT analysis, solving for person and item difficulty estimates involving numerous simultaneous and iterative computations, it is perhaps not surprising that early IRT adopters were somewhat restricted until the 1980s when more powerful software programs and computer capacities were available (Kirisci et al., 2001; Lissak & Wytmar, 1981). IRT software,

while providing numerical output, still relies on graphical representations of these estimates to ease the complexity of interpreting the enormous volumes of data (Thissen et al., 2003). Typically, IRT analysis uses model fit to confirm item and person parameter estimation.

In terms of software package selection for this study, University of Illinois at Urbana-Champaign had an extensive IRT modeling laboratory, which provided free software, called MODFIT, for addressing how well the observed responses fit the selected models (Stark et al., 2001; Zagorsek et al., 2006). MULTILOG software was used to calculate IRT parameters for the GRM specifications (Kirisci, et al., 2001; Scherbaum et al., 2006; Thissen et al., 2003; Zagorsek et al., 2006). GGUM2004 is free software that was used for the GGUM analysis (Roberts et al., 2006). Although other IRT software programs are available, these packages were used in studies involving Likert type scales such as in the MLQ (Chernyshenko et al., 2001; De Ayala, 2009) and in leadership studies (Craig & Gustafson, 1998; Scherbaum et al., 2006; Zagorsek et al., 2006). Additionally, the selection of MULTILOG for IRT analyses in this study was supported in part by Kirisci et al. (2001), who found lower variances in parameter estimates under conditions in which violations of unidimensionality occurred.

Literature Review Summary

The MLQ is the standard for research on transformational leadership but is not free from psychometric inconsistencies (Antonakis et al., 2003). No known replication of Antonakis et al. (2003) nine factor result was published in peer-reviewed journals. In the 25-year history of the MLQ, improvements in the assessment have not resolved the

convergent and discriminate validity issues (Avolio & Bass, 2004a). Among the possible causes for instability are the numerous moderating variables impacting the MLQ responses (Judge & Piccolo, 2004). With the lack of agreement about psychometric properties for heterogeneous samples and limitations in the precision of reliability metrics at the item level (Antonakis et al., 2003); it was time to investigate an augmentation to previous classical test theory analysis.

This study sought, for the first time, to investigate the MLQ's 20 transformational leadership items, using IRT. Investigation into item response functioning could provide additional information at the person, item, and subscale level over classical test theory alone. The IRT gap in literature may have persisted due to the difficulties of applying unidimensionality to an instrument that is designed to be multifactorial (Reckase, 1979).

To minimize violating IRT dimensionality assumptions, factor analysis was employed to discover the optimum association of items to dimensions with a combined calibration sample. The calibration sample proposed was the combinations of three Israeli samples from business and sports subordinates. Once the items were grouped by the appropriate dimension, literature suggested applying Samejima's (1969) GRM and Robert's (2008) GGUM for ranked homogeneous polytomous instruments (Ostini & Nering, 2006; Scherbaum et al., 2006).

This IRT analysis of the MLQ's transformational leadership subscale can provide insights into psychometric properties that had not yet been explored, such as item discrimination and difficulty (Scherbaum et al., 2006; Zagorsek et al., 2006). In summary, this study can contribute to unidimensional IRT analysis, fuller understanding

of the transformational leadership subscale's psychometric properties, and IRT person abilities of business leaders, as well as lead to improved detection of potentially harmful transformational leaders. A discussion about how these Methodological Considerations were operationalized appears in chapter 3.

Chapter 3: Research Method

In this chapter, I incorporate the results and recommendations of statistical and leadership researchers discussed in chapter 2. The MLQ research findings highlighted structural and correlational variations with various sample sources (Antonakis et al., 2003). The transformational leadership theory and the MLQ findings were used to prepare for possible sample analysis issues and determine appropriate decision criteria. Understanding the theory and assumptions of IRT along with the MLQ's structural issues led to an understanding of possible limitations of IRT results. More importantly, these pieces of accumulated knowledge illustrated that the results of this study were dependent on the proper integration of conceptual theory with research findings.

In the remaining chapter, the Research Design and Approach section presents the analysis of the MLQ's 20 transformational leadership items that constituted the boundaries of this study. The Samples and Settings section describes the archival samples that were proposed for use in the IRT analysis. The Instrument and Analytical Software section summarizes the MLQ assessment and the four software programs that were proposed for use in this study. The Data Preparation and Analysis section details the data examination processes and sequential analyses required to meet the objectives of this study. The Ethical Protections section discusses the care of participants in the three samples of archival data. Finally, chapter 3 is summarized to show how the proposed methods, supported the research questions and objectives of performing the IRT analysis for the MLQ's 20 transformational leadership items.

Research Design and Approach

This study is an IRT methodology study of the MLQ's transformational leadership subscale using archival data. The objectives of this study were to: (a) test the fit of IRT models to the 20 item MLQ transformational leadership subscale, (b) estimate the IRT parameters for each of the 20 items, and (c) evaluate changes in the reliability estimation of scores from the subscale when using IRT versus classical test theory analysis. These objectives were met using various traditional and IRT analyses.

The initial determination of how well the data met certain assumptions for traditional and IRT analyses were examined. The three samples of archival data were evaluated for typographical errors, complete lack of responses, adequate category responses, and rater response variance. Typographical errors were replaced with a system missing value. Cases were removed if there was a complete lack of responses to all 20 items. Adequate cell frequency of responses for any category of any item was five or higher (Tabachnick & Fidell, 2007). If fewer than five responses per category to any item exist, the categories would have needed to be collapsed.

For the calibration sample, ICCC analysis determined if rater responses within a leader's group had sufficient variation to retain all rater responses (Walumbwa et al., 2008). If ICCC was .20 or below, all raters were retained ($n = 2,222$). If ICCC was above .20, a random rater would have been selected from each group to ensure greater independence of observations ($n = 357$). There was reason to believe the ICCC would have been .20 or below given the MLQ research of Walumbwa et al. (2008).

Classical test theory analysis was performed for internal consistency and item analysis for the combined sample. The Cronbach's alpha coefficient for corrected item-total correlation was used for the discrimination parameter and the mean item score and standard deviation was used for the difficulty parameter (Scherbaum et al., 2006). These item parameter estimates were used for comparison purposes with IRT parameter estimates to examine differences in reliability.

Maximum likelihood factor analysis was performed to determine the extent of unidimensionality violations related to IRT assumptions. Factor loadings at or above 0.40 and first eigenvalue explaining 20% or more total variation indicated a single dominant dimension (Reckase, 1979). If multiple dimensions are discovered that could not have been resolved by the IRT software, either separate IRT analyses would have been conducted for each dimension using two IRT models, or the item(s) with low factor loadings could have been removed from the analysis based on examination of loadings and IRT discrimination parameters.

Robert's (2008) GGUM and Samejima's (1969) GRM models were used for the three combined calibration samples. IRT item and subscale parameters were produced. Using the item parameters, data to model fit was determined, as were person ability estimates for the calibration sample. The IRT model that best fit the combined sample had the lowest chi-squared over degrees of freedom values for singlet, doublet, and triplet based on MODFIT output. Mean values of three and below indicate excellent data to model fit. Finally, the mean theta of the three individual archival samples was compared using the large telecommunications sample as the anchor (Embretson & Reise, 2000).

The Results and Discussion sections describe the output of analysis and interpretation of results, respectively.

Justification of Design and Approach

This study's design and approach followed Scherbaum et al. (2006), evaluating the MLQ rather than the Multidimensional Leader-Member Exchange assessment. The main difference in methodology was using a combination of three samples from Israel business and sport subordinates instead of U.S. university employees for the calibration sample. Although Scherbaum et al. (2006) used university union workers, the IRT analysis by Zagorsek et al. (2006) drew from a sample of business graduate students. According to Peterson (2001), these university settings may introduce common moderating factors.

From a leadership perspective, this Israeli corporate sample may further leadership IRT. Judge and Piccolo (2004) found differences in the MLQ's estimated true score correlations amongst college, business, public sector, and military settings. Peterson (2001), using over 650,000 participants, found that college student populations were more homogeneous than nonstudent adult populations. Further, effect sizes varied significantly with no discernable pattern. Therefore, Peterson argued for caution when generalizing from university only samples.

In contrast to these results, Schyns et al. (2007) performed a meta-analysis and found no significant differences using university over employee samples once outliers had been removed. However, with outliers, university samples were significantly more homogeneous than nonuniversity samples. In this study, the calibration sample was

composed of three Israeli samples. Raters from a single corporation sample, raters from a sample of leaders from 26 businesses, and raters from professional basketball teams were combined for calibration. Therefore, in contrast to Scherbaum et al. (2006), this study used only business and professional team samples to extend IRT literature of leadership instruments avoiding the various reported effects of using university based samples.

The use of Scherbaum et al.'s (2006) study as a template for this study may also facilitate IRT comparisons with other leadership instruments and thereby extend leadership assessment literature. IRT leadership literature is sparse (Scherbaum et al., 2006). This study may encourage other researchers to compare additional leadership instruments using IRT analysis. As the body of IRT leadership literature develops, comparisons become more meaningful on item, subscale, and person parameters, while increasing knowledge of response behaviors on leadership instruments.

Samples and Procedures

Three archival samples containing responses from each of the MLQ's 36 leadership trait items were analyzed. Only the 20 items of the transformational leadership subscale out of the total of 36 items were reviewed in this study. The samples' owner provided the three samples proposed for this study's IRT analysis (Y. Berson, personal communication, October 14, 2009). The three samples provided the responses to raters' observations of their direct supervisor. As such, the latent trait being measured was the raters' perceptions of their leaders' transformational traits rather than the traits themselves (Barr & Raju, 2003; Craig & Kaiser, 2003). Two of the samples were from Israeli businesses. One business sample was from a large telecommunications design and

manufacturing company ($n = 1,600$). The second business sample was from 26 Israeli companies of various industries. In this second business sample, the chief executive officer was rated by senior vice presidents, who were rated by subordinate executives ($n = 282$). The third sample was from professional Israeli basketball players ($n = 357$) rating their coaches. All three samples were combined for calibration purposes ($n = 2,222$). Each of these samples and procedures are described in more detail.

Description of the Israeli telecommunication sample and procedures comes from several published articles (Berson, 1999; Berson & Avolio, 2004; Berson & Linton, 2005; Berson & Sosik, 2007). Berson, together with the human resource department, administered the Hebrew language paper version of the MLQ Form 5x to 30-60 employees per session over a 2-month period at the company (Berson & Linton, 2005). The efficacy of the translation process was validated by Avolio et al. (1999) and can also be estimated by comparing translated scores with untranslated scores, which is reported in chapter 4. All participants rated their direct supervisor on leadership (Berson, 1999). Employees only provided their unit number to retain anonymity (Berson & Linton, 2005).

Data collection for this telecommunication company occurred around 1998 by Berson (1999). This large Israeli company employed about 2800 employees, of which 2025 completed the MLQ assessment rating their direct ($n = 1,600$) and indirect leaders ($n = 425$). Only direct ratings were used in this study. The leaders being rated were 205 department managers, 33 division and area managers, 10 vice presidents, and the chief executive officer. Males represented 69.5% of the respondents. Education varied from 29.1% with high school degrees or less, 49.2% had a college degree, and 21.7% were

technicians. Berson (1999) also reported tenure information with the company ($n = 1,913$), 23.6% had worked less than two years, 38.6% between 2 to 10 years, and 37.8% over 10 years. This first sample represented about 71% of the combined calibration sample.

The second business sample, containing ratings from a number of Israeli companies, was described in a published article (Berson et al., 2008). In 2001, 139 publically traded Israeli companies were contacted for the survey of which 26 companies responded with the minimal required information. Each company administered the MLQ to their employees independently of other companies. No individually identifying data was collected to preserve anonymity. The Hebrew language paper version of the MLQ Form 5x was used. The efficacy of the translation process was validated by Avolio et al. (1999) and was also estimated by comparing translated scores with untranslated scores, reported in chapter 4. All participants rated their direct supervisor on leadership. Of the participants ($n = 282$), 26 were male chief executive officers, 71 were senior vice presidents, of which 82% were male, and 185 were direct reports of the senior vice presidents, of which 69% were male. Mean age of chief executive officers were 52 ($SD = 7.08$) while mean age of the remaining participants was 44 ($SD = 9.5$). Mean job tenure for the chief executive officers was 7.9 years ($SD = 7.5$) and 12.5 ($SD = 11.24$) for the remaining participants. Tenure in the organization for the entire sample was 8.4 years ($SD = 7.9$). Average number of employee per company was 390 ($SD = 450$). Of the 26 companies spending on research and development, 16 were below 3%. Berson, Oreg, and

Dvir (2008) also reported that the MLQ ratings were for direct leaders only. This second sample represented about 13% of the combined calibration sample.

The third sample was of professional basketball team players rating their coaches. Description of this sample comes from personal communications (Y. Berson, personal communication, November 10, 2009). The Hebrew language paper version of the MLQ Form 5x was used. The efficacy of the translation process was validated by Avolio et al. (1999) and was also estimated by comparing translated scores with untranslated scores, reported in chapter 4. Data were collected around 2000. There were 45 basketball teams represented. All participants and their coaches were male ($n = 357$). Average age of players was 22.0 ($SD = 4.9$). Average tenure on the team was 1.8 years ($SD = 1.8$). Average tenure in basketball was 11.9 years ($SD = 4.5$). Average tenure with the coach was 1.6 years ($SD = 1.4$). There are no published reports for this sample and Berson did not provide a description of the procedures used to collect the respondents' scores. This third sample represented about 16% of the combined calibration sample.

Instrument and Analytical Software

A paper-based Hebrew translation of the MLQ Form 5X short was used in all samples (Y. Berson, personal communication, November 10, 2009). A back translated Russian paper-based version of the MLQ Form 5X short was also used for a few participants in the telecommunications company sample (Berson, 1999). The efficacy of the translation process was validated by Avolio et al. (1999) and was also estimated by comparing translated scores with untranslated scores, reported in chapter 4. Of the 45 questions from the MLQ, all 36 leadership trait questions were presented to all

participants across the three samples. Participants responded to statements describing behaviors and attributes of their direct supervisor. These items are scored in Likert fashion from zero to four, with zero representing *not at all*, one representing *once in a while*, two denoted *sometimes*, three was *fairly often*, and a score of four was *frequently, if not always* (Avolio & Bass, 2003a). Only the 20 transformational leadership items were used in this study. Details of the MLQ instrument, theoretical basis for each factor, and research findings, were discussed in detail in the Literature Review chapter.

Data Preparation and Analysis

The content of this section is organized by data preparation, assumption testing, and five research questions. For each, the purpose of the operation, the procedure(s) followed, and any guiding criteria are presented. In chapter 4, the results are described in the same order.

Data Preparation

Three archival samples were examined before being combined. The overall purpose was to analyze and adjust for inconsistencies in the samples before combining. Three operations were conducted. The first had the purpose of screening for rater only data. The procedure used was to filter out nonsubordinate responses. The guideline used was that only direct subordinate responses should remain. The screening was based on data field values. SPSS was the analytic software used as part of PASW version 18 (SPSS, 2009).

The second operation had the purpose of detecting outliers or typographical errors. The MLQ should only have integer values of zero through four (Avolio & Bass,

2004a). The procedure used was to examine histogram information of each item. Values that were not integers from zero to four were easily detected. The guideline was that all unexpected values were to be replaced by a missing data designation.

The third operation had the purpose of evaluating missing data. The procedure used was calculating the sum of scores for each respondent. The guideline was that respondents with zero sum scores had no usable information and were to be removed from the sample. The three operations were performed on individual samples before combining all three samples for assumption testing.

Assumption Testing

A number of limitations and assumptions are presented in Chapter 1. Six operations to analyze potential impact of testable limitations and assumptions were conducted. SPSS and AMOS software were used for the classical theory analyses; part of PASW version 18 (SPSS, 2009). MULTILOG version 7 (Thissen et al., 2003) was used for equating corporate and athletic samples.

Assumption 1: Translation accuracy. The first operation had the purpose of indicating the degree of mistranslation in the Russian and Hebrew versions of the MLQ used in collecting the archival samples. The procedure used was a mean difference test for percentile comparisons between published normative data for the United States (Avolio & Bass, 2004b) with archival responses. The guideline used for adequate translation was a mean difference p value less than .05.

Assumption 2: Independent observations. The second operation had the purpose of evaluating the degree of violation of independent observations. The procedure

used was to examine the effect of within group correlated observations from subordinates using a one way random effects ICCC model. The guideline was for an ICCC at or below .20 the rater's individual responses were retained. If ICCC was above .20, then a random rater would be selected from each leader ($n = 357$) supporting independence of observations.

Assumption 3: Sufficient category responses. The third operation had the purpose of determining the sufficiency of categorical responses. The procedure used was examining the histogram of each item to determine the number of responses per category. The guideline was that items having four or fewer responses per categories were collapsed with other categories of the same item. Final histograms were to show all categories of each item had five or more responses.

Assumption 4: Normal distribution. The fourth operation, in preparation for factor analysis, examined item values for normal distributions (Tabachnick & Fidell, 2007). Normal item value distributions can be difficult to achieve constrained by a small number of discrete category choices per item. Guidelines for normal distribution are means and standard deviations closest to the theoretical mean for the MLQ of two, with a standard deviation of less than one (Tabachnick & Fidell, 2007). Further, normal distributions should have a skewness and kurtosis of less than an absolute value of one (Tabachnick & Fidell, 2007).

Assumption 5: Unidimensionality. The fifth operation had the purpose of testing for unidimensionality. Before performing factor analysis, the procedures recommended were item-item correlations and item-total correlations (Tabachnick & Fidell, 2007).

Tabachnick and Fidell (2007) suggested guidelines for correlation values between .20 and .80 could lead to interpretable factor analysis results. Further, values below .20, suggest the influence of additional factors and values above .80, suggest redundant items.

Achieving stable parameter estimates using unidimensional IRT models is supported by an assumption of unidimensionality (De Ayala, 2009). Embretson and Reise (2000) emphasized that the primary approach for unidimensionality testing was factor analysis, using exploratory and confirmatory procedures. Exploratory factor analysis for the MLQ used maximum likelihood estimation (Tabachnick & Fidell, 2007). Guidelines for unidimensionality in exploratory factor analysis was factor loadings of .40 or higher, a first eigenvalue of 20% or more of total explained variance, and a second eigenvalue below one (Reckase, 1979). Guidelines for confirmatory factor analysis are model fit indices CFI, RFI, and NFI, at or above .95 for a good model fit, with RMSEA at or below .06 (Hu & Bentler, 1999). If the factor analyses guidelines were not met then separate IRT analysis was to be performed on each dominant factor grouping. De Ayala (2009) suggested that separating the items by unique factor would support the IRT assumption of unidimensionality.

Assumption 6: Sample homogeneity. The sixth and final operation had the purpose of examining differences in mean perceived transformational leadership abilities between the corporate and athletic archival samples. The procedure for equating samples, described by Thissen, Chen, and Bock (2003), uses MULTILOG to anchors one sample to a mean of zero and a standard deviation of one. All other samples' ability means are then computed in relation to the anchor sample mean ability using a single MULTILOG

analysis. The guideline is a classical theory mean difference test with a p value less than .05, indicating no significant mean difference between corporate and athletic samples.

With the completion of assumption testing, research questions were investigated.

Five Research Questions

Data preparation and assumption testing were prerequisites to analyzing each research question. Research questions used SPSS version 18 software for classical theory analysis (SPSS, 2009), MULTILOG version 7 for GRM IRT analysis (Thissen et al., 2003), GGUM2004 for GGUM IRT analysis (Roberts et al., 2006), and MODFIT version 1.1 for data to model fit statistics (Stark et al., 2001). For the GRM model, MULTILOG reported the discrimination parameter in the logistic metric contain the constant, D , equaling 1.702, which was divided from the discrimination parameters before conducting the MODFIT analysis (Embretson & Reise, 2000). MODFIT assumes inputs are in the normal metric (Stark et al., 2001). Roberts and Shim (2008) indicated that GGUM output uses the normal metric and therefore, discrimination parameters were used directly in MODFIT analysis. IRT procedures are straightforward even if the software is not (De Ayala, 2009). A single software operation can yield categorical, item, and subscale level metrics and graphs. To facilitate the report of results, findings, and recommendations in later chapters, the purpose, procedures, and guidelines were organized by research question.

Research question 1: Observed versus expected IRT model responses. The purpose of research question one was to investigate category level estimates of each item comparing the GRM and GGUM models. The procedures started with entering category

responses into MULTILOG and GGUM2004 software to calculate item level parameters. The item parameters were then applied along with categorical responses for two models, the GRM and the GGUM, to determine category separate model fit plots and model fit metrics. Guidelines were that expected category responses were within a 95% confidence interval of observed category responses (De Ayala, 2009). Drasgow et al. (1995) expected that visual inspections would provide additional descriptive information by category, across all items of a dimension.

Research question 2: Best IRT model. The purpose of research question two was to determine which model, the GRM or the GGUM, had the best match between the models' expected and the observed item parameters. The procedure was to examine the X^2/df metrics from the MODFIT procedure used in research question one. The guideline was X^2/df less than three for singlet, doublet, and triplet tests denotes good model fit (Drasgow et al., 1995). The model with lowest X^2/df in singlet, doublet, and triplet tests was the best model (Stark et al., 2001). Some values above three were expected, due to significant amount of variability introduced by moderator variables (Antonakis et al., 2003). Therefore, some violations of functional form were anticipated. However, Kirisci et al. (2001) found MULTILOG was robust to some violations of IRT assumptions. Stark et al. (2001) suggested beside fit metrics results, visual inspection of item option response functions provided another means of investigation data to model fit.

Research question 3: Discrimination and difficulty parameter estimates. The purpose of research question three was to examine the discrimination and difficulty parameters of the best fitting model from research question two. The procedure was to

examine the item parameters from research question one, of the GRM or the GGUM, that best answered research question two. Zagorsek et al. (2006) determined the evenness of item level coverage across the perceived leadership trait continuum from negative three to plus three as was completed for this study for comparison.

Research question 4: Highest trait range reliability estimation. The purpose of research question four was to examine the entire 20 item transformational leadership subscale for reliability and standard error of measure metrics. The procedure was to use the MODFIT's test information function values calculated as a byproduct for research question one for the best fitting model from research question two. Test information function standard error of measure values along the trait continuum were produced as part of the results. Reliability is one minus the square of standard error of measure (Tabachnick & Fidell, 2007). Guidelines were to determine the highest reliability metrics across the perceived transformational leadership range (Samejima, 1977b). Zagorsek et al. (2006) suggested that leadership assessments generally are not reliable, at .95 and above, in the upper trait range and the MLQ was anticipated to confirm this expectation.

Research question 5: IRT versus classical test theory reliability. The purpose of research question five was to compare item level and subscale level reliability using classical test theory and IRT. The procedures for classical test theory were item calculations of mean and standard deviation as an indication of difficulty and corrected item-total correlation for discrimination. Cronbach's alpha was also to be calculated for each item and subscale. The classical test theory item parameters were compared to IRT item parameters from research question three. Classical test theory subscale reliability

was compared to IRT subscale reliability from research question four. De Ayala (2009) suggested that the greater precision afforded by IRT was usable in professional applications.

A final operation, comparing classical test theory with IRT reliabilities, was the calculation of perceived transformational leadership abilities of the combined sample. Individual abilities are computed using, and therefore after, item parameters. The procedure used item parameters from research question one for the best model from research question two and calculating the appropriate person abilities of the combined sample by individual. The Nunnally and Bernstein (1994) guideline was to determine the trait range for reliabilities at .95 or above.

Ethical Protections

Although an institutional review board procedure was completed successfully for the telecommunication company sample (Berson, 1999), there was no mention of this procedure for the 26 company senior management sample (Berson, Oreg, & Dvir, 2008) and basketball player sample. It was not expected that senior managers of Israeli companies or professional basketball players were as vulnerable as other protected class populations. There was no individually identifiable personal information in the archival data analyzed. Two of the three samples had already been used in peer-reviewed publications with no known adverse participant impact (Berson et al., 2008; Berson & Sosik, 2007). With the MLQ leadership assessment completed voluntarily by professionals rating, not themselves but their supervisors, with no identifying information, any potential harm was believed to be negligible. In addition, all three

samples were combined so that item level analysis with no identifiable information could not be traced to any one sample.

The potential harmfulness of any individually identifiable leader was not possible or part of this study. Therefore, no specific protection was required. If a future study does use the MLQ for detection of potentially harmful leaders, new data will be required with additional protections for those leaders identified as potentially harmful as in Khoo and Burch (2007). Potentially harmful leaders identified in future studies may lead to appropriate feedback, development, or separation which might be administered by human resource departments with legal obligations for protection of those leaders. The subordinate employees who currently suffer under harmful leaders may desire additional protections and could receive significant benefits from the future development of this study (Pullen & Rhodes, 2008). The knowledge accumulation on the MLQ's psychometric properties of these samples has already been considerable (Berson & Linton, 2005) and can be extended with this study to benefit future researchers, practitioners, and subordinate employees. The institutional review board assigned identifier 06-17-10-0321129, for this study.

Summary

This study had the objective of discovering the discrimination and difficulty values of each of the MLQ's 20 transformational leadership items and subscale metrics. It was hoped that the results could add practical interpretability for practitioners administering the MLQ, including detection of potentially harmful leaders. The MLQ is known to be inconsistent psychometrically, yet practitioners are increasingly using the

instrument due to its predictive validity (Antonakis et al., 2003). The predictive validity has advanced the MLQ, as the most researched transformational leadership instruments globally (Heinritz et al., 2005). The sequence of data screening, factor analysis, IRT modeling, and data to model fit techniques discussed in this chapter were designed to provide the greatest possibility for interpretable and potential for useful results. These expected item, subscale, and people ability results, meet the bounded study objectives, thereby aiding detection of potentially harmful transformational leaders using the MLQ.

Chapter 4: Results

The five research questions posed in this study are addressed in this chapter. The questions, together with data preparation, an unanticipated limitation, and assumption testing, form the basis of chapter sections. The findings are summarized after presenting the relevant research results and explanations.

Data Preparation

The Sample and Procedures section in chapter 3 presented sample demographics and other descriptive statistics from published reports. The data analyzed for this study included a sample identifier, a group membership identifier, and the individual responses to 20 MLQ transformational leadership items. The group membership identifier was coded to assure anonymity. No other information was included in the analysis.

For this study, three samples were combined. The data preparation is described for each sample. The Israeli telecommunication sample included direct and indirect ratings ($n = 2,199$). Indirect ratings were removed ($n = 425$). Of the remaining direct subordinate ratings ($n = 1,774$), those missing all 20 item responses were removed ($n = 89$), leaving 1,685 subordinate ratings. The remaining data contained only integer category responses from 0 to 4, indicating no obvious typographical errors. Where no responses were recorded, system missing indicators were present for up to 19 items. There were 219 separate groups of subordinates.

The second sample of executives from 26 Israeli companies contained 282 respondents. Chief executive ($n = 26$) and senior vice president ($n = 71$) self-ratings were removed, leaving 185 subordinate ratings, representing 26 independent company groups.

There were four respondents removed due to missing values for all 20 transformational leadership items ($n = 181$). Three subordinates had responded to a single item with intermediate scores (1.5, 2.5, and 3.5), rather than the expected 0 to 4 integers. After consultation with the author of the archival data, these three values were rounded down to 1, 2, and 3, respectively (Y. Berson, personal communication, June 16, 2009).

The final sample was from professional basketball players. There were 357 direct coach ratings. Only one respondent left all 20 transformational leadership items blank and was removed. The remaining 356 respondents represented 45 distinct groups.

An Unanticipated Limitation

A software restriction became evident during analysis. The design of GGUM software includes an upper limit on the number of respondents for a single analysis. GGUM2004 restricts respondents to a maximum of 2,000 (Roberts & Shim, 2008). Given 2,222 respondents in the combined sample, a solution that allowed for comparisons using the same data across models was required. By removing respondents with missing data ($n = 519$), a combined sample size with no missing responses was derived ($N = 1,703$). An analysis was conducted to determine the effect of removing these respondents.

Removing respondents with system missing values had no significant effect on the parameter estimates. The analysis of the 20 items showed 43 to 179 ($M = 96.60$, $SD = 29.61$) missing values per item. The item with the most missing values ($n = 179$) was mlq23; an idealized influence behavioral item. The mlq23, with 8% missing responses, did not seem to be caused by obvious wording or confidentiality concerns. The

comparison between the GRM and the GGUM models was based on the combined sample with no missing data ($N = 1,703$). Classical test theory corrected item-total correlation required list-wise deletion, effectively using the combined sample with no missing data ($N = 1,703$). Therefore, the same sample ($N = 1,703$) was used for comparisons of the classical test theory, the GRM, and the GGUM. For completeness, parallel analyses were performed for the GRM parameter estimates using the combined sample with missing data ($n = 2,222$) and the combined sample with no missing data ($N = 1,703$). Alpha and maximum information location means for the two samples showed no significant difference ($p < .05$). Results of the parallel analyses will be presented in the Discrimination and Difficulty Parameter Estimates section of this chapter.

Assumption Testing

There were six assumptions that were investigated. Assumptions and limitations were discussed in chapter 1 and the purpose, procedures, and guidelines in chapter 3. Interpreting the findings of the assumptions is addressed in chapter 5.

Assumption 1: Translation Accuracy

There was an assumption made that the Hebrew and Russian versions of the MLQ used to gather the samples were correctly translated. Comparing the scores of the translated sample with published untranslated scores can provide some measure of translation effectiveness (Avolio & Bass, 2004a). Percentile scores for subordinate ratings of U.S. norms were nearly identical with the percentiles of the combined sample ($N = 1,703$). Mean difference test was not significant ($p < .05$). Therefore, there was

evidence that the Hebrew and Russian translations effectively conveyed the original constructs.

Assumption 2: Independent Observations

An assumption of independent observations was made. Subordinates rating the same leader should contain enough subjective variation to approximate independent observations. The combined sample ($n = 2,222$) represented 290 independent groups with an average group membership of 9.74 respondents ($SD = 5.62$). An ICCC analysis was performed to determine if there was sufficient evidence to retain individual rater's responses. ICCC examines within and between group variance and tests for independent rater observations. Values above .20 for any item would suggest observations were not sufficiently independent among the subordinates of the same group rating the same leader. All 20 items had significant ANOVA values at $p = .001$. ICCC ranged from .04 to .09, ($M = .07$, $SD = .01$). No item had an ICCC value above .20, indicating that subordinates within a group, rating the same leader, had sufficient individual subjectivity. Therefore, all responses of the individual subordinate raters were retained.

Assumption 3: Sufficient Category Responses

An assumption of sufficient categorical responses was used for matrix computations. A concern with using categorical data is that each category must have enough responses to provide stable estimates. Category by item cell frequency counts relate to matrix algebra stability used in this study's estimation techniques. With cell frequencies below 5, categories should be collapsed for that item. The lowest cell

frequency in the combined sample ($N = 1,703$) was 15 for any category of an item. Therefore, all categories of all 20 transformational leadership items were retained.

Assumption 4: Normal Distribution

An approximation to normal distribution for categorical data was assumed for item level factor analysis in unidimensionality assumption testing. Normal distribution assumption may not be appropriate for dichotomous, nonordered polytomous items, and items of unequal category width (DiStefano, 2002). The selection of any single item category over others influences item difficulty parameter estimates in IRT analysis (Embretson & Reise, 2000). These item difficulties can appear as factors in traditional factor analysis (Wirth & Edwards, 2007). For polytomous variables, such as a 5-point Likert scale, polychoric correlation is one technique used to separate the impact of item difficulty as threshold parameters from item correlations (Flora & Curran, 2004). However, the SPSS (2009) software used in this study did not provide polychoric correlation analysis capability.

Techniques other than polychoric correlation may be used for observed responses. Although Flora and Curran (2004) showed favorable results using polychoric correlation with simulated ordinal data, being a hypothetical estimate, polychoric correlation should be used cautiously with observed data in multivariate analysis (Cohen, Cohen, West, & Aiken, 2003) or not at all (Nunnally & Bernstein, 1994). Wirth and Edwards (2007) suggested using Markov chain Monte Carlo method to avoid these concerns; however, this technique is newer and the software is not readily available.

To provide an indication of techniques used for unidimensionality assumption testing of observed ordinal Likert responses, several pertinent journal articles were examined. Zagorsek et al. (2006) used traditional confirmatory factor analysis with distribution analysis for Leadership Practices Inventory. Scherbaum et al. (2006) used modified parallel analysis, which compares item factor analysis results from observed responses to ideal simulated data using the observed parameter estimates and observed person trait values, for Member–Leader Exchange (Drasgow & Lissak, 1983). Finally, Heinritz et al. (2005) also used modified parallel analysis for an MLQ application of item factor analysis.

Consistent with this study’s proposed approach, an analysis similar to Zagorsek et al. (2006) involving theoretical, normative, and observed response statistical analysis was used. Specifically, theoretical item construction was examined along with distribution of normative data and statistical distribution measures of observed responses to indicate the degree of approximation to normal distribution of each item. The MLQ items have five ordered categories of similar width and are incremental ratios of increasing behavioral frequencies and validated across diverse leadership populations (Avolio & Bass, 2004a). Normative statistics appear to suggest a close approximation to normal distribution with slight negative skewness at the facet level (Avolio & Bass, 2004b).

Finally, Table 1 shows the mean, standard deviation, skewness, and kurtosis of the observed responses to the 20 MLQ study items. On a continuous scale from 0 to 4, a normal distribution would have a mean of 2, a standard deviation of 1, and no skewness or kurtosis. As can be seen in Table 1, all items had means above 2 with slight negative

skewness. Standard deviations ranged around 1, from 0.89 to 1.24. For all items, skewness and kurtosis were less than 1. The theoretical item construction, normative statistics, and observed response distribution measures appear to indicate an approximation to a normal distribution for each item.

Table 1

Distribution Statistics of the MLQ 20 Transformational Items (N = 1,703)

Item	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
mlq10	2.21	1.24	-0.22	-0.89
mlq18	2.78	1.10	-0.76	-0.04
mlq21	2.48	1.07	-0.36	-0.49
mlq25	2.82	1.04	-0.73	-0.04
mlq06	2.45	1.13	-0.36	-0.68
mlq14	2.40	1.16	-0.34	-0.73
mlq23	2.53	1.03	-0.43	-0.32
mlq34	2.73	1.05	-0.60	-0.26
mlq09	2.77	1.02	-0.67	-0.01
mlq13	2.69	1.05	-0.58	-0.28
mlq26	2.22	1.12	-0.20	-0.66
mlq36	2.92	0.89	-0.67	0.20
mlq02	2.51	0.95	-0.26	-0.33
mlq08	2.71	0.98	-0.58	-0.04
mlq30	2.45	1.02	-0.40	-0.26
mlq32	2.53	1.01	-0.39	-0.37
mlq15	2.19	1.18	-0.11	-0.88
mlq19	2.83	1.12	-0.75	-0.25
mlq29	2.46	1.13	-0.39	-0.63
mlq31	2.28	1.12	-0.20	-0.70

Another consideration in determining an approximation of a normal distribution for the 20 items is sample size. DiStefano (2002) reported guidelines for minimum sample sizes with ordered categorical data and asymptotic distributions. For 20 items, 630 responses are considered a minimum sample size. Given the 1,703 responses in this

study, observed responses may be sufficient to approximate a normal distribution, if one exists. Table 1 provides some support for an approximation of normal distribution of each item. If an assumption of normal distribution can be made, confirmatory factor analysis can be employed to test for unidimensionality.

Assumption 5: Unidimensionality

Item-item and item-total correlations were examined in preparation for testing the assumption of unidimensionality. Table 2 shows item-item correlations while Table 3 shows item-total correlations for the 20 MLQ transformational items. Correlation values below 0.20 might indicate more than one leadership construct while values above 0.80 might indicate redundant items. Item-item correlations, from 0.23 to 0.65, and item-total correlations, from 0.54 to 0.71, indicate factor analysis may yield interpretable results.

Item factor analysis showed one dominant transformational leadership dimension and a second minor dimension with eigenvalues above one. Exploratory factor analysis using maximum likelihood estimation with oblimin rotation was performed on the 20 transformational items using SPSS (SPSS, 2009) as shown in Table 4 through Table 7. To provide greater clarity on any violation of unidimensionality implied by the second factor, Scherbaum et al. (2006) recommendation was followed for modified parallel analysis using Drasgow and Lissak (1983) procedure. GRM item parameter estimates and people theta values ($N = 1,703$) of observed responses were used as input to simulate unidimensional response data using WINGEN 3 (Han, 2010).

Table 2
Item-Item Correlations of the 20 MLQ Transformational Leadership Items

Inter-item correlation matrix																					
Item	mlq10	mlq13	mlq21	mlq25	mlq26	mlq34	mlq35	mlq36	mlq39	mlq44	mlq45	mlq46	mlq48	mlq52	mlq53	mlq54	mlq55	mlq56	mlq57	mlq58	
mlq10	1.00	0.42	0.60	0.43	0.43	0.42	0.44	0.45	0.43	0.46	0.44	0.45	0.43	0.45	0.43	0.43	0.45	0.43	0.43	0.43	
mlq13	0.42	1.00	0.47	0.39	0.28	0.26	0.42	0.38	0.27	0.29	0.31	0.34	0.37	0.36	0.39	0.38	0.35	0.38	0.35	0.42	0.38
mlq21	0.60	0.47	1.00	0.52	0.33	0.39	0.52	0.39	0.36	0.34	0.44	0.45	0.43	0.48	0.52	0.43	0.44	0.44	0.44	0.52	0.50
mlq25	0.43	0.39	0.52	1.00	0.34	0.43	0.47	0.41	0.36	0.47	0.51	0.48	0.41	0.39	0.43	0.39	0.38	0.36	0.38	0.38	0.36
mlq26	0.43	0.28	0.33	0.34	1.00	0.51	0.36	0.46	0.34	0.34	0.34	0.31	0.31	0.29	0.30	0.42	0.29	0.23	0.42	0.29	0.23
mlq34	0.45	0.38	0.44	0.46	0.43	1.00	0.42	0.52	0.41	0.52	0.56	0.45	0.37	0.27	0.36	0.34	0.43	0.32	0.43	0.32	0.31
mlq35	0.45	0.42	0.50	0.41	0.38	0.42	1.00	0.39	0.37	0.34	0.42	0.41	0.44	0.43	0.45	0.42	0.39	0.45	0.42	0.39	0.45
mlq36	0.44	0.36	0.39	0.41	0.46	0.51	0.39	1.00	0.39	0.47	0.49	0.51	0.36	0.34	0.39	0.43	0.36	0.33	0.43	0.36	0.33
mlq39	0.45	0.27	0.39	0.36	0.36	0.41	0.37	0.39	1.00	0.46	0.53	0.46	0.32	0.29	0.29	0.31	0.31	0.35	0.29	0.31	0.35
mlq44	0.45	0.29	0.34	0.41	0.48	0.62	0.34	0.47	0.46	1.00	0.54	0.47	0.34	0.31	0.22	0.31	0.39	0.30	0.39	0.30	0.35
mlq45	0.51	0.31	0.44	0.51	0.44	0.56	0.42	0.49	0.52	0.54	1.00	0.49	0.36	0.31	0.36	0.38	0.40	0.34	0.38	0.40	0.34
mlq48	0.45	0.34	0.45	0.45	0.34	0.45	0.47	0.51	0.42	0.46	0.47	0.49	0.43	0.36	0.40	0.40	0.40	0.39	0.42	0.39	0.42
mlq52	0.46	0.37	0.48	0.38	0.34	0.37	0.44	0.36	0.30	0.34	0.36	0.43	0.40	0.45	0.45	0.45	0.41	0.43	0.42	0.43	0.42
mlq53	0.43	0.36	0.48	0.41	0.31	0.37	0.43	0.34	0.39	0.39	0.37	0.43	0.36	0.40	0.51	0.49	0.39	0.38	0.38	0.36	0.41
mlq54	0.45	0.39	0.52	0.39	0.28	0.36	0.45	0.34	0.29	0.32	0.36	0.40	0.45	0.51	1.00	0.57	0.39	0.45	0.39	0.45	0.50
mlq55	0.43	0.36	0.48	0.43	0.30	0.34	0.42	0.39	0.31	0.31	0.36	0.40	0.45	0.49	0.57	1.00	0.43	0.42	0.43	0.42	0.61
mlq56	0.43	0.35	0.44	0.39	0.42	0.43	0.39	0.43	0.31	0.39	0.40	0.47	0.33	0.33	0.39	0.43	1.00	0.42	0.37	0.42	0.52
mlq57	0.49	0.49	0.51	0.38	0.29	0.31	0.45	0.36	0.26	0.30	0.34	0.39	0.43	0.38	0.45	0.42	0.42	1.00	0.65	0.51	0.51
mlq58	0.44	0.38	0.50	0.36	0.23	0.31	0.42	0.33	0.28	0.25	0.32	0.41	0.42	0.36	0.50	0.43	0.37	0.65	1.00	0.58	0.58
mlq59	0.54	0.42	0.54	0.45	0.35	0.42	0.49	0.44	0.35	0.35	0.45	0.45	0.41	0.50	0.61	0.53	0.52	0.53	0.61	1.00	1.00

Table 3

Corrected Item-Total Correlation for the 20 MLQ Transformational Leadership Items

Item	CITC
mlq10	0.71
mlq18	0.55
mlq21	0.70
mlq25	0.63
mlq06	0.54
mlq14	0.64
mlq23	0.63
mlq34	0.62
mlq09	0.55
mlq13	0.59
mlq26	0.64
mlq36	0.64
mlq02	0.61
mlq08	0.58
mlq30	0.63
mlq32	0.63
mlq15	0.60
mlq19	0.63
mlq29	0.59
mlq31	0.71

Note: CITC = Corrected item-total correlation

One dominant factor was confirmed. Table 4 shows a dominant factor with an eigenvalue above 8.0 which explained about 42% of the total variance and a second factor with an eigenvalue above 1.0 with 6% of total variance. These values were then compared to a randomly generated unidimensional data set. The dominant factor had roughly similar eigenvalues and percent variance explained for the observed responses and simulated data.

Violation of unidimensionality was indicated. For modified parallel analysis, an observed secondary factor with eigenvalues higher than the simulated data of 0.42 suggests a violation of unidimensionality. In this study, the secondary factor had an eigenvalue of 1.12 and percent of total variance explained of 5.59, indicating a possible violation of unidimensionality. However, the observed eigenvalue was just above 1.0, indicating a minor secondary factor.

Table 4

Eigenvalues and Percent of Variance Explained in Exploratory Factor Analysis Using Maximum Likelihood Estimation for Observed and Simulated Unidimensional Data

Factors	Observed		Simulated	
	Eigenvalues	% of variance	Eigenvalues	% of variance
1	8.39	41.94	8.32	41.62
2	1.12	5.59	0.42	2.13

Exploratory factor analysis showed a more parsimonious factor structure without oblimin rotation using maximum likelihood estimation. Factor loading test results ($N = 1,703$) are shown in Table 5. Item loading of .40 or higher and a first eigenvalue of 20% or more of total explained variance would indicate support for one higher order construct (Reckase, 1979). Table 5 shows item loadings from .56 and .74 for the observed nonrotated solution for factor one with an eigenvalue of 42% from Table 4, indicating one dominant transformational leadership factor consistent with the MLQ literature. Nonrotated solution of the maximum likelihood estimation was the most parsimonious.

Comparison with randomly generated unidimensional data showed the second factor having generally higher item loadings for observed responses. Table 5 shows, for

no rotation, the second factor had loadings for observed responses from -.44 to .34 and the simulated data were from -.58 to .11. A slightly higher loading for observed nonrotated item responses than for simulated data indicated a possible violation of unidimensionality. Further evidence of a unidimensional violation, shown in Table 6, was indicated through comparing goodness of fit metrics.

Table 5

Loadings for Exploratory Factor Analysis Using Maximum Likelihood Estimation With and Without Oblimin Rotation for Observed and Simulated Data

Item	No rotation				Oblimin rotation			
	Observed		Simulated		Observed		Simulated	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
mlq10	0.73	-0.02	0.73	0.06	0.45	-0.35	0.73	-0.01
mlq18	0.57	0.17	0.54	0.03	0.56	-0.04	0.52	-0.03
mlq21	0.72	0.19	0.73	0.06	0.70	-0.07	0.72	-0.02
mlq25	0.65	-0.05	0.66	0.11	0.36	-0.35	0.70	0.05
mlq06	0.56	-0.30	0.63	-0.58	0.00	-0.64	0.05	-0.83
mlq14	0.67	-0.39	0.62	0.11	-0.03	-0.80	0.66	0.06
mlq23	0.65	0.09	0.63	0.11	0.53	-0.18	0.68	0.06
mlq34	0.64	-0.21	0.63	0.04	0.16	-0.55	0.61	-0.03
mlq09	0.58	-0.22	0.58	-0.01	0.11	-0.54	0.51	-0.09
mlq13	0.62	-0.44	0.56	0.05	-0.12	-0.84	0.56	0.00
mlq26	0.67	-0.31	0.68	0.08	0.06	-0.70	0.69	0.01
mlq36	0.66	-0.11	0.66	0.07	0.30	-0.44	0.67	0.00
mlq02	0.63	0.12	0.64	0.00	0.55	-0.13	0.59	-0.08
mlq08	0.60	0.12	0.61	0.07	0.53	-0.11	0.63	0.01
mlq30	0.66	0.27	0.65	0.02	0.75	0.05	0.61	-0.06
mlq32	0.65	0.22	0.65	0.08	0.69	-0.01	0.66	0.01
mlq15	0.63	-0.01	0.58	0.09	0.39	-0.29	0.61	0.04
mlq19	0.65	0.29	0.66	0.06	0.76	0.08	0.66	-0.01
mlq29	0.63	0.34	0.64	0.04	0.81	0.15	0.62	-0.03
mlq31	0.74	0.23	0.76	0.07	0.75	-0.04	0.76	-0.01

For the same degrees of freedom ($df = 151$), the observed responses had a much higher chi-squared value than the randomly simulated data, as shown in Table 6. Such a large difference may indicate that one or more minor factors are influencing the model. In addition, the factor correlations for maximum likelihood estimation using oblimin rotation are shown in Table 7. The factor correlation matrix shows similar factor relationships for the observed and simulated responses.

Table 6

Goodness of Fit Statistics for Exploratory Factor Analysis Using Maximum Likelihood Estimation for Observed and Simulated Unidimensional Data

	Observed	Simulated
χ^2	1254.89	128.64
χ^2/df	8.31	0.85

Evidence for violation of unidimensionality assumption was indicated. One dominant transformational leadership factor and a second minor factor appear to be influencing the factor structure of observed responses. However, Kirisci et al. (2001) demonstrated that MULTILOG, used in this study, was robust to some unidimensional violations including three dominant factors with intercorrelations between true thetas of .6. The experimental condition used by Kirisci et al. appear to have been a more severe unidimensional violation than for one dominant factor and one minor secondary factor as in this study. The robustness of MULTILOG to unidimensional violations was analyzed using MODFIT and reported in the Best IRT Model section of this chapter.

Another method to test for unidimensionality is to conduct confirmatory analysis. There were two commonly found factor models described in articles for the MLQ in the

last decade. These two structural models of transformational leadership items were analyzed using maximum likelihood estimation with AMOS (SPSS, 2009). Model one, depicted in Figure 2, with one higher order transformational leadership construct, includes three intermediary facets of charisma, intellectual stimulation, and individual consideration. Charisma, in model one, is directly associated with four items on idealized influence attributed, four items on idealized influence behavioral, and four items on inspirational motivation. Model two is similar to model one; however, the charisma facet is replaced by the three facets: idealized influence attributed, idealized influence behavioral, and inspirational motivation. Both models were analyzed using maximum likelihood estimation.

Table 7

Factor Correlation Matrix for Exploratory Factor Analysis Using Maximum Likelihood Estimation with Oblimin Rotation for Observed and Simulated Unidimensional Data

Factor	Observed		Simulated	
	1	2	1	2
1	1.00	-0.67	1.00	-0.63
2	-0.67	1.00	-0.63	1.00

The two models involved a single higher order transformational factor. Model fit indices CFI, RFI, and NFI, should be at or above .95 for a good model fit, with RMSEA at or below .06 (Hu & Bentler, 1999). For model one, with three facets and one higher order transformational construct, the fit indices were $CFI = .88$, $RFI = .86$, $NFI = .87$, and $RMSEA = .08$ ($X^2 = 2,108.31$, $df = 167$). For model two, with five facets and one higher order transformational construct, the fit indices were $CFI = .87$, $RFI = .84$, $NFI = .86$, and

$RMSEA = .09$ ($X^2 = 2,297.65$, $df = 165$). No direct comparison was possible as these two models represent different constructs. However, the purpose was not to select a specific model, only to establish model metrics, as both models contained a single higher order transformational leadership construct.

Given the exploratory and confirmatory factor analysis results, one dominant transformational leadership dimension was tentatively supported. SPSS and AMOS indicated one higher order transformational leadership construct through item-item and item-total correlation values between 0.20 and 0.80. Further, 42% of total variance was explained by a single eigenvalue for the 20 transformational items and loadings for all 20 items were above the .40 guideline using maximum likelihood estimation with no oblimin rotation. Both models found in literature included a single higher order transformational leadership construct and fit metrics indicate a moderate degree of fit. While the presence of a minor second factor violates unidimensionality, MULTILOG appears to be robust to these violations (Kirisci et al., 2001). IRT data to model fit using MODFIT software provides further information to determine the extent of any unidimensional violations. Therefore, all 20 items were used for analysis in the GRM, the GGUM, and classical test theory, for a perceived transformational leadership dimension.

In AMOS, variables are treated as continuous. One consideration for the results of model fit being below the guideline for good fit was that AMOS historically was not designed for categorical items (Antonakis et al., 2003). Review of the user manual and online help features of AMOS did not provide clarity. DiStefano (2002) showed that for

categorical data, moderate levels of negative bias occurred using maximum likelihood in confirmatory factor analysis.

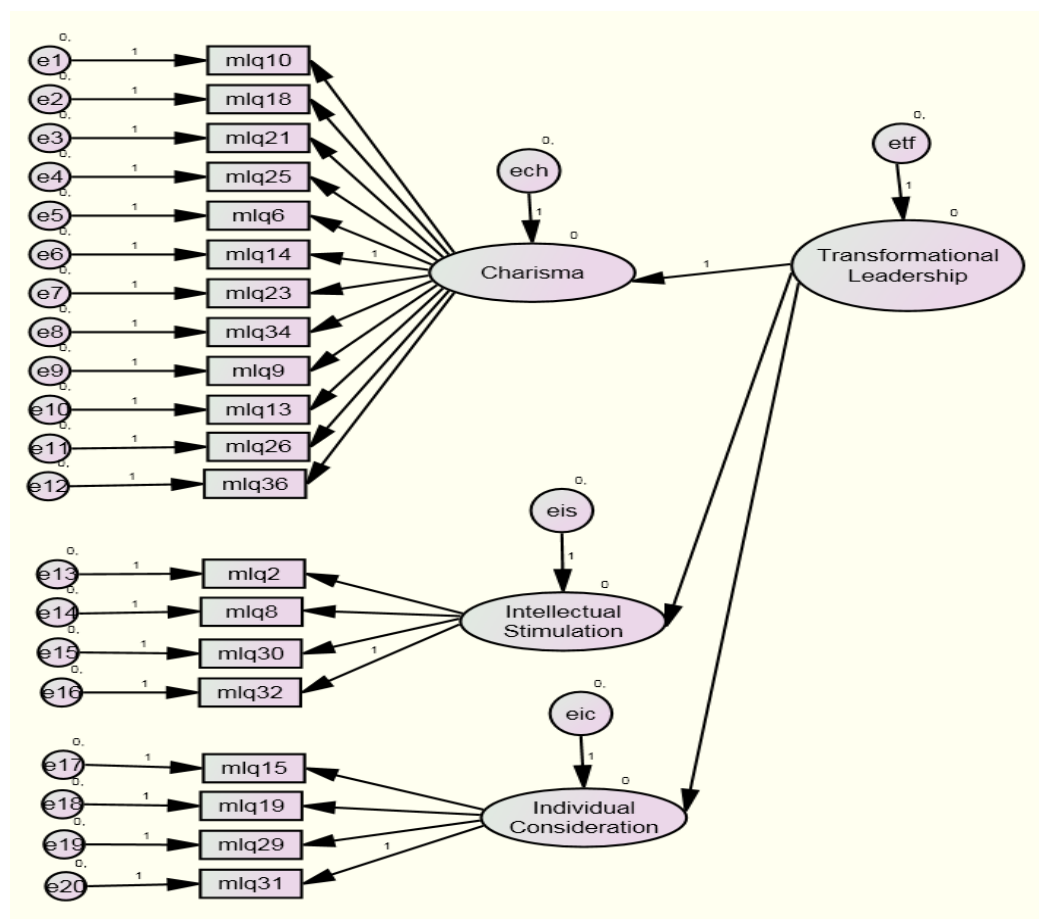


Figure 2. Model one: AMOS transformational leadership structural model showing one higher order factor and three facets of charisma, intellectual stimulation, and individual consideration.

Assumption 6: Sample Homogeneity

The final investigation involved differences among the corporate samples and the athletic sample. Normative samples for the MLQ did not address athletic samples (Avolio & Bass, 2004a). This study was the first to examine mean transformational theta values for an athletic sample. Table 8 shows the classical test theory means and standard

deviations for each sample. The 20 item mean score for the telecommunications sample ($n = 1,248$) was highest, followed by the 26 companies sample ($n = 161$), and the basketball players were lowest ($n = 294$) on a scale from zero to four. The total score means followed the same pattern on a scale from zero to 80.

IRT provides an alternative measure to classical test theory's mean total scores for average sample ability. Using IRT techniques, samples can be equated on the same theta scale with the advantage of direct comparison. MULTILOG provides such a procedure following techniques described by Thissen et al. (2003). The technique of equating involves selecting an anchor sample whose sample mean is set to zero with a standard deviation of one normalized, such as a z score. All other samples' ability means are then computed in relation to the anchor sample mean ability using a single MULTILOG analysis.

Table 8

Item and Total Score Means and Standard Deviations for Three Samples

	Telecommunication		26 companies		Basketball players	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
20 items	2.59	1.06	2.55	1.02	2.36	1.07
Total score	51.79	14.54	51.03	13.68	47.3	12.83

Note: Telecommunication $n = 1,248$, 26 companies $n = 161$, basketball players $n = 294$, item scale from zero to four, total score scale from zero to 80.

Table 9 shows the IRT ability mean and standard deviation for each sample of the subjective leader ratings. IRT analysis produces person parameter estimates that are invariant, as are item parameter estimates. Therefore, once measured, the means of each

sample in relation to the other samples is no longer dependent on the test items (Embretson & Reise, 2000).

The two corporate samples had similar perceived transformational ability means. The second sample, comprised of executives from 26 Israeli companies, was close to the mean of the telecommunication sample ($M = -0.06$, $SD = 0.94$). This might be expected, as both samples were from corporate settings. Therefore, the mean rater's perception of their leader's transformational ability in the corporate samples was roughly equivalent.

The athletic sample's mean significantly was different. The third sample was from Israeli professional basketball players which had a lower average mean ($M = -0.37$, $SD = 0.85$). Their coaches' transformational leadership abilities were perceived as less, on average, than both corporate samples. This result is perhaps not surprising given that the basketball coaches may have been rated on leadership behaviors exclusive of the team's captain, often an active court leader.

Table 9

Mean Sample Theta Differences Using Telecommunications Sample as the Anchor

	Telecommunication		26 companies		Basketball players	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Theta	0	1	-0.06	0.94	-0.37	0.85

Note: Telecommunication $n = 1,248$, 26 companies $n = 161$, basketball players $n = 294$, scale from 0 to 4.

Having differences in mean sample abilities is similar to having third graders and fourth graders taking the same math test. The items are the same. However, the groups taking the test show different math abilities. In the same manner, the corporate and

athletic samples can show different mean transformational abilities, while the item parameter estimates stay the same.

Parallel analyses were conducted to verify item parameter invariance. Appendix A shows the GRM parameter estimation using the two corporate samples without ($n = 1,409$) and with ($N = 1,703$) the basketball player sample. Appendix B shows data to model fit metrics for the GRM and the GGUM without the basketball players' sample ($n = 1,409$) and adjusted to the normative sample size of 3,000 (Drasgow et al., 1995). Mean values of 3.0 and below are considered excellent fit (Drasgow et al., 1995). Comparison of alpha means, maximum information location means, and beta range means, showed no significant difference ($p < .05$). Invariance of mean parameter estimates was supported.

Having completed data preparation, an unanticipated limitation, and testing assumptions, the five research questions were explored. The response to these questions generally follows a pattern of exploring the categories of an item first, then the item, and concluding with the impact on the 20 item transformational leadership subscale. The final question compares IRT with classical test theory for the MLQ's transformational leadership subscale.

Research Question 1: Observed Versus Expected IRT Model Responses

The first research question asks about the degree of overlap between the observed responses and the expected responses for the GRM and the GGUM, starting at the category level. Said another way, the research question asks how well the IRT models represented the actual data. IRT analysis compares models at the category, item, group of items, and test levels. MODFIT produced fit plots for each of five categories, of all 20

items, for both models. Instead of producing all 200 fit plots, Figure 3 shows a typical example comparing the GRM with the GGUM for item mlq14, response category three, *fairly often*.

Of the two models, the GRM came closest to approximating a normal distribution. The solid lines of Figure 3 trace each model's prediction of the response function. The observed responses are the same for both models, since the data used was identical. The vertical lines centered on the observed data, represent 95% confidence intervals. At the positive theta of 2.0, the model differences were more noticeable. At a theta of 2.0, the GRM probability, P , was .19 and for the GGUM, $P = .25$. The GGUM assumes that a positive bias towards the leader contributes to a higher probability of choosing category three, *fairly often*, for higher trait levels. At least for this category, the data did not seem to support the GGUM's positive bias assumption. Notice that the GGUM 95% confidence interval for a theta of 2.0 did not include the predicted trace line.

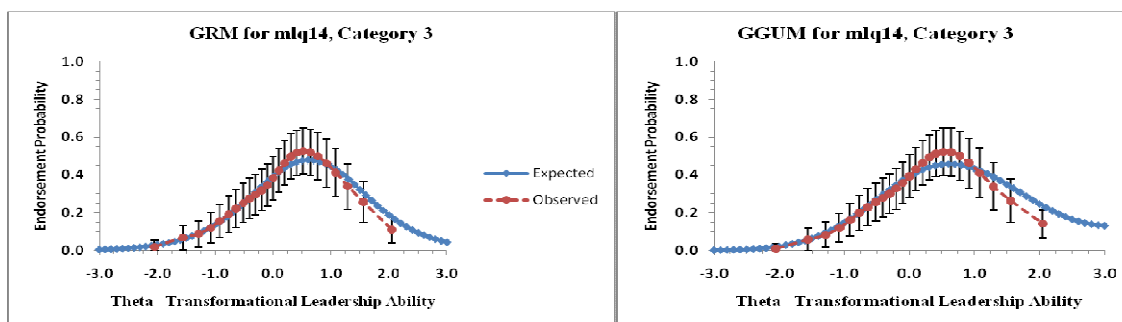


Figure 3. Side by side comparison of the GRM and the GGUM fit plots.

The difference between IRT models is subtle. Figure 4 shows trace lines of each expected category function for item mlq14 with the GRM on the left and the GGUM on the right. Category three, *fairly often*, shows the GGUM ending higher ($P = .13$) than the

GRM ($P = .04$), at a theta of 3.0. In a similar fashion, the GGUM is slightly higher ($P = .24$) for category one, *once in a while*, than the GRM ($P = .16$), at theta of -3.0. The higher probability GGUM at a theta of -3.0 was due to the bias against the leader contributing to a higher probability of choosing a negative category.

Chi-squared metric provides a quantitative measure of data to model fit difference. The accumulated differences between the expected model response functions and the observed responses are measured as chi-squared distributions. Table 10 shows the singlet test of chi-squared and chi-squared over degrees of freedom for each of the 20 transformational leadership items. The best fit metric is a three or lower chi-squared over degrees of freedom mean across all 20 items.

The IRT models are dissimilar using the mean item difference test. The GRM ($M = 0.14$, $SD = 0.10$) and the GGUM ($M = 0.24$, $SD = 0.14$) chi-squared over degrees of freedom mean for all items were significantly different ($p = .007$). For the GGUM items with higher chi-squared values than the GRM, those differences were larger. For instance, item mlq30 was higher for the GGUM ($X^2 = 2.60$) than the GRM ($X^2 = 0.21$).

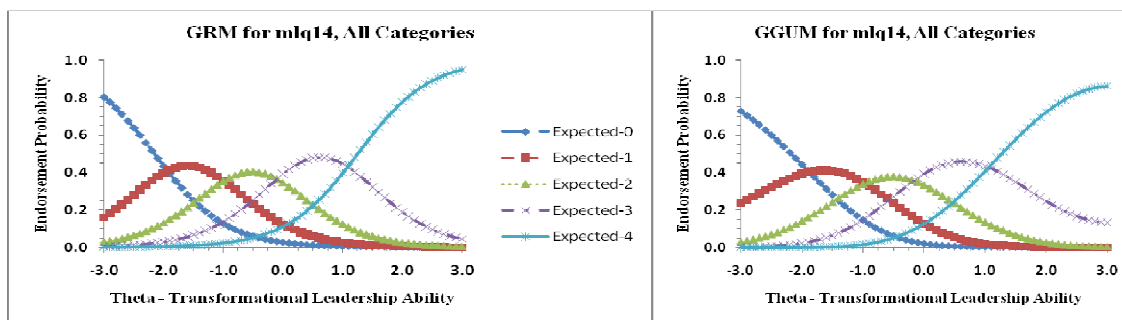


Figure 4. Side by side comparison of the GRM and the GGUM expected response functions for each of the five categories.

Table 10

The GRM and the GGUM Single Item Fit Metrics (N = 1,703, df = 4)

Item	Facet	GRM		GGUM	
		χ^2	χ^2/df	χ^2	χ^2/df
mlq10	IIA1	1.01	0.25	1.49	0.37
mlq18	IIA2	0.27	0.07	0.51	0.13
mlq21	IIA3	0.21	0.05	1.67	0.42
mlq25	IIA4	1.79	0.45	1.07	0.27
mlq06	IIB1	0.56	0.14	0.25	0.06
mlq14	IIB2	0.22	0.06	0.73	0.18
mlq23	IIB3	0.63	0.16	0.91	0.23
mlq34	IIB4	0.67	0.17	0.80	0.20
mlq09	IM1	0.68	0.17	0.61	0.15
mlq13	IM2	0.36	0.09	0.52	0.13
mlq26	IM3	0.40	0.10	0.78	0.20
mlq36	IM4	1.08	0.27	1.07	0.27
mlq02	IS1	0.16	0.04	0.63	0.16
mlq08	IS2	0.32	0.08	0.57	0.14
mlq30	IS3	0.21	0.05	2.60	0.65
mlq32	IS4	0.89	0.22	0.83	0.21
mlq15	IC1	0.63	0.16	0.48	0.12
mlq19	IC2	0.69	0.17	1.12	0.28
mlq29	IC3	0.26	0.07	0.59	0.15
mlq31	IC4	0.31	0.08	1.78	0.44
<i>M</i>		0.57	0.14	0.95	0.24
<i>SD</i>		0.40	0.10	0.56	0.14

Note. The MLQ facets = idealized influence attributed, idealized influence behavioral, inspirational motivation, intellectual stimulation, and individual consideration.

The first research question explored the difference between the observed subordinate responses and the expected responses described by the GRM and the GGUM. The GRM expects the respondent to start at the lowest category and proceed to higher categories until a selection is made that coincides with their subjective view of their leader's transformational leadership ability as depicted in the item statement (Samejima,

1969). The GGUM expects the respondent to start either at the lowest or highest category and proceed toward the middle depending on the subordinate's subjective bias toward the leader (Roberts & Shim, 2008). Both models seemed to predict the observed responses fairly well in the singlet test with mean chi-squared over degrees of freedom values at or below three.

There is more to model fit testing. In examining the difference between observed responses and expected responses at the category, item, and 20 item test levels, both models represented the actual responses fairly well. The GRM performed better with a lower average chi-squared over degrees of freedom value, than the GGUM. There are additional steps that need to be taken when comparing items along the transformational ability scale. These additional steps involve matching two or more items of different difficulties and will be discussed as part of the second research question.

Research Question 2: Best IRT Model

The second research question asks which of the two IRT models used in this study best represents the response patterns of the combined sample ($N = 1,703$). Although the mean and standard deviation of all 20 items in the previous research question was indicative of the answer, doublet and triplet tests provided additional comparisons using MODFIT (Stark et al., 2001). As with the singlet test, a mean chi-squared over degrees of freedom value of three or less indicates excellent fit between the observed responses and the expected responses.

The doublet test examines balanced pairs of items. Instead of measuring the data to model difference for each item individually, as in the singlet test, the doublet test

compares two items at opposite ends of the theta continuum, an easy item matched with a hard item. Comparing two items at different points on the ability axis, for observed versus expected responses, allows these differences to achieve a form of mean weighting (Drasgow et al., 1995). For instance, MODFIT compares mlq10 with mlq18 for the GRM which had an approximate ability location center at -0.17 and -0.92, respectively. Although not opposites on the scale, the MLQ's 20 transformational items had relative betas below 0.0 for both models, necessarily limiting the theta range available for matching. This may have produced an over sensitive doublet test, however, the same sensitivity applies to both IRT models.

The triplet test examines items spread across the available range. The triplet test adds a middle theta item to the doublet comparison, balancing the extremes of opposite theta locations. For instance, one triplet test involved mlq36, mlq25, and mlq14 for the GRM. The item's relative betas were -1.33, -1.02, and -0.49, respectively. By adding mlq25, centered at -1.02, a more evenly weighted middle is included for comparing observed versus expected model differences.

Both models showed excellent data to model fit. The results of these singlet, doublet, and triplet tests are presented in Table 11 for the combined sample ($N = 1,703$). The mean singlet test for the GRM was 0.14 ($SD = 0.10$) and the GGUM was 0.24 ($SD = 0.14$). The mean doublet test for the GRM was 2.11 ($SD = 0.84$) and the GGUM was 2.99 ($SD = 1.21$). The mean triplet test for the GRM was 1.56 ($SD = 0.37$) and the GGUM was 2.04 ($SD = 0.65$). In each case, the mean chi-squared over degrees of freedom values

were better for the GRM than the GGUM. However, both models are deemed an excellent fit given the mean guideline criteria of three or below (Drasgow et al., 1995).

Table 11

Comparing the Fit of the GRM and the GGUM (N = 1,703)

	GRM frequency table of X^2/df							Mean	SD
	<1	1<2	2<3	3<4	4<5	5<7	>7		
Singlet	20	0	0	0	0	0	0	0.14	0.10
Doublet	1	13	7	2	1	0	0	2.11	0.84
Triplet	0	11	1	0	0	0	0	1.56	0.37
	GGUM frequency table of X^2/df							Mean	SD
	<1	1<2	2<3	3<4	4<5	5<7	>7		
Singlet	20	0	0	0	0	0	0	0.24	0.14
Doublet	0	5	8	7	3	1	0	2.99	1.21
Triplet	0	7	4	1	0	0	0	2.04	0.65

The GGUM is typically used with a different Likert scale. The GGUM scale is usually anchored between *strongly disagree* and *strongly agree*, where item bias may be more pronounced (Roberts & Shim, 2008). In this study, a behavioral frequency scale anchored between *not at all* and *frequently, if not always* seems to be a new application for the unfolding model. The subjectivity of scales and items will be discussed further in Chapter 5.

The GGUM did not meet guidelines when extrapolated to a larger sample size. Table 12 shows chi-squared metric of observed versus expected difference for a sample size extrapolated to 3,000 responses (Drasgow et al., 1995). Only the GRM remains at or below the mean chi-squared over degrees of freedom guideline value of three. This mathematical extrapolation to 3,000 responses allows other researchers using different

assessments to compare on a similar sample size basis. This use of a normative 3,000 sample size is especially useful as chi-squared metric is sensitive to sample size (DiStefano, 2002). However, as noted, the narrow negative range of relative beta locations creates an over sensitivity in the doublet and triplet tests. In the doublet test, the GGUM mean value is over the guideline of three ($M = 4.50$, $SD = 2.13$). The GRM mean value for the doublet test is just under the guideline value of 3.0 ($M = 2.96$, $SD = 1.49$).

Overall, the GRM was a better model fit than the GGUM. In the singlet, doublet, and triplet test, the GRM had lower mean differences and standard deviations than the GGUM. The GRM model was better at explaining the responses to the 20 transformational leadership items for all comparative analyses. Therefore, the GRM will be used for the remaining research questions with the GGUM values presented in the appendices.

Table 12

Comparing the GRM and the GGUM Fit Adjusted to $n = 3,000$

GRM frequency table of X^2/df									
	<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	SD
Singlet	20	0	0	0	0	0	0	0.00	0.01
Doublet	1	4	9	7	1	1	1	2.96	1.49
Triplet	0	6	5	1	0	0	0	1.99	0.65
GGUM frequency table of X^2/df									
	<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	SD
Singlet	20	0	0	0	0	0	0	0.02	0.09
Doublet	0	2	3	5	7	5	2	4.50	2.13
Triplet	0	2	6	3	0	1	0	2.84	1.15

Research Question 3: Discrimination and Difficulty Parameter Estimates

The third research question asks for the discrimination and difficulty parameter estimates of the MLQ's 20 transformational leadership items. Parameter estimates for the GRM were calculated using MULTILOG 7.0 (Thissen et al., 2003). Program defaults were changed to 91 quadrature points for more precise estimation and an increase in estimation cycles to allow convergence (M. Edwards, personal communication, June 24, 2010). IRT marginal reliability was .94. Table 13 depicts estimates, in logistic form, of the GRM discrimination value alpha, the four categorical boundary beta values, and the location along theta of the maximum IIF values for two samples ($n = 2,222$, $N = 1,703$) of the 20 transformational items.

The 20 items are easier than average. Item parameters were estimated for the GRM using combined samples with missing data ($n = 2,222$) and with no missing data ($N = 1,703$) in Table 13. Only comparative sample results ($N = 1,703$) will be discussed. Total beta range across items was -3.37 to 1.47 with a relative beta mean of -0.73 ($SD = 0.33$), indicating generally easier behavioral items than the 0.0 average. An item of average difficulty is defined by the theta scale, like a z-score, with a mean of 0.0 and standard deviation of 1.0. IIF maximum locations ranged from -1.94 to 0.01 ($M = -1.22$, $SD = 0.49$). The maximum IIF location is the point where the information for an item peaks and is derived from category boundary values. The discrimination parameter estimates ranged from 1.28 to 2.24 ($M = 1.72$, $SD = 0.27$), indicating higher discrimination than the 1.0 average slope.

Table 13

The GRM Item Parameter Estimates for Two Samples

Items	Facet	GRM including system missing data ($n = 2,222$)						GRM incomplete data removed ($N = 1,703$)					
		α	δ_1	δ_2	δ_3	δ_4	IIF	α	δ_1	δ_2	δ_3	δ_4	IIF
mlq10	IIA1	2.24	-1.49	-0.75	0.14	1.14	-0.85	2.23	-1.54	-0.75	0.17	1.20	-0.86
mlq18	IIA2	1.47	-2.58	-1.69	-0.60	0.78	-1.69	1.45	-2.65	-1.74	-0.60	0.81	-1.75
mlq21	IIA3	2.18	-2.19	-1.17	-0.10	1.09	-1.21	2.17	-2.21	-1.14	-0.06	1.17	-1.16
mlq25	IIA4	1.80	-2.57	-1.58	-0.58	0.71	-1.55	1.77	-2.66	-1.61	-0.57	0.76	-1.56
mlq06	IIB1	1.36	-2.51	-1.20	-0.06	1.38	-0.75	1.28	-2.72	-1.26	-0.07	1.45	-0.71
mlq14	IIB2	1.72	-2.12	-1.02	-0.05	1.21	-0.71	1.67	-2.16	-1.04	-0.01	1.25	-0.76
mlq23	IIB3	1.84	-2.47	-1.37	-0.21	1.24	-1.45	1.76	-2.54	-1.37	-0.16	1.31	-1.37
mlq34	IIB4	1.72	-2.63	-1.53	-0.48	0.89	-1.38	1.63	-2.81	-1.57	-0.47	0.96	-1.32
mlq09	IM1	1.35	-3.08	-1.91	-0.58	1.06	-2.02	1.35	-3.19	-1.92	-0.60	1.05	-1.94
mlq13	IM2	1.49	-2.78	-1.58	-0.44	1.07	-1.41	1.45	-2.95	-1.63	-0.46	1.10	-1.30
mlq26	IM3	1.73	-1.95	-0.89	0.25	1.57	-0.97	1.72	-2.04	-0.89	0.28	1.60	-0.88
mlq36	IM4	1.85	-3.20	-1.99	-0.82	0.81	-1.93	1.86	-3.37	-2.00	-0.78	0.85	-1.88
mlq02	IS1	1.63	-3.13	-1.56	-0.12	1.46	-1.48	1.62	-3.13	-1.57	-0.06	1.53	-1.54
mlq08	IS2	1.57	-3.01	-1.80	-0.50	1.08	-1.92	1.55	-3.05	-1.79	-0.48	1.16	-1.86
mlq30	IS3	1.84	-2.42	-1.32	-0.06	1.43	-1.47	1.85	-2.40	-1.31	-0.04	1.47	-1.47
mlq32	IS4	1.77	-2.67	-1.40	-0.18	1.34	-1.33	1.78	-2.69	-1.34	-0.16	1.36	-1.17
mlq15	IC1	1.52	-2.10	-0.79	0.28	1.46	-0.09	1.52	-2.09	-0.75	0.34	1.55	0.01
mlq19	IC2	1.80	-2.40	-1.46	-0.58	0.55	-1.22	1.84	-2.44	-1.45	-0.57	0.55	-1.16
mlq29	IC3	1.62	-2.30	-1.18	-0.13	1.20	-0.99	1.62	-2.32	-1.16	-0.08	1.23	-0.96
mlq31	IC4	2.27	-1.92	-0.87	0.11	1.27	-0.83	2.24	-1.93	-0.83	0.18	1.33	-0.77

Note. All GRM values in logistic metric, IIA = idealized influence attributed, IIB = idealized influence behavioral, IM = inspiration motivation, IS = intellectual stimulation, IC = individual consideration, α = discrimination, δ_i = category boundaries, IIF = location along theta for the maximum value of the item information function.

The alpha value represents the slope of a given item; its discrimination. With steeper slopes, sharper differentiations can be made between respondent's latent abilities.

The discrimination values are shown in Table 13. All alpha values were above the 1.0 standard for normally discriminating items.

It is not straightforward to interpret the GRM beta category boundaries. Figure 5 provides a graphical indication of alpha and relative beta parameters for the 20 item

subscale. Beta is shown as a single, relative difficulty value, δ_r , for each item and is the average of the four category beta values δ_1 to δ_4 . Beta values represent the point where the probability of two adjacent category selections are equally possible ($P = .50$). The labels for each item represent one of four items of a transformational facet where *IIA* identifier is for idealized influence attributed, *IIB* is idealized influence behavioral, and *IM* is inspiration motivation. These three facets were part of the charisma factor identified in Figure 2. In addition, identifier *IS* represents intellectual stimulation and *IC* for the individual consideration facet.

In Figure 5, the four idealized influence attributed items ranged in relative beta from -1.04 to -0.23, idealized influence behavioral from -0.97 to -0.49, and inspirational motivation from -1.33 to -0.26. Four intellectual stimulation items ranged from -1.04 to -0.57 and individual consideration from -0.98 to -0.24. Using relative beta as an indicator, leaders, whose abilities ranged from about -1.4 to 0.5, should find these 20 MLQ items relatively reliable in differentiating their transformational leadership ability.

Item modifications are indicated. Figure 5 is useful for posing questions about which items provided the least amount of additional information. A candidate item, mlq6, is from the idealized influence behavioral facet marked as *IIB1*. *IIB1* was lower in alpha than mlq14 marked *IIB2* or mlq23 marked *IIB* with similar relative betas. This suggests that *IIB2* and *IIB3* represented the idealized influence behavioral facet with greater discrimination over a relatively similar theta range than *IIB1*. If *IIB1* was modified with a relative beta above 0.0, the facet and subscale would benefit from increased reliability and information content. Therefore, item mlq6 (*IIB1*) is a candidate for modification.

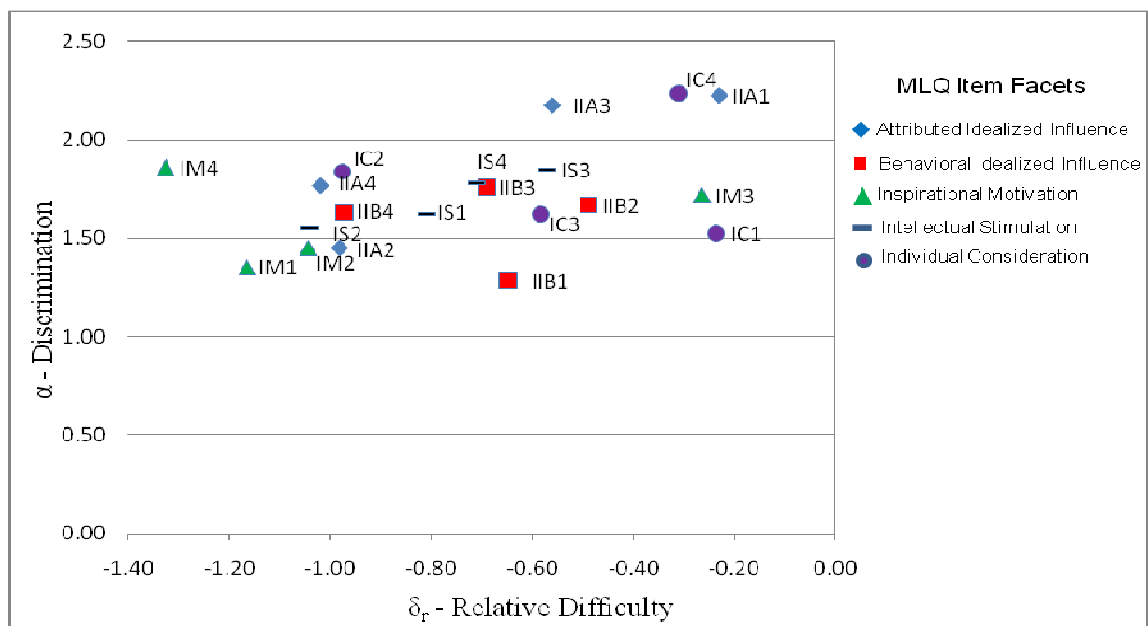


Figure 5. The GRM item parameters estimates ($N = 1,703$) in logistic metric.

In the same manner, four additional items may be candidates for modification. Items mlq9, mlq18, mlq2, mlq15, marked *IM1*, *IIA2*, *IS1*, and *IC1*, respectively, are represented by other items of the same facet with larger alphas similar relative beta values. Together with mlq6, each of the redundant items represented one of the five facets of the 20 item subscale. If these five items were reworded difficulty above 0.0, the reliability of each facet and the information for the 20 item subscale would increase.

The interpretation of the GGUM unfolding parameters is not intuitive. For comparison purposes, the GGUM parameter estimates are provided in Appendix C. Program defaults were used with the GGUM2004. IRT marginal reliability was .94. Appendix C shows the GGUM parameter estimates for discrimination alpha, location parameter beta, and four subjective response thresholds. In addition, the location of the maximum IIF is shown for each item. Interpretations of the GGUM beta and threshold

values are not directly comparable to the GRM values. For example, item mlq10, an idealized influence attributed item, had a beta of 3.14 and a first subjective response threshold of $\tau_1 = -5.17$. With $\tau_1 = (\theta - \delta)$ or $-5.17 = (\theta - 3.14)$, then $\theta = -2.03$. This means that a leader with a subjective transformational ability below -2.03 would have a higher than 50% probability of being marked as *not at all* by a subordinate rater in response to the first idealized influence attributed item.

The GGUM alphas had relatively average discriminations. The slope α , of the GRM items had a higher range of 1.28 to 2.24 ($M = 1.72$, $SD = 0.27$) compared to the GGUM range of 0.72 to 1.49 ($M = 1.08$, $SD = 0.22$). The higher GRM item slopes show that at the crossover point, from endorsing one category to the adjacent category, there was more information for the GRM to differentiate leaders. The GRM IIF maximum locations ranged from -1.94 to 0.01 ($M = -1.22$, $SD = 0.49$) and the GGUM IIF maximum locations ranged from -1.64 to -0.11 ($M = -1.06$, $SD = 0.41$) on the same theta scale. The GRM had a greater range of information than the GGUM, for the 20 item subscale.

The third research question was answered by exploring parameter estimates of the 20 transformational leadership items. Alphas, betas, and IIF maximum location values were estimated, presented, and interpreted. The data for the 20 items used in this study were generally below average in difficulty and higher in discrimination. Further item and subscale information results will be presented as part of the next research question.

Research Question 4: Highest Trait Range Reliability Estimation

The fourth research question asks about the information content of the 20 items as a subscale and asks how the subscale information relates to reliability and standard error

of measurement. In order to answer these questions, it is useful to return to the item level results that build up to the 20 item subscale findings. Figure 6 shows the item characteristic curves of mlq14 and the associated IIF.

Item information is greater at category boundaries. The information content for the 20 item transformational leadership subscale is the simple additive information content of each item, at every theta point (De Ayala, 2009). Each item's information increases at category boundaries. At the intersection of two category boundaries, such as 0 and 1 in Figure 6, a respondent with that theta value of -1.98, had the same chance of selecting either category ($P_0 = .41, P_1 = .41$). The other category selections are less probable ($P_2 = .14, P_3 = .04, P_4 = .00$). On either side of the category boundary, there was information about which category, 0 or 1, a respondent was likely to select. As the respondent selected one of the two categories, the relative theta of their response became more certain. Therefore, someone that selected category one on mlq14 was likely to have a theta higher than -1.98 but less than the next higher category boundary at -1.0.

Category boundaries mark points of decision for respondents. Over multiple items, a respondent's theta becomes more reliably known. For highly discriminating items, the range between category boundaries narrows, with positive kurtosis, allowing more precision in person ability estimates. Therefore, category boundaries provide additive information about a person's ability location.

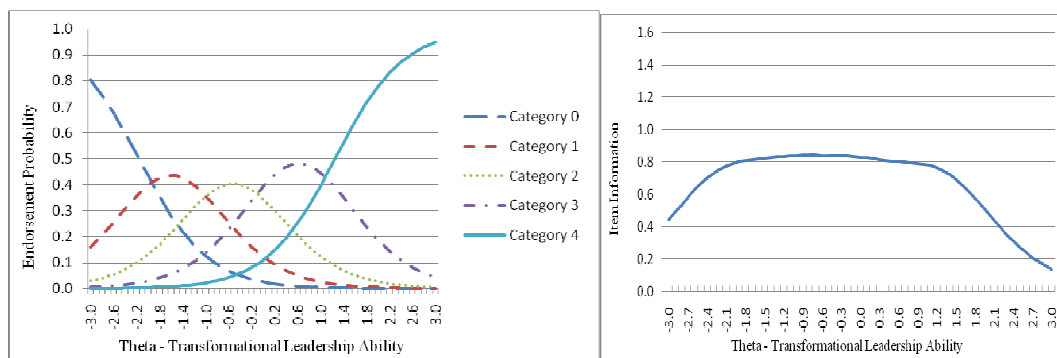


Figure 6. Side by side comparison of the GRM item characteristic curves and IIF for mlq14, an idealized influence behavioral item ($N = 1,703$).

The GGUM IIFs had positive kurtosis and positive skewness. Appendices D through H show the GRM graphs of characteristic curves on the left hand side and the corresponding IIFs on the right, for each of the 20 items. Appendices I through M show the corresponding GGUM graphs for characteristic curves and IIFs. Each Appendix represents four items of a facet. Examination of the IIF graphs reveals positive kurtosis and positive skewness for the GGUM versus the GRM. In general, the GGUM had higher item information values over reduced theta ranges than the GRM.

Standard error and item information are mathematically related. The information for an item adds to other items' information along the theta scale to form the total information function of the 20 item subscale. The information function for all 20 items is called the *test information function* and is shown in Figure 7 for the GRM, using the left hand axis. Also shown on Figure 7, using the right hand axis is the corresponding standard error of the measure. Standard error is the reciprocal of the square root of information along the theta axis (Embretson & Reise, 2000). The test information function's maximum was 17.68 at $\tau = -1.0$. The standard error is the reciprocal of the

square root of 17.68, or 0.24. The 20 item subscale's standard error of measure is shown as the lower dashed line.

Reliability goes up as the standard error goes down. As can be seen in Figure 7 using the right hand axis, the standard error of the 20 item subscale was relatively low, about 0.25, from theta of -2.3 to 1.2. Reliability is calculated as one minus the square of the standard error of measure (Embretson & Reise, 2000). For a standard error of .25 the reliability is .94. The test information function for the GGUM is shown in Appendix N.

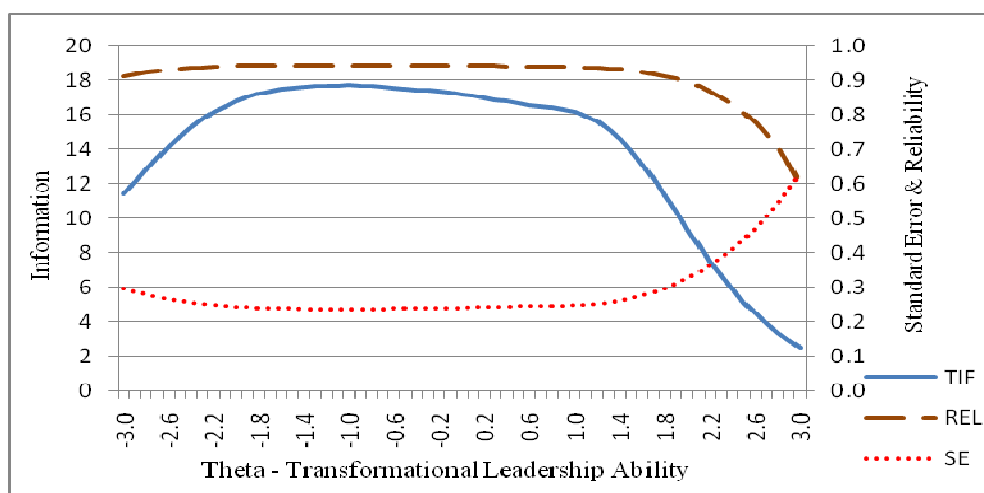


Figure 7. The GRM test information function, standard error, and reliability for all 20 items of the perceived transformational leadership ability ($N = 1,703$).

The 20 items of the MLQ had a standard error of measure that varied with theta. An expanded trait range from -2.7 to 1.5 had a standard error of not more than 0.27 and a reliability of not less than .93, with maximum information at a theta of -1.0. As Figure 7 shows, at the upper latent trait range, standard error increases and reliability decreases with a steep slope. At a theta of 3.0, the reliability was only .60. From a leadership

perspective, the 20 item subscale best measured the transformational leadership abilities within a theta range from -2.4 to 1.3 at a standard error of .24 and a reliability of .95.

Research question four concerned the information and standard error for the entire transformational leadership subscale. Standard error changed at a slower, reciprocal rate as the information content changed. For the 20 transformational items of the MLQ, greater than 1.5 standard deviations above the sample mean, the standard error increased quickly. Although reliability's relation to standard error was discussed, the next research question explores reliability in more detail.

Research Question 5: IRT Versus Classical Test Theory Reliability

The fifth and last research question asked for the reliability estimation differences between classical test theory analysis and IRT for the MLQ's transformational leadership subscale. Classical test theory uses many test level descriptive indicators. For instance, the respondents' total score of the sample with no missing responses ($N = 1,703$) had a mean of 50.94, a standard deviation of 14.27, and a mode of 52. The skewness was -0.33, kurtosis was -0.21, and the total score range was 2 to 80. The overall internal consistency of the 20 item subscale was .93. The internal consistency of item-total with item deleted was .93 for all items.

Classical test theory showed above average discrimination and easier items. The classical test theory indices of item discrimination and difficulty are presented in Table 14. Discrimination is measured by corrected item-total correlation (Scherbaum et al., 2006). Values above 0.5 are more discriminating. Classical test theory item discrimination ranged from 0.54 to 0.71 ($M = 0.62$, $SD = 0.05$), indicating more

discriminating items. Classical test theory item difficulty is measured by score means. Mean values above the midpoint of 2.0 are considered easier items. For the 20 items of this study, item difficulty ranged from 2.19 to 2.92 ($M = 2.55$, $SD = 0.22$), indicating easier items.

Table 14

Classical Test Theory Item Analysis: Corrected Item-Total Correlations for Discrimination, Mean Item Scores for Difficulty, and Cronbach's Alpha for Reliability

Item	Facet	CITC	<i>M</i>	<i>SD</i>	α - item deleted
mlq10	IIB1	0.71	2.21	1.24	0.93
mlq18	IIA2	0.55	2.78	1.10	0.93
mlq21	IM1	0.70	2.48	1.07	0.93
mlq25	IS2	0.63	2.82	1.04	0.93
mlq06	IM2	0.54	2.45	1.13	0.93
mlq14	IC3	0.64	2.40	1.16	0.93
mlq23	IC1	0.63	2.53	1.03	0.93
mlq34	IS1	0.62	2.73	1.05	0.93
mlq09	IIB4	0.55	2.77	1.02	0.93
mlq13	IIA4	0.59	2.69	1.05	0.93
mlq26	IS4	0.64	2.22	1.12	0.93
mlq36	IC2	0.64	2.92	0.89	0.93
mlq02	IS3	0.61	2.51	0.95	0.93
mlq08	IIB3	0.58	2.71	0.98	0.93
mlq30	IIB2	0.63	2.45	1.02	0.93
mlq32	IM4	0.63	2.53	1.01	0.93
mlq15	IM3	0.60	2.19	1.18	0.93
mlq19	IIA3	0.63	2.83	1.12	0.93
mlq29	IIA1	0.59	2.46	1.13	0.93
mlq31	IC4	0.71	2.28	1.12	0.93

Note. CITC = corrected item - total correlation, IIA = idealized influence attributed, IIB = idealized influence behavioral, IM = inspiration motivation, IS = intellectual stimulation, IC = individual consideration.

IRT and classical test theory share similar item discrimination patterns. For instance, mlq10 and mlq31 had the highest corrected item-total correlations of .71. The IRT discrimination parameters for these two items were also the highest at 2.23 and 2.24, respectively. More generally, the relative ranking of MLQ items was roughly the same, from lowest to highest discrimination.

A single distribution measures the classical test theory item difficulty parameter. The classical test theory parameter estimates are therefore, limited. For instance, item mlq14 in Table 14 shows a classical test theory difficulty mean score of 2.40 and a standard deviation of 1.16. For additional distribution information, classical test theory also provided a skewness of -0.34 and a kurtosis of -0.73. The findings from classical test theory related to a single difficulty distribution. In IRT, each category had its own probability distribution as shown in Figure 6, five per item.

IRT calculates the distribution function for each category along the theta scale. The category zero is a monotonically decreasing slope, categories one through three are similar to normal distributions, and category four is a monotonically increasing function, as shown in Figure 6. Category distributions intersect at boundaries b_1 to b_4 measured by mean and standard error metrics. Table 13 showed mlq14 mean values for category boundaries $b_1 = -2.16$ ($SE = 0.11$), $b_2 = -1.04$ ($SE = 0.06$), $b_3 = -0.01$ ($SE = 0.05$), $b_4 = 1.25$ ($SE = 0.07$). Providing distribution information at a category level provides greater precision in reliability estimates over the latent trait range than classical test theory.

Reliability is not a constant, although for comparison, IRT provided a marginal reliability of .94. In Table 14, for classical test theory, all items had the same 0.93

reliability, unrelated to theta values. Instead of a single reliability value, IRT calculates an information function along with a standard error function that varies over theta for each item and for the 20 item subscale, as was seen in Figure 7. Reliability, which is one minus the square of the standard error, decreases quickly at the top of the theta range for the 20 item subscale. For instance, Figure 7 shows that from a theta from 2.0 to 3.0, the reliability decreased from .89 to .60. Classical test theory calculated a single value, .93, for reliability of the 20 item subscale with no ability to incorporate theta. Without being able to model the effect of theta changes, classical test theory must qualify results under the conditions in which the results were recorded. IRT parameters, however, are invariant because they completely describe the item, independent of the measurement conditions.

Unlike classical test theory, item difficulty locations and the perceived transformational leadership abilities are described on the same x-axis. The IRT scale is similar to a z-score metric with a mean of 0 and a standard deviation of 1. IRT estimates for mlq14 showed respondents had the maximum probability of choosing category one for their leader's behaviors at a theta of -1.7, those choosing category two at -0.6, and category three at 0.6, on the perceived transformational leadership scale.

Item parameters and person abilities are invariant. In classical test theory, item mlq14 could not be added to a different transformational leadership assessment such as the Leadership Practices Inventory and have the same parameters. New parameters would have to be calculated as classical test theory considers any item change a new test. IRT item parameters are transportable. Given the calibration sample, each item's discrimination and difficulty parameters are independent of any other item in the test or

of a new test, dependent only on measuring the same construct. The same is true of person abilities.

IRT person ability invariance has practical applications. The IRT analysis presented so far had been primarily concerned with item parameter estimates. Individual abilities, along the latent trait axis, may also be determined with a standard error of measure. Those individual abilities are independent of the original test and original conditions because ability parameter estimates retain the uncertainty as part of the standard error of measure. The following example, shown in Figure 8, illustrates the usefulness of determining individual abilities with precision.

Precise individual ability differentiation is not available with classical test theory total score method. The improved differentiation of transformational leadership ability using IRT parameters versus classical test theory analysis can be seen in Figure 8. Classical test theory uses total score to make criterion based personnel decisions. In keeping with this tradition, the total score mode of 52 forms the center of Figure 8 and depicts an excerpt of the MLQ respondents ($N = 1,703$) who rated their leader's transformational abilities. Two total score points on either side of this mode, from 50 to 54, are depicted as horizontal lines. The total score scale is represented by the y-axis. On the x-axis is theta, the perceived transformational leadership ability as derived by IRT estimates for the leaders. Therefore, each vertical mark represents a leader being rated.

Total score method leads to cutoff errors. For each total score, the leaders who were rated the same total score did not have the same transformational leadership ability. There is as much as a third of a standard deviation separating leaders' abilities for the

same total score. If the MLQ had a transformational leadership subscale cutoff criterion of 52, with the intention of promoting all leaders who possessed at least average transformational leadership abilities, there would be leaders promoted who did not possess at least average abilities. Also, some leaders would be promoted having less than average abilities and others would not be promoted that had more than average transformational abilities. It is similar to being unaware of the extent of Type I and Type II error while making personnel decisions that impact careers.

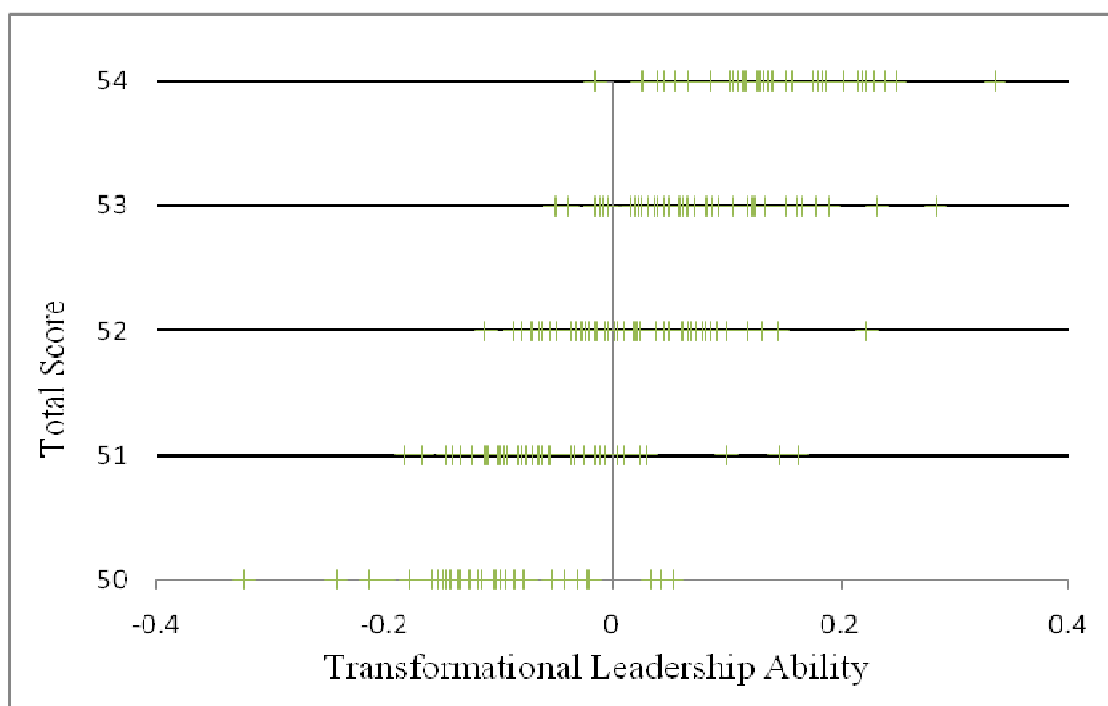


Figure 8. Classical test theory total score versus IRT theta estimates measuring raters' perception of their leaders' transformational leadership ability on the 20 MLQ items.

There can be errors in using classical test theory total score method alone. For instance, the individual leader in Figure 8 with a total score of 52 was the farthest to the left, at about -0.1, on transformational leadership ability scale. That individual would

have passed a cutoff criterion of 52 and above, however, they were generally less able on the leadership trait than most of the leaders scoring 51 and less able than half of those scoring 50. It is this greater precision of detecting the latent ability of individuals that sets IRT apart from classical test theory.

IRT provides a reliability measure for individual ability scores. The reliability of individual latent ability estimates varies by theta. Table 15 shows an excerpt of IRT person parameter estimates for the combined sample ($N = 1,703$) along with classical test theory total score values for comparison. Each IRT leader's perceived transformational leadership ability estimate also had a standard error of measure. That standard error changed depending on the precision of item information. For instance, the leader being rated by respondent 766 had a transformational ability estimate of -0.5, indicating a half of a standard deviation below the average of the sample. The standard error of measure was 0.22 which equates to a reliability of .95.

Table 15

A 20 item Excerpt of the Classical Test Theory Total Score Plus the IRT Individual Ability Theta and Standard Error Estimates for the MLQ Respondents ($N = 1,703$)

ID	Score	θ	SE
763	17	-2.01	0.22
764	68	+0.99	0.23
765	80	+2.68	0.45
766	44	-0.50	0.22
767	77	+1.99	0.31
768	60	+0.48	0.24
769	35	-1.04	0.22
770	77	+2.01	0.31

The high reliability value was due to many items having category crossover points in the -0.5 theta region. Therefore, the precision of determining this individual's perceived transformational leadership ability was relatively high. For the responses from the combined sample ($N = 1,703$), the highest reliability of a person parameter estimate was .96 over a theta range of -1.7 to 1.3 and decreased to no less than .95 over a range of -2.4 to 1.3. However, the leader being rated by respondent 765 had an ability estimate of 2.68 or over two and a half standard deviations above the mean of this sample. Figure 7 showed that the MLQ's 20 item subscale had a low reliability, about .75, at a theta of 2.68. Classical test theory had only a total score measure which did not vary by theta.

The fifth and last research question explored the difference in reliability between classical test theory and IRT. IRT models retained item category and individual latent ability precision. Therefore, IRT provided specific reliability measures across theta which classical test theory did not. This increased precision is why IRT is used in assessment construction, validation, modification, and person's ability detection, as has been shown with the 20 transformational leadership items of the MLQ.

Summary

The usefulness of IRT in determining item and person parameter estimates has been demonstrated for the 20 items of the MLQ's transformational leadership ability. The data preparation of the three samples was instrumental in identifying three data input irregularities in the sample of Israeli executives of 26 companies. Removal of self-reports from the samples of an Israeli telecommunications company and the executives of 26 companies allowed the research to consider only direct subordinate ratings. Both

corporate samples and the professional basketball players' sample had respondents that were removed due to lack of any information for all 20 of the MLQ items. These changes provided a total combined sample ($n = 2,222$) of subordinates rating their leaders on perceived transformational leadership ability.

An unanticipated limitation arose in running the GGUM2004 software. There was a maximum limit of 2,000 cases per analysis. Given a sample size of 2,222, a decision was made to remove the cases with system missing values. Therefore, a combined sample with no missing values ($N = 1,703$) was used throughout this study for comparisons between IRT models and classical test theory. Testing the difference between these two combined samples ($n = 2,222$, $N = 1,703$) showed no significant effect. Where helpful for illustration purposes, the combined sample with system missing data ($n = 2,222$) was evaluated and presented alongside results from the comparison sample ($N = 1,703$).

Several tests were conducted to determine the appropriateness of five key assumptions. The first assumption was testing the validity of the Hebrew and Russian MLQ translations used in obtaining the archival data. Comparative metrics were examined to U.S. norms published by Avolio and Bass (2004a). Results showed that the combined sample was not significantly different to the published percentile scores of U.S. norms for subordinate raters.

The second assumption tested was one of independent observations. Respondents rating the same leader within a group are not independent. However, there was sufficient subjectivity in each subordinate's perspective of their leader's transformational leadership behaviors to support using all rater responses. The latent trait or theta being

measured, therefore, was the subordinate's perception of their leader's transformational leadership ability rather than the leader's actual ability.

The third assumption was sufficient responses in all categories of all items to make analyses meaningful. Categories must be collapsed when responses from raters are below five. The smallest number of responses to any category of any item was 15 for the combined sample used for analysis and comparison purposes ($N = 1,703$).

The fourth assumption was testing for unidimensionality. Exploratory factor analysis was performed on observed and simulated unidimensional data. IRT assumption of unidimensionality was not supported. One dominant factor and one minor factor had eigenvalues above one. All 20 items of the MLQ's transformational leadership subscale had loading above .40 in nonrotated maximum likelihood estimation with the dominant factor representing 42% of the total variance. Confirmatory factor analysis using maximum likelihood estimation was used on two different models found in literature. The model with one higher order transformational leadership factor and three lower order facets best fit the data. MULTILOG was found robust (Kirisci et al., 2001) to the levels of unidimensionality violation described in this study as all 20 items showed chi-squared over degrees of freedom values at or below three for all GRM conditions.

The fifth and final investigation was the comparison of mean ability values of the three samples. The difference between the three Israeli person parameter means was tested using the largest sample as an anchor. The two corporate samples were roughly equivalent; however, the mean of the basketball players' sample was 0.37 standard deviations lower. IRT analysis was conducted with and without the basketball sample for

both IRT models. The results showed no significant difference. Possible interpretations for the difference in corporate and athletic mean transformational leadership ability perceptions will be discussed in Chapter 5.

Having completed the data preparation and making research decisions based upon software usage and assumption testing, the five research questions were explored. The first of these research questions showed how IRT can be used graphically and quantitatively to determine the degree of observed versus expected model fit. Initial indications were that the GRM best represented the respondents' perceptions of their leaders' transformational abilities.

The second research question provided additional quantitative measures to determine which IRT model best fit the responses. The GRM best represented the observed responses from subordinates. The GGUM was also an excellent fit except when the sample size was adjusted to a 3,000 size benchmark. For the extrapolation to 3,000 cases, only the GRM continued to be an excellent fit. That both IRT models adequately represented the responses will be discussed in Chapter 5.

The third research question was used to determine the GRM item parameter estimates. Both discrimination and difficulty parameters were estimated. The 20 transformational items of the MLQ were generally easier than average in difficulty and more than average in discrimination. Greater information was available, therefore, to those whose leaders were perceived to be from at least one standard deviation below average to average in transformational leadership. The subscale was best at

differentiating leaders in this range of transformational abilities. Chapter 5 will discuss IRT parameter comparisons to another transformational leadership assessment.

The fourth research question involved examining the information content and reliability for the 20 item subscale. The range associated with higher information content and a reliability of .94 was from -2.3 to 1.2 standard deviations either side of the mean for the combined sample ($N = 1,703$). There was more information to differentiate participants at the low end than high end, of perceived transformational leadership ability.

The fifth and final research question compared reliability for classical test theory and IRT of the MLQ's 20 item transformational subscale. Results were presented to show that IRT had greater precision at the item, subscale, and individual ability levels. Specifically, IRT was shown to reduce errors in differentiating latent abilities than classical test theory total score method. A personnel example of cutoff criterion was presented in Figure 8, comparing classical test theory total score method with IRT individual latent trait estimates.

Further discussion of results will be presented in Chapter 5. Having detailed the findings of this study, Chapter 5 will discuss possible interpretations. In addition, implications for social change will be reviewed along with recommendation for action and further research. These topics and others form the basis for Chapter 5.

Chapter 5: Discussion, Conclusions, and Recommendations

Overview

The MLQ has been used to detect harmful leadership behaviors (Judge & Piccolo, 2004). For instance, behaviors that are considered avoidant, coercive, and corrective have been shown to lead to decreased satisfaction, loss of effectiveness, and reduced job satisfaction (Avolio & Bass, 2004a). Avolio and Bass (2004a) did not provide advice or standards to detect harmful transformational leaders, relying instead on the unit weighting of four individual consideration items. Little has been written about detecting transformational leaders who are potentially harmful. Khoo and Burch (2007) were an exception and conducted a preliminary study finding that harmful transformational leaders might be detected. The MLQ facet analysis showed that a combination of high idealized influence attributed and low individual consideration scores correlated with narcissistic transformational leadership. One problem with the research from Khoo and Burch was the use of composite facet measures rather than individual item parameters. Precision was lost in Khoo and Burch's classical test theory approach to detecting harmful transformational leaders.

IRT analysis for the MLQ was not available. This study was designed to expand the understanding of the MLQ's 20 item transformational leadership subscale using IRT. By increasing the knowledge of item and person parameters for the MLQ, detecting harmful transformational leaders may be improved. Therefore, the 20 transformational leadership items were analyzed with two IRT models using a combination of three Israeli archival samples ($N = 1,703$).

Research questions led to the results, presented in chapter 4, that the MLQ's 20 item transformational leadership subscale is better at differentiating lower level abilities. Broadly, the research questions covered three objectives: (a) test the fit of IRT models for the 20 item MLQ transformational leadership subscale, (b) estimate the IRT parameters for each of the 20 items, and (c) evaluate changes in the reliability estimation of scores when using IRT versus classical test theory analysis. Research questions one and two examined both graphical and quantitative measures of the IRT models, at the category and item level, to determine whether the GRM or the GGUM was a better fit to the response data. The 20 item transformational leadership responses from the combined sample ($N = 1,703$) showed the GRM to be the best model for all conditions tested.

Quantitative item and subscale parameter estimates were considered for research questions three and four. These IRT measures answered the second objective of calculating the parameter estimates using the GRM. Results showed that items were more discriminating but easier than average. Therefore, this subscale would not be suitable to differentiate those whose transformational abilities were greater than 1.2 standard deviations above the mean.

Supplementing classical test theory with IRT analysis achieved the best overall results for reliability precision. The third objective coincided with the final research question. Classical test theory was used to test numerous assumptions. IRT was used for category, item, and subscale precision measures. It was the combination of the two approaches that achieved the results of this study. Although classical test theory showed a constant reliability across items for all respondents, IRT analysis demonstrated that the

reliability of the subscale information varied from .94 to .60 along the transformational leadership ability scale. IRT models retain precision at the category, item, and subscale levels.

Interpretation of Findings

The results from chapter 4 have implications that are discussed in this section. The organization follows the outline of chapter 4. Where appropriate, results from chapter 4 and literature from chapter 2, is referenced to provide context and justification for interpretations.

Data Preparation

Of the combined samples, 16% of the initial data were discarded. Primarily, the responses removed were indirect ratings or self-ratings. A concern with inadequate sample size was described in chapter 2. Literature suggested a sample size above 500 for two parameter models (Reise & Yu, 1990). With an initial combined sample size of 2,222, the parameter estimates were stable indicated by higher IRT reliabilities of items and abilities than with classical test theory for a limited part of the trait range. It is not known if IRT reliabilities would increase significantly with much larger sample sizes.

An Unanticipated Limitation

The reduction in combined sample size to 1,703 was necessitated by an upper limit of 2,000 responses for the GGUM2004 software. The 23% of eliminated data were all partial responses. It was not known the cause of the incomplete responses. Generally, larger sample sizes produce more stable estimates with lower standard error of measure and, therefore, higher reliabilities (Emberetson & Reise, 2000). Post hoc tests for the

GRM on the initial sample of 2,222 produced similar item reliability values as the 1,703 sample across the latent trait. Perhaps the added information of the 2,222 sample did not increase the reliability due to the amount of missing responses or because the difference in sample size was not significant.

Assumption Testing

Assumption of unidimensionality. Violation of IRT assumption of unidimensionality was supported in Table 4 and Table 6 of chapter 4. One dominant primary factor and a minor secondary factor were reported. Kirisci et al. (2001) asserted that MULTILOG was robust to unidimensionality violations greater than shown in this study. The stability of item and parameter estimates shown by reliability estimates at or above .94 and χ^2/df values below the guideline of three, indicated support for the robustness of MULTILOG. The conclusion is that the violations of unidimensionality found in this study did not negatively impact the stability of item parameters or ability estimates.

Assumption of sample homogeneity. Assumption testing in Table 9 of chapter 4 found the basketball player sample to be at least 0.31 standard deviations below the mean of both corporate samples for perceived transformational leadership ability. The conclusions in this study, on average, are that basketball players rated their coaches as having less transformational leadership ability than subordinates rated their corporate leaders. There are several plausible explanations for these results. One possible reason, noted in chapter 4, was that coaches and captains are often two different persons; splitting the role of leader. Another possible reason might be, in general, that coaches had less

transformational leadership ability. Perhaps, as a population, leaders seeking careers in professional athletic coaching had less latent transformational abilities. In other words, the 20 MLQ items measured what they were supposed to measure and the basketball players perceived their coaches as less transformational because coaches were, on average, less transformational.

A further possible interpretation might be that the MLQ's 20 item subscale did not measure transformational coaching well. The explanation could be that the MLQ's item development did not take into account athletic responses. Athletic samples were not listed in the MLQ's 1999 or 2003 norms (Avolio & Bass, 2004a). In athletics, a winning coach may exhibit transformational behaviors that are, in some essential manner, different than corporate, academic, and military leaders, which were used for normative samples (Avolio & Bass, 2004a). Said another way, the transformational leadership construct assessed by the MLQ's 20 item subscale may be sufficiently different, thereby reducing the average athletic coaches' perceived transformational abilities.

Finally, basketball players might not be similarly transformed by their coaches. The possibility is that the perception of what is transformational may be different for different populations. The professional players might be driven more by other factors such as their own perceived merits, rankings, and publicity and be less influenced or notice a coach's transformational behaviors than an employee would take notice of their supervisors in corporations. Players may also not fully identify with transformational behaviors of the coaches. These possible explanations are not exhaustive or mutually exclusive. The proposed reasons may combine to lower the average professional

basketball players' perceptions of their coaches' transformational abilities. Further research is needed to determine some of the reasons for any differences in mean perceived ability levels of athletic versus other leader populations.

Research Question 1: Observed Versus Expected IRT Model Responses

How do the observed responses differ from IRT models' expected patterns for each of the five categories of the MLQ's 20 transformational leadership items? MODFIT category graphs ($n = 200$) of observed responses were similar to the GRM and the GGUM expected patterns, as shown in the example graph of Figure 3 in chapter 4. Specifically, of the 2,500 expected category data points graphed for each model, the GRM had seven and the GGUM had nine, outside the 95% confidence interval. The conclusion was that the GRM and the GGUM closely approximated the MLQ's 20 transformational leadership category responses.

Research Question 2: Best IRT Model

Which of the selected IRT models best represents the response patterns observed in the sample? Although the GRM was a slightly better fit than the GGUM, the comparison was close; with analysis from Table 11 and Table 12 in chapter 4 showing both models fit reasonably well. The GRM has X^2/df values below three for singlet, doublet, and triplet adjusted tests and the GGUM did not. The GGUM assumes that respondents approach an item from either end of the scale and move toward the middle. The GRM assumes the respondents start at the lower end of the scale and move upward. One explanation for two different models being able to describe the observed data similarly relates to results shown in Figure 7 and Appendix N, that the subscale

information peak location and maximum for both the GRM (-1.0, 17.68) and the GGUM (-1.0, 19.53), were similar. Therefore, although the models predict slightly different beginning and ending points, the middle is much the same. It is the middle categories that represent the majority of responses and model weighting. The similarity of the middle category prediction between the two models may help to explain the general similarity of model to data fit results. The conclusion of the results is that the GRM fits the observed data better than the GGUM.

Scale type may matter with the GGUM. The observed frequency of transformational leadership behavior represented by the scale anchored at each end by *not at all* and *frequently, if not always*, may not have evoked sufficient subjective differentiation for the GGUM to detect. Typically, the GGUM is used with a scale anchored at either end by *strongly disagree* and *strongly agree* (Roberts & Shim, 2008). In addition, items can be written to appeal to feelings rather than observed behaviors. For instance, Scherbaum et al. (2006) described an item on the Leader-Member Exchange assessment used to compare the GRM with the GGUM, “I like my supervisor very much” (p. 378). The mlq30 from Avolio and Bass was “Gets me to look at problems from many different angles” (2004a, p. 107) has less of an emotional appeal. Although the construct is entirely different, the sense of asking for an emotionally subjective response is clearer in the Leader-Member Exchange item. Evoking emotional responses to a self-rating scale may provide significantly increased discrimination for the GGUM. Scherbaum et al. concluded that the GGUM was a better fit than the GRM for Leader-Member Exchange self-assessment using the emotionally subjective item and scale combination. The MLQ’s

transformational leadership items appeared not to evoke high emotive responses from subordinates than the Leader-Member Exchange, decreasing the predictive accuracy of the GGUM over the GRM.

Research Question 3: Discrimination and Difficulty Parameter Estimates

What are the discrimination and difficulty parameters of each of the MLQ's 20 transformational leadership items? Discrimination and difficulty parameter estimates for the GRM were shown in Table 13 and Figure 5 of chapter 4. Item discrimination ($N = 1,703$) ranged from 1.28 to 2.24, above the 1.00 relative scale average. The conclusion is that the MLQ's 20 transformational leadership items are generally better at distinguishing between individual trait abilities than average (Embretson & Reise, 2000). The implication for higher item discrimination, as noted by De Ayala (2009), is that professional application of detection using the 20 items of this study allows higher reliability estimates and, therefore, greater confidence in behavioral predictability.

Although discrimination was above average, the overall information content was low. For instance, the GRM analysis for the Leader-Member Exchange showed the maximum information was 45.15 with 12 items (Scherbaum et al., 2006), compared to the MLQ's 20 item maximum of 17.68; from Figure 7 in chapter 4. However, another transformational leadership assessment, the Leadership Practices Inventory, appeared to have a maximum total information function of approximately 19 with 30 items (Zagorsek et al., 2006), similar to the MLQ's. Transformational leadership assessments with behavioral scales may have lower maximum information. Even if the current information content for the MLQ's 20 item subscale was low, additional discrimination from

modifications at the upper end of the scale could certainly increase the information content.

For item difficulty, from Table 13 and Figure 5 of chapter 4, the highest category boundary ($N = 1,703$) was 1.60 for mlq26 and the lowest was -3.37 for mlq36, both inspirational motivation items. Category responses centered below the middle of the trait range for all 20 items. The conclusion is that the difficulty of all 20 items is relatively easy (Zagorsek et al., 2006). The ease of answering the 20 items was related to subordinates perceiving greater frequency of observed transformational leadership behaviors (Avolio & Bass, 2004a). An implication of all 20 items being easier is that upper level trait abilities remain undifferentiated (Zagorsek et al., 2006). Said another way, those at the top of the perceived transformational leadership range cannot be as accurately measured on their ability as those in the low to middle levels. Only modification of the study items will increase reliability of detection at upper levels of perceived transformational leadership.

Recommendations for modification of the 20 item MLQ transformational subscale was described in chapter 4. IRT analysis showed that all 20 items were easier than average. The five items, one from each facet, were candidates for modification. Modifying mlq2, mlq6, mlq9, mlq15, and mlq18 to increase difficulty would improve differentiation at higher thetas, all else being equal. As described in the Discrimination and Difficulty Parameter Estimates section of chapter 4, modification to increase difficulty above a theta of 1.5 would improve the information content and reduce standard error. Loss of information would be noticeable in lower theta ranges if these

item were removed, however, the information gained by placing the items above a theta of 1.5 would increase the overall subscale detection effectiveness (Zagorsek et al., 2006). Item modifications, using IRT, is discussed as part of Recommendations for Further Research in this chapter.

Research Question 4: Highest Trait Range Reliability Estimation

What portion of the transformational leadership trait range has the highest reliability estimates? For item reliability, Figure 7 of chapter 4 shows a Cronbach's alpha of .94 from a trait range of -2.3 to 1.2. The conclusion is that IRT analysis does not provide one reliability number for the entire range as does classical test theory (Samejima, 1977b). IRT analysis increases precision by reporting reliabilities associated with specific trait ranges. Further, IRT item parameters are invariant and can be directly applied by future researchers (Embretson & Reise, 2000). The implication is that all of the future research with different participants and incorporating one or all of the 20 items can change without impacting each item's parameter estimates from this study.

Person ability estimates, reported in chapter 4, was a Cronbach's alpha of .95 for a trait range of -2.4 to -1.3 and .96 from a narrower range of -1.7 to 1.3. The conclusion is that prediction of a leader's perceived transformational leadership ability, within the range from -2.4 to 1.3, meets the guidelines for minimal reliability in detection (Nunnally & Bernstein, 1994). The implication is that this study provides the selection reliability for harmful transformational leaders' detection.

Research Question 5: IRT Versus Classical Test Theory Reliability

What are the differences in reliability estimation of the MLQ's transformational leadership subscale using IRT versus classical test theory analysis? In Chapter 4, Figure 7 shows IRT item reliability estimates and Table 14 reports the classical test theory item measures. IRT reliability estimates vary by trait range with maximums at .94 for items and .96 for abilities. The classical test theory measures are a constant .93 for items and provide no ability estimates. Classical test theory provides one single value (.93) for all 20 study items across the entire trait continuum. The constancy of the classical test theory measure is due to averaging the reliability across the trait range, losing precision (Samejima, 1977b). The conclusion is that the IRT has greater reliability for item and ability estimates. The implication is that IRT can provide improved detection of transformational leaders using higher reliability estimates than classical test theory. Figure 8 is an example of utilizing IRT in a professional selection application to precisely identify transformational leadership abilities versus classical test theory alone.

In the Implications for Social Change section of this chapter, an application for the detection of harmful transformational leaders is presented. This application demonstrates, as is described throughout this study, that IRT is effective in conjunction with classical test theory, not independently (Embretson & Reise, 2000). Many of the IRT assumptions are supported or rejected by classical test theory analysis (Samejima, 1977a). Detection of harmful transformational leader may require an initial facet cutoff score from classical test theory analysis before applying item level and ability level IRT analysis; such as with the suggested selection criteria. Each theory has unique strengths

and limitations (De Ayala, 2009). The conclusion is that IRT and classical test theory are mutually supportive and should be used together in professional detection and intervention applications and for further assessment research efforts.

Additional Finding Interpretation

Comparison of the MLQ findings from Chapter 4 with other transformational leadership assessments is the final topic of this section. The only other study found that performed an IRT analysis of a transformational leadership assessment was the 30-item Leadership Practices Inventory by Zagorsek et al. (2006). Zagorsek et al. provided IRT GRM data for five facets, each with six items rather than treating all 30 items as one transformational dimension. The GRM fit statistics showed the mean to be less than three for all facets ($N = 801$). Only one facet, encouraging the heart, was more than three for the adjusted fit metric ($n = 3,000$). All 30 items had relative location parameters between -4.0 and 0.0, with most between -2.0 and 0.0, indicating easier items. Discrimination parameters ranged from 0.75 to 1.81 ($M = 1.25$, $SD = 0.27$). Maximum total information function was estimated at 19. Reliability ranged from .64 to .91. The IRT results mentioned were for the Leadership Practices Inventory by Zagorsek et al. (2006). The IRT data presented provides an opportunity to compare transformational assessments on an item and construct basis.

The published data from Zagorsek et al. (2006) for the Leadership Practices Inventory was used for this comparison. The Leadership Practices Inventory facets had better GRM fit statistics than the MLQ's transformational subscale. Although both assessments had easier items, the Leadership Practices Inventory was even easier, on

average. Mean discrimination was less with Leadership Practices Inventory, as was mean difficulty. The parameter estimate differences were relatively small. As noted, the classical test theory Cronbach's alpha was less than the MLQ's 20-items subscale and IRT reliability was less, achieving a high of .91 for the Leadership Practices Inventory versus .94 for the MLQ's 20 item subscale. Zagorsek et al. (2006) provided enough IRT information on the Leadership Practices Inventory to determine that the two assessments had similar limitations of item difficulty in the upper ability levels. Reliability was a strong differentiator of the two assessments, favoring the MLQ's transformational leadership subscale.

Implications for Social Change

A reliable instrument and clear criteria are needed to detect harmful transformational leaders. The consequences of misidentification require a high degree of reliability in differentiating the beneficial transformational leader from the potentially harmful one. The MLQ's 20 item transformational subscale will require additional research before being fully relied upon for narcissistic transformational leader detection. Figure 9 shows an IRT example of an unidentified leader from the combined sample ($N = 1,703$) who might have narcissistic transformational leader tendencies as described in Chapter 1. The example is used to illustrate proposed selection criteria that can advance social change through detection and intervention of narcissistic transformational leaders for training and development or supplementing hiring and promotion decisions. Narcissism was correlated to high idealized influence attributed facet scores and low individual consideration facet scores (Khoo & Burch, 2008). From a classical test theory

perspective the anonymous leader depicted in Figure 9, had score a total of 52, the mode. This score was unremarkable. However, the idealized influence attributed facet score was 15 and the individual consideration facet score was 5. The difference was remarkable and may suggest a narcissistic tendency according to research by Khoo and Burch (2006). Considering that the average facet score was about 10, the unidentified leader might have warranted further testing. The Hogan Development Survey was designed to assess dysfunctional behavior such as narcissism in the workplace, though not transformational leadership (Khoo & Burch, 2006). Khoo and Burch used a combination of assessments to achieve their findings.

From an IRT perspective, total score or facet score have inadequate precision. IRT item and person parameters can be used to detect inter-item correlations across theta, suggesting potentially harmful transformational leaders. In Figure 9, the individual's subjective transformational theta was above average ($\tau = 0.79$, $SE = 0.24$), shown as a horizontal dashed line. Individual responses to each of the four idealized influence attributed items and four individual consideration items were 4,4,4,3,2,2,1, and 0, respectively. The selection criteria suggested by this example are that all idealized influence attributed items are marked as *fairly often* (3) or above and individual consideration items are *sometimes* (2) or below. Support for these criteria was found in Khoo and Burch (2006) correlation study augmented by the IRT analysis for the 8 invariant items. These eight items are marked on the graph as beta values. For instance, idealized influence attributed item 4 marked *IIA4* was rated *fairly often* (3). The greatest probability of a three response lies between category boundaries δ_3 and δ_4 . The $\delta_3 = -0.57$

($SE = .05$) value would be the initial beta and $\delta_4 = 0.76$ ($SE = 0.06$) would be the final beta in Figure 9.

For those items with only a final value, the responses were the extreme of either 0 or 4. In the case of a four response, the boundary from category three to four marks the final beta value. Above the highest category is a theta of positive infinity; a monotonically increasing category function. This is depicted in Figure 9 as an upward arrow pointing beyond the graph. In the same manner, the only response below 0 is negative infinity, depicted with a downward arrow pointing beyond the graph. The range of responses for all 8 items had a reliability of at least .99, equivalent to a 99% confidence interval.

The four individual consideration items are at least 0.5 to 2.5 standard deviations below the leader's subjective theta ability measure. Khoo and Burch (2006) measured at the facet level and provided only correlations with the Hogan Development Survey (Hogan & Hogan, 1997) facets. Therefore, it is not clear what range of theta or theta differences between items was significant in detecting narcissism. Further IRT research, using the MLQ together with a test for dysfunctional behaviors such as the Personality Disorder Scales of the Minnesota Multiphasic Personality Inventory (Morey, Waugh, & Blashfield, 1985) or Hogan Development Survey (Hogan & Hogan, 1997) is required to validate the suggested selection criteria.

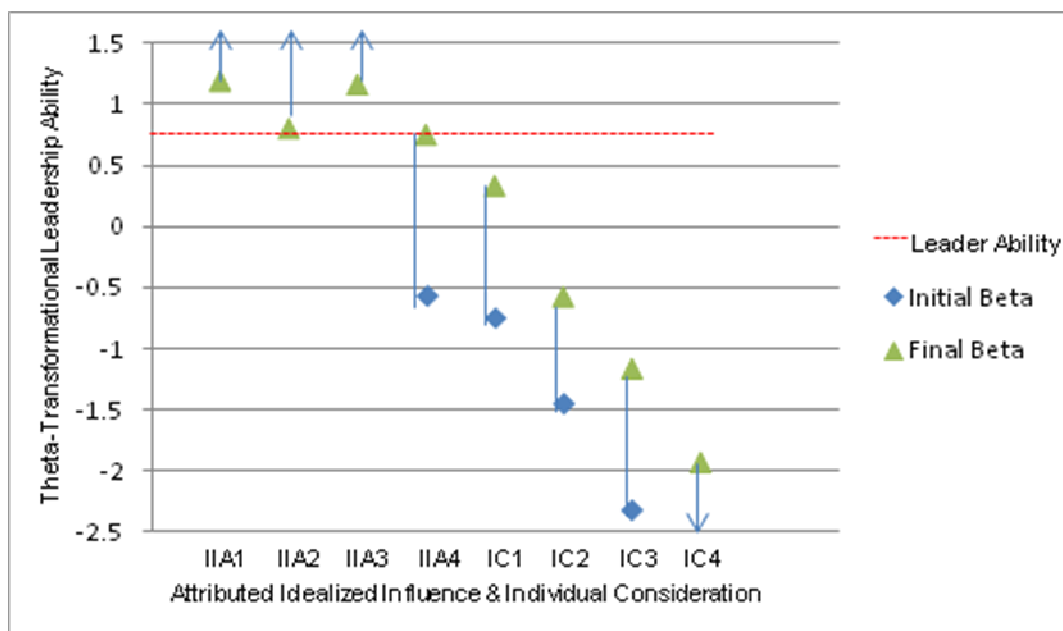


Figure 9. Subordinate's response to a possible narcissistic transformational leader.

As described in Chapter 1, harmful transformational leaders have been responsible for billions in economic costs (Edid, 2004), significant job displacement (Post, 2008), and large scale psychological suffering (Hetland et al., 2007). Khoo and Burch (2006) demonstrated significant correlations between narcissism and transformational leadership. A limitation of the experiment was the exclusive use of classical test theory for analysis. Imprecise facet level results were reported (Khoo & Burch, 2006). This study has advanced the precision of item parameters of the transformational subscale where future research can determine the precise responses that indicate a potentially harmful transformational leader. By detecting these narcissistic leaders, it is hoped that early intervention can reduce the costs, job losses, and suffering of hundreds of thousands of subordinates.

Recommendations for Action

Action Recommendation 1: Disseminate Results

Disseminate findings from this study to human resource professionals, assessment researchers, and leadership researchers. Professional conferences, leadership assessment centers, and professional psychology based internet sites may be able to provide effective distribution networks for information from this study. Multiple dissemination methods can support an integrative response to this study's findings and recommendations.

Avolio, Walumbwa et al. (2009) called for a holistic approach to studying leadership within the organizational context. Emphasis was placed on a multidiscipline approach.

Once disseminated, this study can be useful to human resource professionals for detailed discussions with leaders about assessment scores, by assessment researchers on IRT analysis, and by leadership researchers for additional study recommendations. However, an integrative approach with all three groups can lead to improved professional application of detection and intervention of harmful transformational leaders. Avolio, Mhatre et al. (2009) suggest that an integrative approach appeared to increase effect sizes for leadership intervention.

Action Recommendation 2: Integrate IRT in the MLQ's Research

Classical test theory alone is not sufficient for research on the MLQ. As this study demonstrated, a combination of classical test theory and IRT is essential for studying the MLQ, especially for transformational leadership. Psychologists researching the MLQ assessment should attend to the findings in this study as an example of additional precision in reporting the results of their own studies. IRT can supplement classical test

theory in assessment research (Zagorsek et al., 2006). Item parameter estimates are invariable and can be directly used or compared across research settings, providing accelerated knowledge accumulation (De Ayala, 2009). Sharing IRT information will be especially useful if item modification efforts are pursued for the MLQ since single item changes do not require retesting of the entire assessment (Embretson & Reise, 2000).

Action Recommendation 3: Integrate IRT in Psychology Classes

Introduce IRT analysis along with classical test theory in psychology education.

The Literature Review section of Chapter 2 discusses IRT as a newer theory or set of models (De Ayala, 2009). Baker (2001) suggested that IRT's greater initial complexity might restrict coverage in statistical courses. As this study shows, psychological research is enhanced with the use of IRT in studying assessments and latent traits. Psychology education can be an important context to expose new researchers to models and applications of IRT. The retention of categorical information, along with the practical examples used in this study, provides awareness of ways IRT can be used in combination with classical test theory to improve social conditions. Students and teachers should request that IRT be taught in statistical classes, especially in psychology education.

Action Recommendation 4: Improve Transformational Leader Assessment

Resist reporting single composite scores as adequate representations of transformational leadership abilities. Given the example of cutoff scores and IRT ability estimation shown in Figure 8 of Chapter 4, human resource practitioners should be concerned with reporting only composite scores. Using IRT in transformational leadership ability estimation is supported by this study; however, reporting classical test

theory composite scores for decision criteria was not. Accuracy in decision criteria can be legally defended (Kleiman & Faley, 1978). Human resource practitioners should request IRT ability estimates as part of data processing from the MLQ copyright holder, Mind Garden Incorporated, or perform their own IRT analysis on the responses.

Action Recommendation 5: Improve Organizational Responsibility

Use multiple assessments to detect harmful transformational leaders until the MLQ is more fully researched as a single source. As documented in Chapter 1, harmful transformational leaders have damaged hundreds of thousands of lives. Organizations have a responsibility to detect and intervene for the protection of vulnerable populations. Until detection of harmful transformational leaders using the MLQ becomes practical, multiple assessments and other qualitative tools should be relied upon to reduce the many negative impacts.

Human resource managers may receive complaints or observe harmful leaders throughout the organization. At every level, highly valid and reliable assessments are required to provide information for impactful personnel decisions (Terpstra et al., 1999). The MLQ has had a positive association with optimum leadership styles (Kanste et al., 2007). Given the social desirability of being a transformational leader, the MLQ was voluntarily completed with response rates above 50% (Cole et al., 2006). The dysfunctional leaders, often at the top of the organizational hierarchy, can be difficult to successfully dislodge without a credible assessments, voluntarily taken (Chatterjee & Hambrick, 2007). The MLQ can be a part of a larger set of inputs to detect and intervene with harmful leadership behaviors.

Detecting potentially harmful transformational leaders at an early stage of development may improve intervention strategies and facilitate leadership hiring, promotion, or separation. Support for corrective behaviors in those leaders who have a tendency towards narcissism or other dysfunctional behaviors enhanced subordinate welfare (Avolio, Mhatre et al., 2009). The organization could further benefit through greater productivity (Avolio & Bass, 2004a), shareholders could increase the stability of their holdings (Hayward & Hambrick, 1997), and the costs to society due to job dislocation could also be reduced (Edid, 2004). Edid (2004) calculated the benefits of detecting potentially harmful transformational leaders and intervening on behalf of the subordinates, organizations, and shareholders were in the billions of dollars. Human Resource managers are encouraged to seek multiple methods of detecting and then intervening with harmful transformational leaders until the MLQ can be shown to provide a reliable single source of detection.

Recommendations for Further Study

Future Study Recommendation 1: Extend Khoo and Burch (2007) Study

For practical detection of harmful transformational leaders, IRT item and person level replication of Khoo and Burch (2007) study must be conducted using this study's invariant item parameters to test proposed selection criteria. Khoo and Burch originated the study of harmful transformational leaders using the MLQ and the Hogan Development Survey. However, the classical test theory approach did not provide enough reliability at the item or person level for adequate detection and intervention. The recommendation is to employ classical test theory in combination with IRT to study item

and person level correlations of harmful transformational behaviors using the MLQ and the Hogan Development Survey. IRT can be used to determine which patterns of responses predict harmful behaviors. Perhaps significant patterns include other facet items in addition to idealized influence attributed items and individual consideration items. Even in facets known to have correlations to harmful behaviors, item patterns need to be studied to improve predictability. Future research to test suggested item level cutoff scores for beneficial transformational leadership behaviors can be an essential tool for organizational governance and human resource management.

Future Study Recommendation 2: Modify Five MLQ Items

Five items in the MLQ's transformational leadership subscale are candidates for modification. From results of this study, there are known gaps in reliable detection higher than 1.2 of the latent trait range. Modification of redundant items related to the description of Figure 5 in Chapter 4, can lead to increase differentiation of upper level transformational leaders (Zagorsek et al., 2006). Specifically, if the five facet items, mlq2, mlq6, mlq9, mlq15, and mlq18 were modified to increase difficulty, detection reliability would increase at the top end of the continuum. The unmodified items could be used as IRT anchors for comparisons (Thissen et al., 1983). New item testing using IRT would be more efficient due to the invariant nature of item parameters over classical test theory methods (De Ayala, 2009). Items with higher difficulty and discrimination would increase the information and reliability across the theta range (Embretson & Reise, 2000). To increase item reliability, rephrasing may be used to increase discrimination while retaining the essence of the behavioral facet (Thissen et al., 1983). Future research is

needed to improve the range over which transformational leadership can reliably be tested above a theta of 1.2.

Future Study Recommendation 3: Extend This Study

This study should be extended to include other MLQ items, multidimensional software, and untested populations. This study is an exploratory IRT analysis of the MLQ's 20 item transformational leadership subscale and should be verified and expanded. Other MLQ items in the assessment require further research using software based on multidimensional models (De Ayala, 2009). Chapter 2 discusses that the transactional and laissez-faire items have not been analyzed using IRT methods. Multidimensional models are required to study the MLQ transactional items, given the complex nature described in Chapter 2. Multidimensional models are not readily available in software form; however, are an active area of development (De Ayala, 2009).

Analysis of different population samples would yield additional IRT ability parameter estimates. For instance, Avolio and Bass (2004a) described scoring differences due to culture. Non-Israeli samples could be used to determine the precise relationship between culture and transformational leadership ability. Athletic, corporate, military, and academic settings may produce unique person parameters, as inclusion of an athletic sample demonstrated in this study. The estimated differences may also be due to differential item functioning for groups within populations of interest, such as gender (Güler & Penfield, 2009). Further research is required to understand the MLQ 20 item parameter differences across cultures, settings, and groups within a sample.

Future Study Recommendation 4: Connect This Study to Derailment

Derailment and harmful transformational leadership may share some common characteristics and should be investigated. There is a body of psychological knowledge about manager derailment (Hogan & Hogan, 2001). Khoo and Burch (2006) made an initial connection between derailment behaviors and transformational leadership using the Hogan Development Survey and the MLQ. Further research is needed to verify and extend this exploratory work at the item level. Derailment may differ from harmful transformational leadership in specific ways; however, the relationship is unclear. Further research is needed to investigate possible connections between derailment and harmful transformational leadership.

Future Study Recommendation 5: Improve Leaders' Response Rates

Study the separation of the transformational leadership items of the MLQ as a distinct test. A final area for further research is increasing senior executive response rates using only the 20 transformational leadership items from this study. IRT item parameters are invariant and therefore can form part of a new test for the same construct (Embretson & Reise, 2000). This parsimonious new test should take about 7 to 10 minutes to complete. Getting response rates for top executives up around 75% to 80% would be desirable for screening harmful transformational leaders and increasing sample size. Given executive schedules, this may be accomplished online (Cole et al., 2006). Future research can address whether high response rates are practical with a 20 item online transformational leadership test.

These five recommendations for future study can advance the IRT body of knowledge of harmful transformational leadership. Harmful leaders abound in many organizations (Edid, 2004). Improving the detection and intervention at an early stage can improve job satisfaction (Avolio & Bass, 2004a), decrease cynicism and exhaustion (Hetland et al., 2007), preserve jobs (Edid, 2004), and decrease sexual abuse (Ronan, 2008). Without further research, the MLQ 20 item subscale cannot be successfully used for detection of harmful transformational leaders.

Conclusions

The social cost from transformational leaders that exhibit narcissistic behaviors has been extensive. Severe physical, psychological, and financial damage has been inflicted on vulnerable individuals, institutions, and societies by harmful transformational leaders (Edid, 2004; Post 2008; Ronan, 2008). Separating harmful from beneficial transformational leaders is not straightforward; narcissism and transformational abilities share common characteristics (Judge, Piccolo, & Kosalka, 2009). Khoo and Burch (2007) used the MLQ and the Hogan Development Survey to determine significant correlations of harmful transformational leaders at a facet level. A gap in literature existed at the item level to lay the foundation for detection of harmful transformational leaders (Hetland et al., 2007). This study's IRT results showed item level analysis provided increased precision for detection over classical test theory item analysis.

This is the first study to apply IRT to the MLQ. IRT item level analysis had not been performed for the MLQ during its 25-year history as is shown in chapter 2. Classical test theory in conjunction with IRT analysis of the 20 transformational leadership items

was used to achieve research objectives. Three Israeli corporate and athletic subordinate samples were combined ($N = 1,703$) for this study. A number of critical assumptions were tested including independence of observations, sample homogeneity, and IRT unidimensionality. Five research questions focused on using the IRT model with the greatest reliability to estimate invariant item parameters. Results showed that the GRM model provided the best item difficulty and discrimination estimates based on lower χ^2/df values, and item parameters can be used by other researchers independent of differences in respondents, sample sizes, administration settings, or with other transformational items (Embretson & Reise, 2000). Further, coaches were perceived as having less transformational ability ($M = -0.37$, $SD = 0.85$) than corporate leaders. Ability detection using the invariant item parameter estimates from this study became possible with sufficiently high reliabilities. Nunnally and Bernstein (1994) suggested reliability should be at least .95 for supportable detection and intervention. Results showed GRM item parameters had a reliability of .94 from -2.3 to 1.2 and abilities had a reliability of .96 from -1.7 to 1.3 of perceived transformational leadership. The selection criteria suggested by 8 invariant item parameters was that all idealized influence attributed items are marked as *fairly often* (3) or above and individual consideration items are *sometimes* (2) or below. These narcissistic leadership selection criteria were demonstrated using an individual example.

Research is often the building of knowledge toward a positive social change (Avolio, Walumbwa et al., 2009). The identification of transformational leaders with narcissistic tendencies was advanced by increasing the reliability of item parameters used

in detection. The reliability of ability selection and proposed criteria may encourage other researchers to further improve detection and intervention. Individual subordinate workers, corporate organizations, religious institutions, and entire segments of societies are irrevocably damaged by individual and group killings, sexual assaults, or other brutal victimization by narcissistic transformational leaders (Edid, 2004; Post 2008; Ronan, 2008). Therefore, the motivation to continue research in detection is related to the hundreds of thousands of vulnerable and distressed individuals, organizations, and societies subjected to narcissistic transformational leaders.

References

- Acton, G., Kunz, J., Wilson, M., & Hall, S. (2005). The construct of internalization: Conceptualization, measurement, and prediction of smoking treatment outcome. *Psychological Medicine, 35*(3), 395-408. doi:10.1017/S0033291704003083.
- Alban-Metcalfe, R., & Alimo-Metcalfe, B. (2000). An analysis of the convergent and discriminant validity of the Transformational Leadership Questionnaire. *International Journal of Selection & Assessment, 8*(3), 158-175. doi:10.1111/1468-2389.00144.
- Allen, T., Barnard, S., Rush, M., & Russell, J. (2000). Ratings of organizational citizenship behavior: Does the source make a difference? *Human Resource Management Review, 10*(1), 97-114. doi:10.1016/S1053-4822(99)00041-8.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (DSM-IV-TR-4th edition, text revision). Washington, DC: Author. Washington, DC: American Psychiatric Association.
- Antonakis, J., Avolio, B., & Sivasubramaniam, N. (2003). Context and leadership: An examination of the nine-factor full-range leadership theory using the multifactor leadership questionnaire. *The Leadership Quarterly, 14*(3), 261-295. doi:10.1016/S1048-9843(03)00030-4.
- Asia: Saving the children. (2000, October). *The Economist, 357*(8191), 51. doi: 62426705.

- Atkins, P., & Wood, R. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55*(4), 871-904. doi:10.1111/j.1744-6570.2002.tb00133.x.
- Atwater, L., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self- and follower ratings of leadership. *Personnel Psychology, 48*(1), 35-59. doi:10.1111/j.1744-6570.1995.tb01745.x.
- Atwater, L., Waldman, D., Ostroff, C., Robie, C., & Johnson, K. (2005). Self-Other Agreement: Comparing its Relationship with Performance in the U.S. and Europe. *International Journal of Selection and Assessment, 13*(1), 25-40. doi:10.1111/j.0965-075X.2005.00297.x.
- Avolio, B. (2008). Bernard (Bernie) M. Bass (1925—2007). *American Psychologist, 62*(0). Retrieved from <http://search.ebscohost.com.ezp.waldenulibrary.org>
- Avolio, B., & Bass, B. (1995). Individual consideration viewed at multiple levels of analysis: A multi-level framework for examining the diffusion of transformational leadership. *Leadership Quarterly, 6*(2), 199-218. doi:10.1016/1048-9843(95)90035-7.
- Avolio, B. J., & Bass, B. M. (2004a). *Multifactor leadership questionnaire manual and sample set (3rd ed.)*. Menlo Park, CA: Mind Garden. Retrieved from <http://www.mindgarden.com/index.htm>
- Avolio, B. J., & Bass, B. M. (2004b). *MLQ international normative samples*. Menlo Park, CA: Mind Garden. Retrieved from <http://www.mindgarden.com/index.htm>

- Avolio, B., Bass, B., & Jung, D. (1999). Re-examining the components of transformational and transactional leadership using the Multifactor Leadership Questionnaire. *Journal of Occupational and Organizational Psychology*, 72(4), 441-462. doi:10.1348/096317999166789.
- Avolio, B., Mhatre, K., Norman, S., & Lester, P. (2009). The moderating effect of gender on leadership intervention impact: An exploratory review. *Journal of Leadership & Organizational Studies*, 15(4), 325-341. doi:10.1177/1548051809333194.
- Avolio, B., Walumbwa, F., & Weber, T. (2009). Leadership: Current theories, research, and future directions. *Annual Review of Psychology*, 60(1), 421-449. doi:10.1146/annurev.psych.60.110707.163621.
- Baker, F., & ERIC Clearinghouse on Assessment and Evaluation, C. (2001, January 1). *The Basics of Item Response Theory. Second Edition.* (ERIC Document Reproduction Service No. ED458219). Retrieved from <http://search.ebscohost.com.ezp.waldenulibrary.org>
- Barling, J., Weber, T., & Kelloway, E. (1996). Effects of transformational leadership training on attitudinal and financial outcomes: A field experiment. *Journal of Applied Psychology*, 81(6), 827-832. doi:10.1037/0021-9010.81.6.827.
- Barr, M., & Raju, N. (2003). IRT-based assessments of rater effects in multi-source feedback instruments. *Organizational Research Methods*, 6(1), 15-43. doi:10.1177/1094428102239424.
- Bass, B. (1985). *Leadership and performance beyond expectations*. New York, NY: Free Press.

- Bass, B. (1990). *Bass & Stogdill's handbook of leadership: Theory, research, and managerial applications (3rd ed.)*. New York, NY: Free Press.
- Bass, B. (1997). Does the transactional-transformational leadership paradigm transcend organizational and national boundaries? *American Psychologist*, *52*(2), 130-139. doi:10.1037/0003-066X.52.2.130.
- Bass, B., Jung, D., Avolio, B., & Berson, Y. (2003). Predicting unit performance by assessing transformational and transactional leadership. *Journal of Applied Psychology*, *88*(2), 207-218. doi:10.1037/0021-9010.88.2.207.
- Berson, Y. (1999). A comprehensive assessment of leadership using triangulation of qualitative and quantitative methods. Ph.D. dissertation, State University of New York at Binghamton, New York. Retrieved from ProQuest database.
- Berson, Y. & Avolio, B. (2004). Transformational Leadership and the Dissemination of Organizational Goals: A Case Study of a Telecommunication Firm. *Leadership Quarterly*, *15*(5), 625-646. doi:10.1016/j.leaqua.2004.07.003.
- Berson, Y., & Linton, J. (2005). An examination of the relationships between leadership style, quality, and employee satisfaction in R&D versus administrative environments. *R&D Management*, *35*(1), 51-60. doi:10.1111/j.1467-9310.2005.00371.x.
- Berson, Y., Oreg, S., & Dvir, T.. (2008). CEO values, organizational culture and firm outcomes. *Journal of Organizational Behavior*, *29*(5), 615-634. doi:10.1002/job.499.

- Berson, Y., Shamir, B., Avolio, B., & Popper, M. (2001). The relationship between vision strength, leadership style, and context. *The Leadership Quarterly*, *12*, 53-73. doi:10.1016/S1048-9843(01)00064-9.
- Berson, Y., & Sosik, J. (2007). The relationship between self-other rating agreement and influence tactics and organizational processes. *Group & Organization Management*, *32*(6), 675-698. doi:10.1177/1059601106288068.
- Bertrand, M., & Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *Quarterly Journal of Economics*, *118*(4), 1169-1208. doi:10.1162/003355303322552775.
- Bjorner, J., Chang, C., Thissen, D., & Reeve, B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research*, *16*, 95-108. doi:10.1007/s11136-007-9168-6.
- Blankenship, V., Vega, C., Ramos, E., Romero, K., Warren, K., Keenan, . . . Sullivan, A. (2006). Using the multifaceted Rasch model to improve the TAT/PSE measure of need for achievement. *Journal of Personality Assessment*, *86*(1), 100-114. doi:10.1207/s15327752jpa8601_11.
- Bloom, N., & Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries. *Quarterly Journal of Economics*, *122*(4), 1351-1408. doi:10.1162/qjec.2007.122.4.1351.
- Bono, J., & Judge, T. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal of Applied Psychology*, *89*(5), 901-910. doi:10.1037/0021-9010.89.5.901.

- Bycio, P., Hackett, R., & Allen, J. (1995). Further assessments of Bass's (1985) conceptualization of transactional and transformational leadership. *Journal of Applied Psychology, 80*(4), 468-478. doi:10.1037/0021-9010.80.4.468.
- Bureau of Economic Analysis. (2010, June 17). International Economic Accounts. Retrieved July 10, 2010 from <http://www.economicindicators.gov>
- Burns, J. M. (1978). *Leadership*. New York, NY: Harper & Row.
- Campbell, D. (2002). The history and development of the Campbell Interest and Skill Survey. *Journal of Career Assessment, 10*(2), 150-168. doi:10.1177/1069072702010002002.
- Campbell, D. P., Hyne, S. A., & Nilsen, D. L. (1992). *Manual for the Campbell Interest and Skill Survey*. Minneapolis, MN: National Computer Systems.
- Careless, S. (1998). Assessing the discriminant validity of transformational leader behavior as measured by MLQ. *Journal of Occupational and Organizational Psychology, 71*, 353-358. Retrieved from <http://www.bpsjournals.co.uk/journals/joop>
- Carless, S. (2001). Assessing the discriminant validity of the Leadership Practices Inventory. *Journal of Occupational and Organizational Psychology, 74*(2), 233-239. doi:10.1348/096317901167334.
- Chan, K., & Drasgow, F. (2001). Toward a theory of individual differences and leadership: Understanding the motivation to lead. *Journal of Applied Psychology, 86*(3), 481-498. doi:10.1037/0021-9010.86.3.481.

- Chatterjee, A., & Hambrick, D. (2007). It's all about me: Narcissistic chief executive officers and their effects on company strategy and performance. *Administrative Science Quarterly*, 52(3), 351-386. Retrieved from <http://www.johnson.cornell.edu/publications/asq>
- Chernyshenko, O., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36(4), 523-562.
doi:10.1207/S15327906MBR3604_03.
- Clinton, J., & Lapinski, J. (2006). Measuring Legislative Accomplishment, 1877–1994. *American Journal of Political Science*, 50(1), 232-249. doi:10.1111/j.1540-5907.2006.00181.x.
- Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.
- Cohen, R., & Swerdlik, M. (2005). *Psychological testing and assessment: an introduction to test and measurements (6th ed.)*. New York: NY: McGraw-Hill Publishing.
- Cole, M., Bedeian, A., & Field, H. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods*, 9(3), 339-368.
doi:10.1177/1094428106287434.

- Cook, J. (2005, January 14). Save tsunami orphans from the dangers of prostitution: [USA 2ND EDITION]. *Financial Times*, p. 12. doi: 778755501.
- Costa, P. T., & McCrae, R. R. (1992). *Professional manual for the NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.
- Cox, S., & Sergejew, A. (2003). The development of a unidimensional continuum measure of positive and negative affective experience. *Australian Journal of Psychology*, 55, 76-77. Retrieved from <http://www.informaworld.com/smpp/title~content=t713701010>
- Craig, S., & Gustafson, S. (1998). Perceived leader integrity scale: An instrument for assessing employee perceptions of leader... *Leadership Quarterly*, 9(2), 127-144. doi:10.1016/S1048-9843(98)90001-7.
- Craig, S., & Kaiser, R. (2003). Applying item response theory to multisource performance ratings: What are the consequences of violating the independent observations assumption? *Organizational Research Methods*, 6(1), 44-60. doi:10.1177/1094428102239425.
- Cudeck, R., & Browne, M. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, 18(2), 147-167. doi:10.1207/s15327906mbr1802_2.

- Culp, G., & Smith, A. (2005). Leadership Effectiveness and Behavior. *Leadership & Management in Engineering*, 5(2), 39-48. doi:10.1061/(ASCE)1532-6748(2005)5:2(39).
- Davies, S. & Wadlington, P.L. (2006). Factor and parameter invariance of a Five Factor personality test across proctored/unproctored computerized administration. Society for Industrial and Organizational Psychologists 21st Annual Convention, May 5th, 2006. Dallas, TX.
- Davies, S., & Wadlington, P. (2007). *Interactions in Test Administration Settings: The Effect of Applicant Personality*. Society for Industrial and Organizational Psychologists 22nd Annual Convention May, 2007. New York, NY.
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.
- Den Hartog, D. N., Van Muijen, J. J., & Koopman, P. L. (1997). Transactional versus transformational leadership: An analysis of MLQ. *Journal of Occupational and Organizational Psychology*, 70, 19-34. Retrieved from <http://www.bpsjournals.co.uk/journals/joop>
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, 9(3), 327-346. doi:10.1207/S15328007SEM0903_2.

- DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An Empirical review. *Journal of Psychoeducational Assessment, 23*, 225-241. doi:10.1177/073428290502300303.
- Donovan, M., Drasgow, F., & Probst, T. (2000). Does computerizing paper-and-pencil job attitude scales make difference? New IRT analyses offer insight. *Journal of Applied Psychology, 85*(2), 305-313. doi:10.1037/0021-9010.85.2.305.
- Drasgow, F., Levine, M., Tsien, S., Williams, B., & Mead, A. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*(2), 143-165. doi:10.1177/014662169501900203.
- Drasgow, F., & Lissak, R. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363-373. doi:10.1037/0021-9010.68.3.363.
- Eagly, A., Johannesen-Schmidt, M., & van Engen, M. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin, 129*(4), 569. Retrieved from <http://search.ebscohost.com.ezp.waldenulibrary.org>
- Edid, M (2004). Ethical leadership and the price of bad behavior. Cornell University ILR School. Retrieved July 10, 2010 from <http://digitalcommons.ilr.cornell.edu/briefs/3>
- Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

- Edelen, M., & Reeve, B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *16*, 5-18. doi:10.1007/s11136-007-9198-0.
- Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. London, England: Hodder & Stoughton.
- Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, *6*(1), 56-83. doi:10.1080/10705519909540119.
- Felfe, J., & Schyns, B. (2002). The relationship between employees' occupational self-efficacy and perceived transformational leadership: Replication and extension of recent results. *Current Research in Social Psychology*, *7*(9), 137-162. Retrieved from <http://www.uiowa.edu/~grpproc/crisp/crisp.html>
- Fleishman, E. A. (1953). The description of supervisory behavior. *Personnel Psychology*, *37*, 1-6. doi:10.1037/h0056314.
- Flora, D., & Curran, P. (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. *Psychological Methods*, *9*(4), 466-491. doi:10.1037/1082-989X.9.4.466.
- Gentry, W., Hannum, K., Ekelund, B., & de Jong, A. (2007). A study of the discrepancy between self- and observer-ratings on managerial derailment characteristics of European managers. *European Journal of Work and Organizational Psychology*, *16*(3), 295-325. doi:10.1080/13594320701394188.

- Goleman, D. (1995). *Emotional intelligence*. New York, NY: Bantam.
- Gough, H. G. (1987). *California Psychological Inventory administrator's guide*. Palo Alto, CA: Consulting Psychological Press.
- Grimm, L. G., & Yarnold, P. R. (2000). *Reading and understanding more multivariate statistics*. Washington, D.C.: American Psychological Association.
- Guilford, J. S., Zimmerman, W. S., & Guilford, J. P. (1976). *The Guilford-Zimmerman Temperament Survey handbook: Twenty-five years of research and application*. San Diego, CA: EDITS.
- Güler, N., & Penfield, R. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, 46(3), 314-329. doi:10.1111/j.1745-3984.2009.00083.x.
- Han, K. T. (2010). WINGEN: Version 3.01.414 for Windows [Computer program]. University of Massachusetts Amherst. Retrieved from <http://www.umass.edu/remf/software/simcata/wingen/manualF.html>.
- Hater, J., & Bass, B. (1988). Superiors' evaluations and subordinates' perceptions of transformational and transactional leadership. *Journal of Applied Psychology*, 73(4), 695-702. doi:10.1037/0021-9010.73.4.695.
- Hayward, M., & Hambrick, D. (1997). Explaining the premiums paid for large acquisitions: Evidence of CEO hubris. *Administrative Science Quarterly*, 42(1), 103-127. doi:10.2307/2393810.

- Heinitz, K., Liepmann, D., & Felfe, J. (2005). Examining the Factor Structure of MLQ: Recommendation for a Reduced Set of Factors. *European Journal of Psychological Assessment, 21*(3), 182-190. doi:10.1027/1015-5759.21.3.182.
- Hetland, H., & Sandal, G. (2003). Transformational leadership in Norway: Outcomes and personality correlates. *European Journal of Work & Organizational Psychology, 12*(2), 147. doi:10.1080/13594320344000057.
- Hetland, H., Sandal, G., & Johnsen, T. (2007). Burnout in the information technology sector: Does leadership matter? *European Journal of Work and Organizational Psychology, 16*(1), 58-75. doi:10.1080/13594320601084558.
- Hinkin, T., & Schriesheim, C. (2008a). An examination of 'nonleadership': From laissez-faire leadership to leader reward omission and punishment omission. *Journal of Applied Psychology, 93*(6), 1234-1248. doi:10.1037/a0012875.
- Hinkin, T., & Schriesheim, C. (2008b). A theoretical and empirical examination of the transactional and non-leadership dimensions of the multifactor leadership questionnaire (MLQ). *Leadership Quarterly, 19*(5), 501-513. doi:10.1016/j.leaqua.2008.07.001.
- Hogan, R., & Hogan, J. (1995). *Hogan Personality Survey Manual (2nd ed.)*. Tulsa, OK: Hogan Assessment Systems.
- Hogan, R., & Hogan, J. (1997). *Hogan Development Survey Manual*. Tulsa, OK: Hogan Assessment Systems.

- Hogan, R., & Hogan, J. (2001). Assessing leadership: A view from the dark side. *International Journal of Selection and Assessment*, 9(1-2), 40-51. doi:10.1111/1468-2389.00162.
- Hogan, R., & Kaiser, R. (2005). What We Know About Leadership. *Review of General Psychology*, 9(2), 169-180. doi:10.1037/1089-2680.9.2.169.
- Hogan, R., & Tett, R. (2003). Leadership personality. In *Encyclopedia of Psychological Assessment* (Vols. 1-2), pp. 548-553). London, England: Sage.
- Howell, J., Neufeld, D., & Avolio, B. (2005). Examining the relationship of leadership and physical distance with business unit performance. *Leadership Quarterly*, 16(2), 273-285. doi:10.1016/j.leaqua.2005.01.004.
- Hu, L., & Bentler, P. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424-453. doi:10.1037/1082-989X.3.4.424.
- Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55. doi:10.1080/10705519909540118.
- Hunt, J. (1999). Transformational/charismatic leadership's transformation of the field: An historical essay. *The Leadership Quarterly*, 10(2), 129-144. doi:10.1016/S1048-9843(99)00015-6.
- Javidan, M., & Dastmalchian, A. (2009). Managerial implications of the GLOBE project: A study of 62 societies. *Asia Pacific Journal of Human Resources*, 47(1), 41 - 58. doi:10.1177/1038411108099289.

- Judge, T., & Piccolo, R. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology, 89*(5), 755-768. doi:10.1037/0021-9010.89.5.755.
- Judge, T., Piccolo, R., & Kosalka, T. (2009). The bright and dark sides of leader traits: A review and theoretical extension of the leader trait paradigm. *The Leadership Quarterly, 20*(6), 855-875. doi:10.1016/j.leaqua.2009.09.004.
- Kaiser, R., Hogan, R., & Craig, S. (2008). Leadership and the fate of organizations. *American Psychologist, 63*(2), 96-110. doi:10.1037/10003-066X.63.2.96.
- Kanste, O., Miettunen, J., & Kyngäs, H. (2007). Psychometric properties of the multifactor leadership questionnaire among nurses. *Journal of Advanced Nursing, 57*(2), 201-212. doi:10.1111/j.1365-2648.2006.04100.x.
- Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146-162. doi:10.1177/01466210122031975.
- Kleiman, L., & Faley, R. (1978). Assessing content validity: standards set by the court. *Personnel Psychology, 31*(4), 701-713. doi:10.1111/j.1744-6570.1978.tb02119.x.
- Khoo, H., & St. J. Burch, G. (2008). The 'dark side' of leadership personality and transformational leadership: An exploratory study. *Personality and Individual Differences, 44*(1), 86-97. doi:10.1016/j.paid.2007.07.018.
- Koenig, J., & Roberts, J. (2007). Linking parameters estimated with the generalized graded unfolding model: A comparison of the accuracy of characteristic curve

methods. *Applied Psychological Measurement*, 31(6), 504-524.

doi:10.1177/0146621606297315.

Kouzes, J., & Posner, B. (1988). *The Leadership Practices Inventory*. San Diego, CA: Pfeiffer.

Langan-Fox, J., & Grant, S. (2006). The Thematic Apperception Test: Toward a standard measure of the big three motives. *Journal of Personality Assessment*, 87(3), 277-291. doi:10.1207/s15327752jpa8703_09.

Lazarsfeld, P., & Robinson, W. (1940). The quantification of case studies. *Journal of Applied Psychology*, 24(6), 817-825. doi:10.1037/h0058384.

Lim, B., & Ployhart, R. (2004). Transformational Leadership: Relations to the Five-Factor Model and Team Performance in Typical and Maximum Contexts. *Journal of Applied Psychology*, 89(4), 610-621. doi:10.1037/0021-9010.89.4.610.

Ling, Y., Simsek, Z., Lubatkin, M., & Veiga, J. (2008). The impact of transformational CEOs on the performance of small- to medium-sized firms: Does organizational context matter? *Journal of Applied Psychology*, 93(4), 923-934. doi:10.1037/0021-9010.93.4.923.

Lissak, R. & Wytmar, R. (1981). CARIF: a computational program for *item* information functions and *item* characteristic curves. *Behavior Research Methods and Instrumentation*, 13(3), 360. Retrieved from <http://brm.psychonomic-journals.org>

- Lord, F. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989–1020. doi:10.1177/001316446802800401.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lowe, K. (2000). Ten years of the leadership quarterly: Contributions and challenges for the future. *Leadership Quarterly, 11*(4), 459. doi:10.1016/S1048-9843(00)00059-x.
- Lowe, K., Kroeck, K., & Sivasubramaniam, N. (1996). Effectiveness correlates of transformation and transactional leadership: A meta-analytic review of MLQ literature. *Leadership Quarterly, 7*(3), 385-425. doi:10.1016/S1048-9843(96)90027-2.
- Mayers, A., Khoo, S., & Svartberg, M. (2002). The Existential Loneliness Questionnaire: Background, development, and preliminary findings. *Journal of Clinical Psychology, 58*(9), 1183-1193. doi:10.1002/jclp.10038.
- McAlearney, A. (2005). Leadership development in health care: Results of two nationwide studies. *Academy of Management Proceedings, USA*, H1-H6. Retrieved from <http://www.aomonline.org>
- McArthur, C. (1956). The dynamic model. *Journal of Counseling Psychology, 3*(3), 168-171. doi:10.1037/h0042027.
- McGraw, K., & Wong, S. (1996a). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46. doi:10.1037/1082-989X.1.1.30.

- McGraw, K., & Wong, S. (1996b). 'Forming inferences about some intraclass correlations coefficients': Correction. *Psychological Methods*, 1(4), doi:10.1037/1082-989X.1.4.390.
- Morey, L., Waugh, M., & Blashfield, R. (1985). MMPI scales for DSM-III personality disorders: Their derivation and correlates. *Journal of Personality Assessment*, 49(3), 245-251. doi:10.1207/s15327752jpa4903_5.
- Morgan, C. D., & Murray, H. A. (1938). Thematic Apperception Test. In H. A. Murray (Ed.), *Explorations in personality* (pp. 530-545). New York, NY: Oxford University Press.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176. doi:10.1177/014662169201600206.
- Murray, H., & MacKinnon, D. (1946). Assessment of OSS personnel. *Journal of Consulting Psychology*, 10(2), 76-80. doi:10.1037/h0057480.
- Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologist Press.
- Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, 31(1), 47-73. doi:10.1177/0146621605287691.
- Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory* (3rd Edition). McGraw-Hill Series in Psychology, McGraw-Hill, Inc., New York: NY, 264-265.

- Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment, 14*(1), 50-59. doi:10.1037/1040-3590.14.1.50.
- Osborn, R. N., & Marion, R. (2009). Contextual leadership, transformational leadership and the performance of international innovation seeking alliances. *The Leadership Quarterly, 20*(2009), 191- 206. doi:10.1016/j.leaqua.2009.01.010.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks: CA: Sage.
- Penfield, R., & Bergeron, J. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement, 29*, 218-233. doi:1.1177/0146621604270412.
- Peterson, R. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research, 28*, 450-461. doi:10.1086/323732.
- Peterson, S., Walumbwa, F., Byron, K., & Myrowitz, J., (2009). CEO positive psychological traits, transformational leadership, and firm performance in high-technology start-up and established firms. *Journal of Management, 35*(2), 348-368. doi:10.1177/0149206307312512.
- Pfizer fights back against Trovan study allegations. (2001, October). *Medical Marketing and Media, 36*(10), 24-26. doi:88073413.
- Pilisuk, M. (1998). The hidden structure of contemporary violence. *Peace and Conflict: Journal of Peace Psychology, 4*(3), 197-216. doi:10.1207/s15327949pac0403_1.

- Pittenger, D. (2005). Cautionary comments regarding the Myers-Briggs Type Indicator. *Consulting Psychology Journal: Practice and Research*, 57(3), 210-221.
doi:10.1037/1065-9293.57.3.210.
- Posner, B., & Kouzes, J. (1988). Development and validation of the Leadership Practices Inventory. *Educational and Psychological Measurement*, 48(2), 483-496.
doi:10.1177/0013164488482024.
- Porter, S., & Whitcomb, M. (2007). Mixed-mode contacts in Web surveys: Paper is not necessarily better. *Public Opinion Quarterly*, 71(4), 635-648.
doi:10.1093/poq/nfm038.
- Post, J. (2008). Kim Jong-Il of North Korea: In the shadow of his father. *International Journal of Applied Psychoanalytic Studies*, 5(3), 191-210. doi:10.1002/aps.167.
- Purvanova, R., & Bono, J. (2009). Transformational leadership in context: Face-to-face and virtual teams. *Leadership Quarterly*, 20(3), 343-357.
doi:10.1016/j.leaqua.2009.03.004.
- Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19(1), 49-57.
Retrieved from <http://www.bpsjournals.co.uk/journals/bjmosp>
- Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2008). An IRT analysis of the personal optimism scale. *European Journal of Psychological Assessment*, 24(1), 49-56. doi:10.1027/1015-5759.24.1.49.

- Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4(3), 207-230.
doi:10.2307/1164671.
- Reeve, B., Hays, R., Chang, C., & Perfitto, E. (2007). Applying item response theory to enhance health outcomes assessment. *Quality of Life Research*, 16, 1-3.
doi:10.1007/s11136-007-9220-6.
- Reise, S., Smith, L., & Furr, R. (2001). Invariance on the NEO PI-R neuroticism scale. *Multivariate Behavioral Research*, 36(1), 83-110.
doi:10.1207/S15327906MBR3601_04.
- Reise, S., & Waller, N. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, 8(2), 164-184.
doi:10.1037/1082-989X.8.2.164.
- Reise, S., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. *Journal of Educational Measurement*, 27(2), 133-44. doi:
10.1111/j.1745-3984.1990.tb00738.x.
- Resick, C., Whitman, D., Weingarden, S., & Hiller, N. (2009). The bright-side and the dark-side of CEO personality: Examining core self-evaluations, narcissism, transformational leadership, and strategic influence. *Journal of Applied Psychology*, 94(6), 1365-1381. doi:10.1037/a0016238.
- Roberts, J. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement*, 32(5), 407-423.
doi:10.1177/0146621607301278.

- Roberts, J., Fang, H., Cui, W., & Wang, Y. (2006). GGUM2004: A Windows-Based Program to Estimate Parameters in the Generalized Graded Unfolding Model. *Applied Psychological Measurement, 30*(1), 64-65.
doi:10.1177/0146621605280141.
- Roberts, J., & Shim, H. S. (2008). *GGUM2004 technical reference manual*. Retrieved February 8, 2010 from
<http://www.psychology.gatech.edu/unfolding/FreeSoftware.html>
- Ronan, M. (2008). The clergy sex abuse crisis and the mourning of American Catholic innocence. *Pastoral Psychology, 56*(3), 321-339. doi:10.1007/s11089-007-0099-5.
- Rosen, H., & Rosen, R. (1955). A comparison of parametric and non-parametric analyses of opinion data. *Journal of Applied Psychology, 39*(6), 401-404.
doi:10.1037/h0041250.
- Rosenthal, S., & Pittinsky, T. (2006). Narcissistic leadership. *The Leadership Quarterly, 17*(6), 617-633. doi:10.1016/j.leaqua.2006.10.005.
- Rowold, J., & Heinitz, K. (2007). Transformational and charismatic leadership: Assessing the convergent, divergent and criterion validity of MLQ and the CKS. *The Leadership Quarterly, 18*(2), 121-133. doi:10.1016/j.leaqua.2007.01.003.
- Russell, D. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in personality and social psychology bulletin. *Personality and Social Psychology Bulletin, 28*, 1629-1646. doi:10.1177/014616702237645.

- Sala, F. (2003). Executive blind spots: Discrepancies between self- and other-ratings. *Consulting Psychology Journal: Practice and Research*, 55(4), 222-229. doi:10.1037/1061-4087.55.4.222.
- Sala, F., & Dwight, S. (2002). Predicting executive performance with multirater surveys: Whom you ask makes a difference. *Consulting Psychology Journal: Practice and Research*, 54(3), 166-172. doi:10.1037/1061-4087.54.3.166.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement 17*, 34(4) pt. 2.
- Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, 38(2), 203-219. doi:10.1007/BF02291114.
- Samejima, F. (1977a). A use of the information function in tailored testing. *Applied Psychological Measurement*, 1(2), 233-247. doi:10.1177/014662167700100209.
- Samejima, F. (1977b). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, 42(2), 193-198. doi:10.1007/BF02294048.
- Schriesheim, C. (1982). The great high consideration-high initiating structure leadership myth: Evidence on its generalizability. *Journal of Social Psychology*, 116(2), 221-228. Retrieved from <http://www.heldref.org/pubs/soc/about.html>
- Schyns, B., Felfe, J., & Blank, H. (2007). Is charisma hyper-romanticism? Empirical evidence from new data and a meta-analysis. *Applied Psychology: An International Review*, 56(4), 505-525. doi:10.1111/j.1464-0597.2007.00302.x.
- Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493. doi:10.1177/0265532208094273.

- Scherbaum, C., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research. *The Leadership Quarterly, 17*(4), 366-386. doi:10.1016/j.leaqua.2006.04.005.
- Shrout, P.E., & Fleiss, J.L. (1979). "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin 86* (2): 420–428. doi:10.1037//0033-2909.86.2.420.
- Smith, A., Rush, R., Velikova, G., Wall, L., Wright, E., Stark, D., . . . Sharpe, M. (2007). The initial development of an item bank to assess and screen for psychological distress in cancer patients. *Psycho-Oncology, 16*(8), 724-732. doi:10.1002/pon.1117.
- Snodgrass, J., Douthill, S., Ellis, R., Wade, S., & Plemons, J. (2008). Occupational therapy practitioners' perceptions of rehabilitation managers' leadership styles and the outcomes of leadership. *Journal of Allied Health, 37*(1), 38-44. Retrieved from <http://www.ingentaconnect.com/content/asahp/jah>
- SPSS (2009). SPSS: Version18 for Windows [Computer program]. Chicago: SPSS, Inc.
- Stark, S, Chernyshenko, S., Chuah, D., Lee, W., & Wadlington, P. (2001). MODFIT (version 1.1) [Computer program]. Retrieved from http://io.psych.uiuc.edu/irt/mdf_modfit.asp
- Stodgill, R. M. (1963). *Manual for leadership Behavior Description Questionnaire – Form XII: An Experimental Revision*, Bureau of Business Research, The Ohio State University, Columbus, OH.

- Sullivan, R., & Lee, K. (2008). Organizing immigrant women in America's sweatshops: Lessons from the Los Angeles Garment Worker Center. *Signs*, 33(3), 527-531. doi:10.1086/523807.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics (5th ed.)*. Needham Heights, MA: Allyn & Bacon.
- Tejeda, M., Scandura, T., & Pillai, R. (2001). MLQ revisited: Psychometric properties and recommendations. *Leadership Quarterly*, 12(1), 31-52. doi:10.1016/S1048-9843(01)00063-7.
- Teresi, J., & Fleishman, J. (2007). Differential item functioning and health assessment. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, 16, 33-42. doi:10.1007/s11136-007-9184-6.
- Terpstra, D., Mohamed, A., & Kethley, R. (1999). An analysis of federal court cases involving nine selection devices. *International Journal of Selection and Assessment*, 7(1), 26-34. doi:10.1111/1468-2389.00101.
- Thissen, D., Chen, W., & Bock, R. (2003). MULTILOG (version 7.0.2327.3) [Computer program]. Lincolnwood, IL: Scientific Software.
- Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, 104(3), 385-395. doi:10.1037/0033-2909.104.3.385.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99(1), 118-128. doi:10.1037/0033-2909.99.1.118.

- Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement*, 7(2), 211-226.
doi:10.1177/014662168300700209.
- Thornton, G., & Gibbons, A. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review*, 19(3), 169-187.
doi:10.1016/j.hrmr.2009.02.002.
- Thurstone, L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(7), 433-451. doi:10.1037/h0073357.
- Thurstone, L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273-286. doi:10.1037/h0070288.
- University of Maryland. (2009). The James MacGregor Burns Academy of Leadership. Retrieved September 17, 2009 from
<http://www.academy.umd.edu/People/facultyStaffindividual.asp?DBID=89>
- Van Iddekinge, C., Ferris, G., & Heffner, T. (2009). Test of a multistage model of distal and proximal antecedents of leader performance. *Personnel Psychology*, 62(3), 463-495. doi:10.1111/j.1744-6570.2009.01145.x.
- Vidotto, G., Carone, M., Jones, P., Salini, S., & Bertolotti, G. (2007). Maugeri Respiratory Failure questionnaire reduced form: A method for improving the questionnaire using the Rasch model. *Disability and Rehabilitation: An International, Multidisciplinary Journal*, 29(13), 991-998.
doi:10.1080/09638280600926678.

- Walter, V. (2000). *16PF-Fifth edition: Personal career development profile*. Champaign, IL: Institute for Personality and Ability Testing.
- Walumbwa, F., Avolio, B., & Zhu, W. (2008). How transformational leadership weaves its influence on individual job performance: The role of identification and efficacy beliefs. *Personnel Psychology, 61*(4), 793-825. doi:10.1111/j.1744-6570.2008.00131.x.
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. doi:10.1037/0003-066X.54.8.594.
- Wirth, R., & Edwards, M. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58-79. doi:10.1037/1082-989X.12.1.58.
- Woods, C. M. (2008). Ramsay-curve item response theory for the three-parameter logistic item response model. *Applied Psychological Measurement, 32*, 447-465. doi:10.1177/0146621607308014.
- Wright, B. (1977). Misunderstanding the Rasch model. *Journal of Educational Measurement, 14*(3), 219-225. doi:10.1111/j.1745-3984.1977.tb00039.x.
- Wright, K. (2005). Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication, 10*(3), article 11. Retrieved from <http://jcmc.indiana.edu/vol10/issue3/wright.html>

- Wylie, D., & Gallagher, H. (2009). Transformational leadership behaviors in allied health professions. *Journal of Allied Health, 38*(2), 65-73. Retrieved from <http://www.ingentaconnect.com/content/asahp/jah>
- Yukl, G. (2006). *Leadership in organizations* (6th ed.). Upper River, NJ: Prentice Hall Inc.
- Yukl, G., & Lepsinger, R. (1990). Preliminary report on validation of the Managerial Practices Survey. *Measures of Leadership, 223-237*. West Orange, NJ: Leadership Library of America.
- Zagorsek, H., Stough, S., & Jaklic, M. (2006). Analysis of the reliability of the Leadership Practices Inventory in the item response theory framework. *International Journal of Selection & Assessment, 14*(2), 180-191.
doi:10.1111/j.1468-2389.2006.00343.x.
- Zeidner, M., Roberts, R., & Matthews, G. (2008). The science of emotional intelligence: Current consensus and controversies. *European Psychologist, 13*(1), 64-78.
doi:10.1027/1016-9040.13.1.64.

Appendix A: The GRM estimates without and with basketball players

Items	Facet	Corporate samples ($n = 1,409$)						Corporate and athletic samples ($N = 1,703$)					
		α	δ_1	δ_2	δ_3	δ_4	IIF	α	δ_1	δ_2	δ_3	δ_4	IIF
mlq10	IIA1	2.15	-1.60	-0.81	0.10	1.16	-0.92	2.23	-1.54	-0.75	0.17	1.20	-0.86
mlq18	IIA2	1.43	-2.74	-1.83	-0.67	0.76	-1.86	1.45	-2.65	-1.74	-0.60	0.81	-1.75
mlq21	IIA3	2.02	-2.51	-1.32	-0.18	1.10	-1.26	2.17	-2.21	-1.14	-0.06	1.17	-1.16
mlq25	IIA4	1.80	-2.61	-1.61	-0.65	0.68	-1.49	1.77	-2.66	-1.61	-0.57	0.76	-1.56
mlq06	IIB1	1.58	-2.34	-1.06	0.01	1.42	-0.65	1.28	-2.72	-1.26	-0.07	1.45	-0.71
mlq14	IIB2	1.87	-2.01	-0.99	-0.08	1.12	-0.74	1.67	-2.16	-1.04	-0.01	1.25	-0.76
mlq23	IIB3	1.77	-2.64	-1.45	-0.28	1.21	-1.43	1.76	-2.54	-1.37	-0.16	1.31	-1.37
mlq34	IIB4	1.91	-2.53	-1.41	-0.38	0.98	-1.28	1.63	-2.81	-1.57	-0.47	0.96	-1.32
mlq09	IM1	1.55	-2.95	-1.77	-0.53	1.00	-1.84	1.35	-3.19	-1.92	-0.60	1.05	-1.94
mlq13	IM2	1.71	-2.63	-1.42	-0.36	1.09	-1.14	1.45	-2.95	-1.63	-0.46	1.10	-1.30
mlq26	IM3	1.85	-1.92	-0.86	0.25	1.54	-0.88	1.72	-2.04	-0.89	0.28	1.60	-0.88
mlq36	IM4	1.97	-3.26	-1.94	-0.81	0.81	-1.78	1.86	-3.37	-2.00	-0.78	0.85	-1.88
mlq02	IS1	1.64	-3.15	-1.65	-0.18	1.42	-1.63	1.62	-3.13	-1.57	-0.06	1.53	-1.54
mlq08	IS2	1.57	-3.28	-2.07	-0.67	1.01	-2.27	1.55	-3.05	-1.79	-0.48	1.16	-1.86
mlq30	IS3	1.88	-2.64	-1.55	-0.20	1.38	-1.77	1.85	-2.40	-1.31	-0.04	1.47	-1.47
mlq32	IS4	1.81	-2.85	-1.48	-0.32	1.24	-1.28	1.78	-2.69	-1.34	-0.16	1.36	-1.17
mlq15	IC1	1.65	-1.98	-0.69	0.40	1.51	0.3	1.52	-2.09	-0.75	0.34	1.55	0.01
mlq19	IC2	1.87	-2.64	-1.53	-0.67	0.46	-1.13	1.84	-2.44	-1.45	-0.57	0.55	-1.16
mlq29	IC3	1.66	-2.53	-1.37	-0.23	1.13	-1.33	1.62	-2.32	-1.16	-0.08	1.23	-0.96
mlq31	IC4	2.26	-2.05	-0.91	0.10	1.29	-0.83	2.24	-1.93	-0.83	0.18	1.33	-0.77

Note. All GRM values in logistic metric, IIA = idealized influence attributed, IIB = idealized influence behavioral, IM = inspiration motivation, IS = intellectual stimulation, IC = individual consideration, α = discrimination, δ_i = category boundaries, IIF = location along theta for the maximum value of the item information function.

Appendix B: The GRM and the GGUM fit metrics without basketball players.

Table B1

The GRM and the GGUM Fit Metrics for Telecommunications and Executives of 26 Companies (n = 1,409)

GRM frequency table of X ² /df									
	<1	1<2	2<3	3<4	4<5	5<7	>7	M	SD
Singlet	20	0	0	0	0	0	0	0.09	0.09
Doublet	4	14	4	0	1	1	0	1.84	1.12
Triplet	1	10	1	0	0	0	0	1.54	0.49
GGUM frequency table of X ² /df									
	<1	1<2	2<3	3<4	4<5	5<7	>7	M	SD
Singlet	20	0	0	0	0	0	0	0.18	0.06
Doublet	1	10	8	2	1	2	0	2.56	1.47
Triplet	0	8	3	1	0	0	0	2.02	0.60

Table B2

The GRM and the GGUM Fit Metrics for Telecommunications and Executives of 26 Companies (n=1,407) Adjusted to Normative Sample Size of 3,000

GRM frequency table of X ² /df									
	<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	SD
Singlet	20	0	0	0	0	0	0	0.00	0.00
Doublet	4	6	8	1	3	0	2	2.79	2.39
Triplet	1	5	3	2	1	0	0	2.15	1.03
GGUM frequency table of X ² /df									
	<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	SD
Singlet	20	0	0	0	0	0	0	0.00	0.00
Doublet	1	2	6	5	5	2	3	4.33	3.14
Triplet	0	2	5	2	2	1	0	3.18	1.27

Appendix C: The GGUM parameter estimates.

Item	Facet	GGUM with incomplete data removed ($N = 1703$)						IIF
		α	δ	τ_0	τ_1	τ_2	τ_3	
mlq10	IIA1	1.32	2.91	-4.16	-3.70	-2.78	-1.78	-0.57
mlq18	IIA2	0.80	3.17	-4.95	-5.01	-3.99	-2.45	-1.38
mlq21	IIA3	1.49	3.02	-5.02	-4.16	-3.15	-1.88	-1.08
mlq25	IIA4	1.13	2.94	-5.15	-4.49	-3.70	-2.27	-1.37
mlq06	IIB1	0.72	3.29	-5.71	-4.41	-3.63	-1.85	-0.71
mlq14	IIB2	1.00	3.07	-4.95	-4.05	-3.23	-1.89	-0.65
mlq23	IIB3	1.16	2.80	-5.03	-4.19	-3.08	-1.45	-1.22
mlq34	IIB4	1.05	2.99	-5.41	-4.45	-3.64	-2.10	-1.21
mlq09	IM1	0.85	2.29	-4.74	-4.14	-3.11	-1.21	-1.55
mlq13	IM2	0.90	3.15	-5.63	-4.60	-3.85	-2.07	-1.22
mlq26	IM3	1.05	3.37	-5.17	-4.32	-3.18	-1.77	-0.66
mlq36	IM4	1.36	3.03	-6.03	-4.87	-3.97	-2.21	-1.64
mlq02	IS1	1.11	2.98	-5.83	-4.56	-3.10	-1.42	-1.41
mlq08	IS2	0.99	3.12	-5.65	-4.89	-3.77	-1.92	-1.57
mlq30	IS3	1.12	3.12	-5.18	-4.50	-3.25	-1.58	-1.26
mlq32	IS4	1.19	2.87	-5.37	-4.18	-3.17	-1.46	-1.07
mlq15	IC1	0.88	3.24	-5.17	-3.91	-3.04	-1.81	-0.11
mlq19	IC2	1.08	2.67	-4.76	-4.00	-3.40	-2.29	-1.06
mlq29	IC3	0.93	3.14	-5.17	-4.28	-3.38	-1.93	-0.80
mlq31	IC4	1.52	3.10	-4.89	-3.93	-3.00	-1.81	-0.56

Note. All GGUM values in normal metric, IIA= idealized influence attributed, IIB=idealized influence behavioral, IM=inspiration motivation, IS=intellectual stimulation, IC=individual consideration, α =discrimination, δ =location parameter, τ_1 - τ_4 =subjective response thresholds, IIF=location along theta of the maximum value of the item information function.

Appendix D: The GRM graphs for idealized influence attributed items.

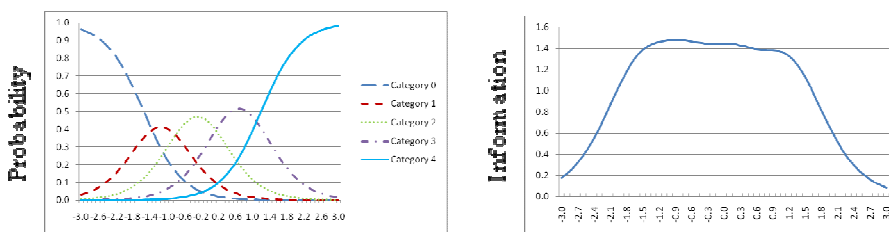


Figure D1. The GRM mlq10 characteristic curves and IIF.

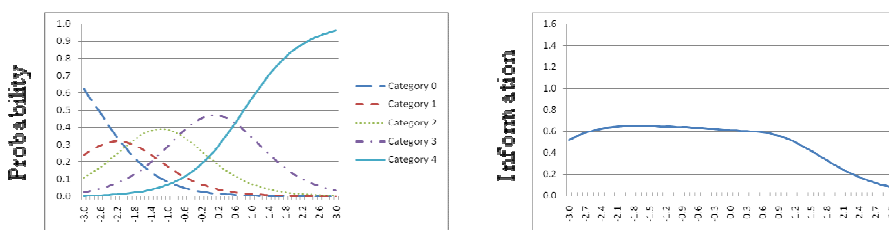


Figure D2. The GRM mlq18 theta characteristic curves and IIF.

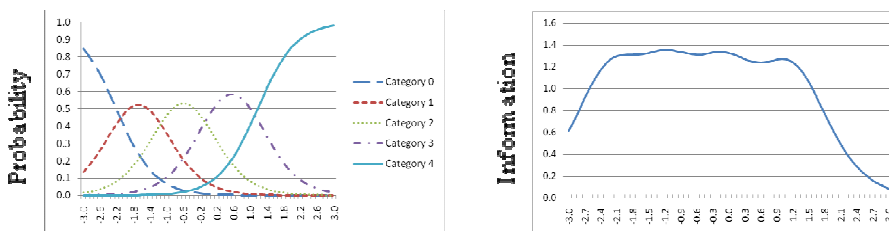


Figure D3. The GRM mlq21 theta characteristic curves and IIF.

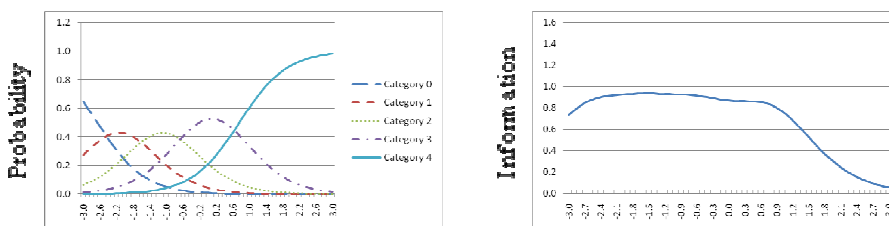


Figure D4. The GRM mlq25 theta characteristic curves and IIF.

Appendix E: The GRM graphs for idealized influence behavioral items.

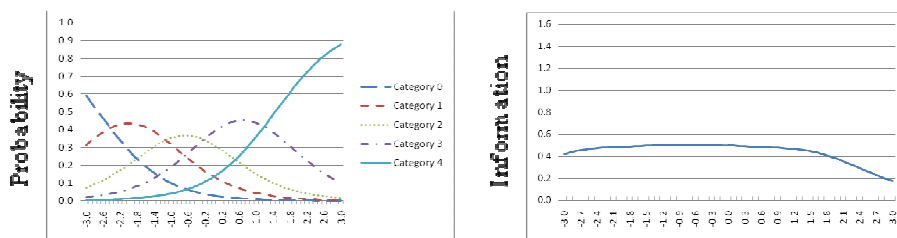


Figure E1. The GRM mlq6 theta characteristic curves and IIF.

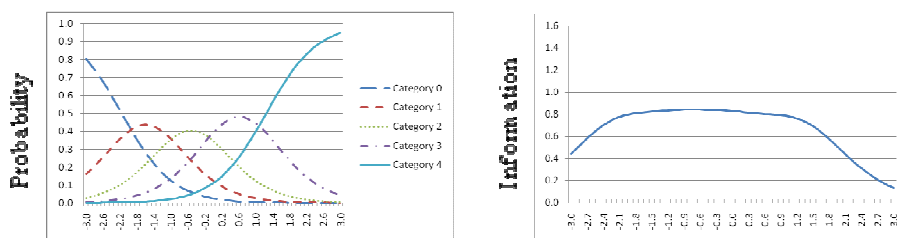


Figure E2. The GRM mlq14 theta characteristic curves and IIF.

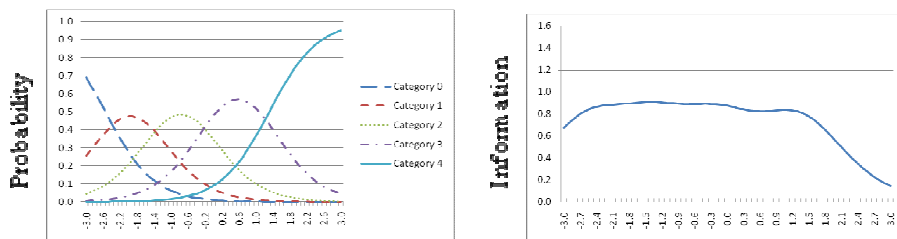


Figure E3. The GRM mlq23 theta characteristic curves and IIF.

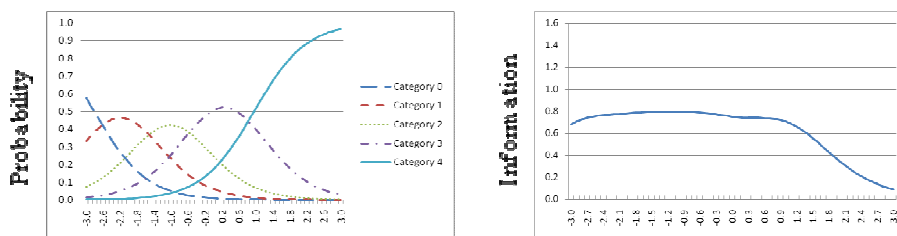


Figure E4. The GRM mlq34 theta characteristic curves and IIF.

Appendix F: The GRM graphs for inspirational motivation items.

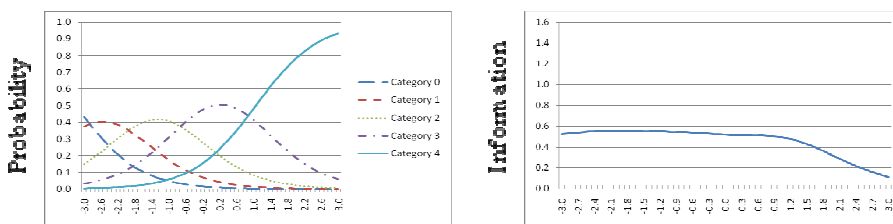


Figure F1. The GRM mlq9 theta characteristic curves and IIF.

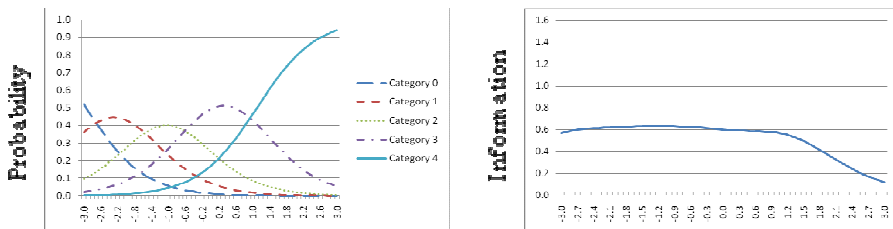


Figure F2. The GRM mlq13 theta characteristic curves and IIF.

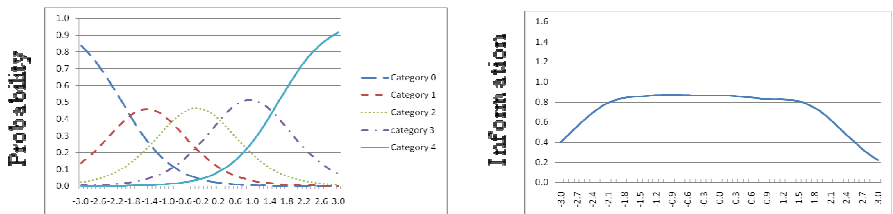


Figure F3. The GRM mlq26 theta characteristic curves and IIF.

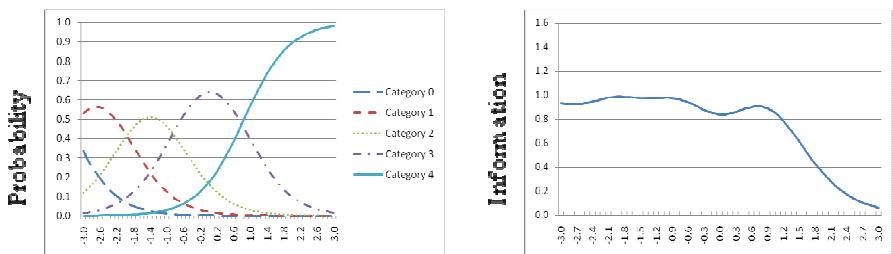


Figure F4. The GRM mlq36 theta characteristic curves and IIF.

Appendix G: The GRM graphs for intellectual stimulation items.

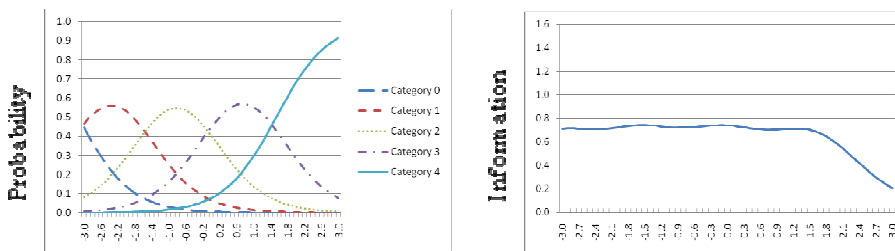


Figure G1. The GRM mlq2 theta characteristic curves and IIF.

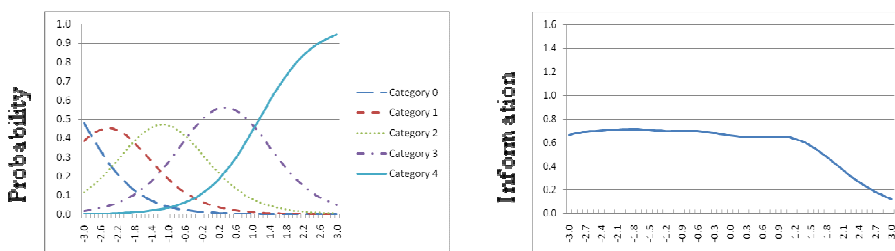


Figure G2. The GRM mlq8 theta characteristic curves and IIF.

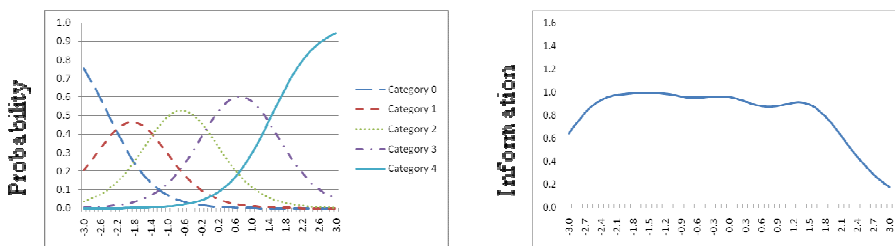


Figure G3. The GRM mlq30 theta characteristic curves and IIF.

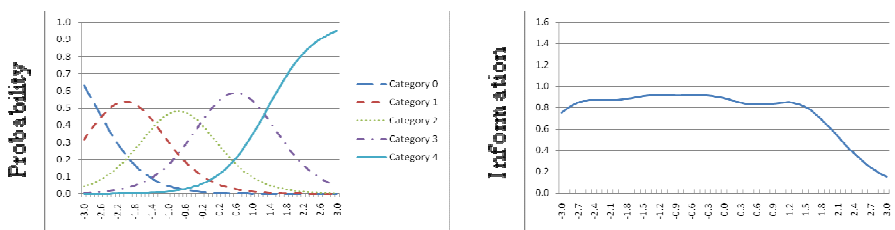


Figure G4. The GRM mlq32 theta characteristic curves and IIF.

Appendix H: The GRM graphs for individual consideration items.

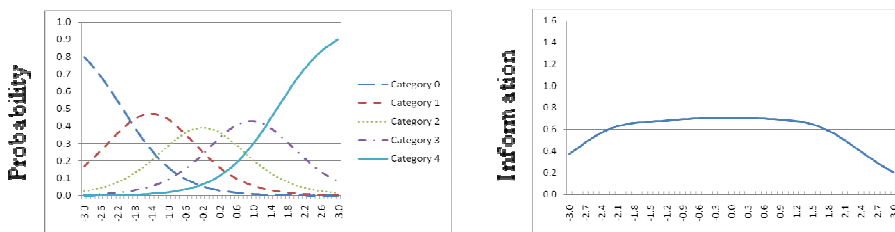


Figure H1. The GRM mlq15 theta characteristic curves and IIF.

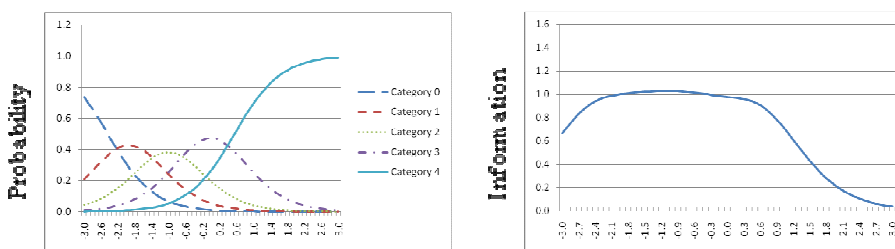


Figure H2. The GRM mlq19 theta characteristic curves and IIF.

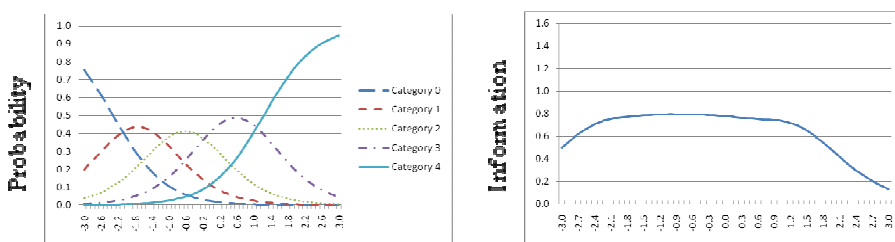


Figure H3. The GRM mlq29 theta characteristic curves and IIF.

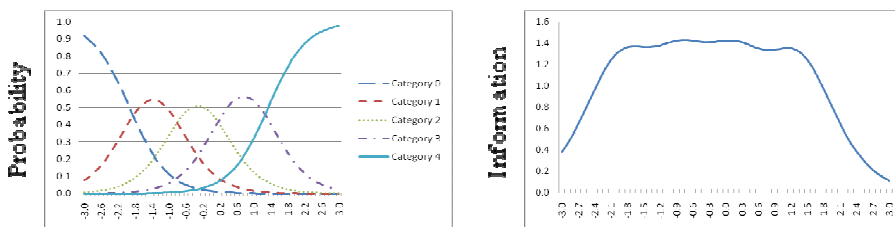


Figure H4. The GRM mlq31 theta characteristic curves and IIF.

Appendix I: The GGUM graphs for idealized influence attributed items.

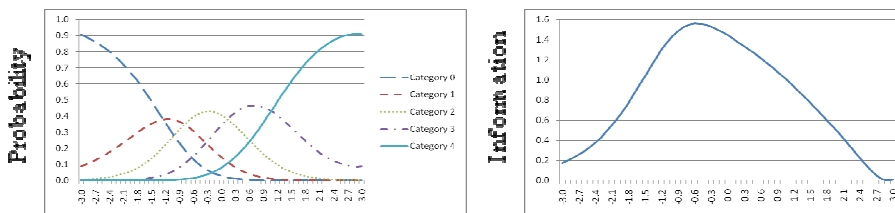


Figure 11. The GGUM mlq10 theta characteristic curves and IIF.

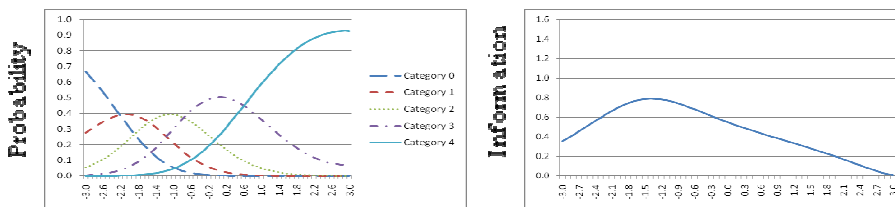


Figure 12. The GGUM mlq18 theta characteristic curves and IIF.

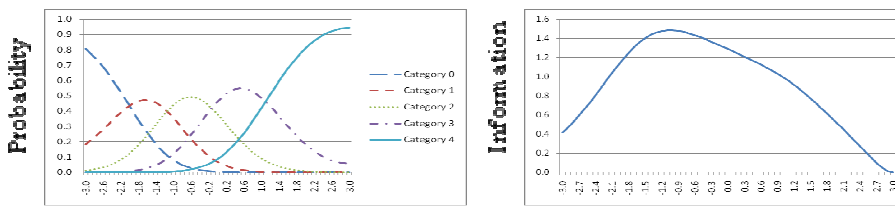


Figure 13. The GGUM mlq21 theta characteristic curves and IIF.

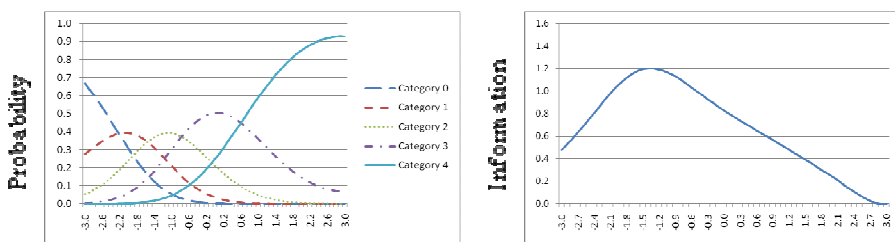


Figure 14. The GGUM mlq25 theta characteristic curves and IIF.

Appendix J: The GGUM graphs for idealized influence behavioral items.

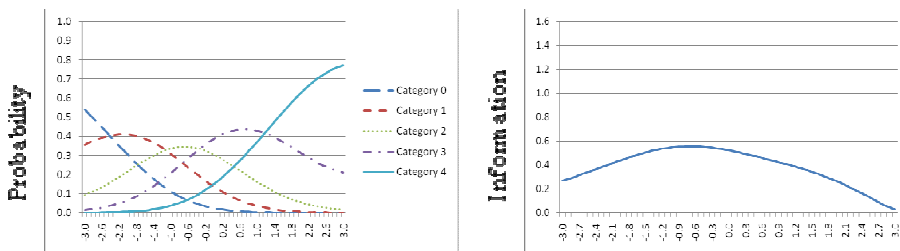


Figure J1. The GGUM mlq6 theta characteristic curves and IIF.

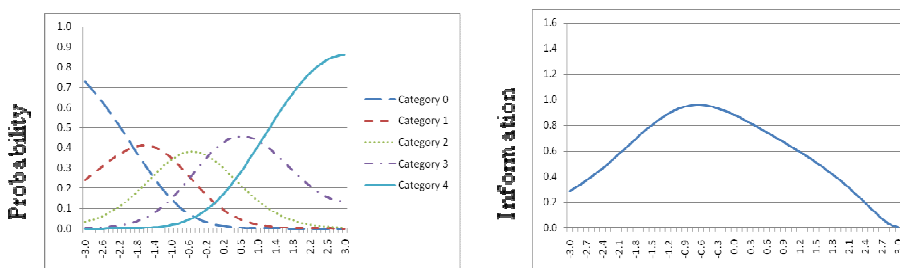


Figure J2. The GGUM mlq14 theta characteristic curves and IIF.

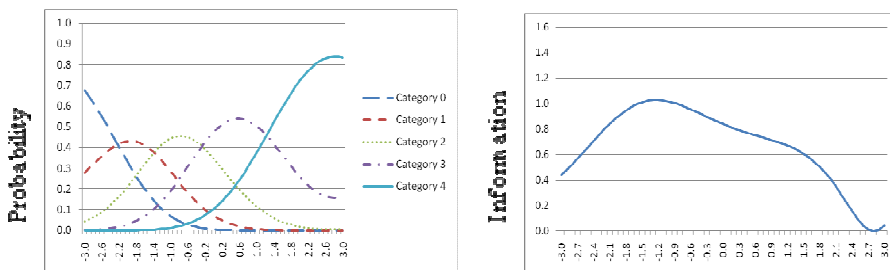


Figure J3. The GGUM mlq23 theta characteristic curves and IIF.

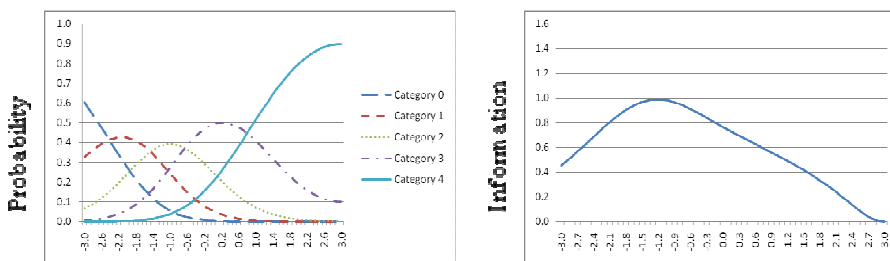


Figure J4. The GGUM mlq34 theta characteristic curves and IIF.

Appendix K: The GGUM graphs for inspirational motivation items.

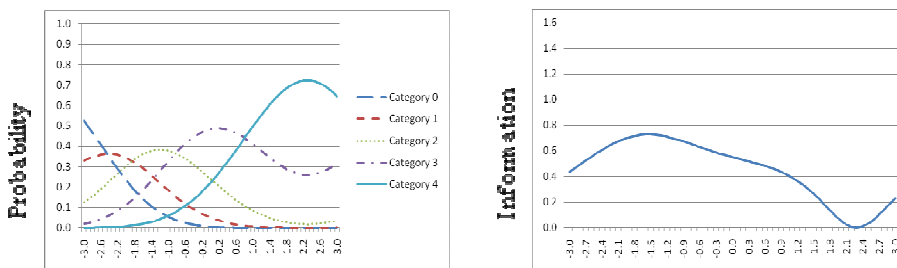


Figure K1. The GGUM mlq9 theta characteristic curves and IIF.

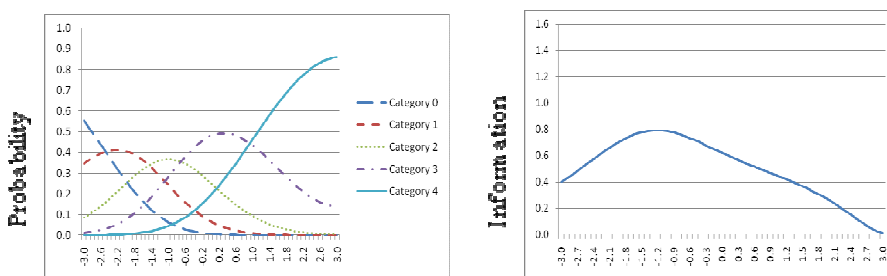


Figure K2. The GGUM mlq13 theta characteristic curves and IIF.

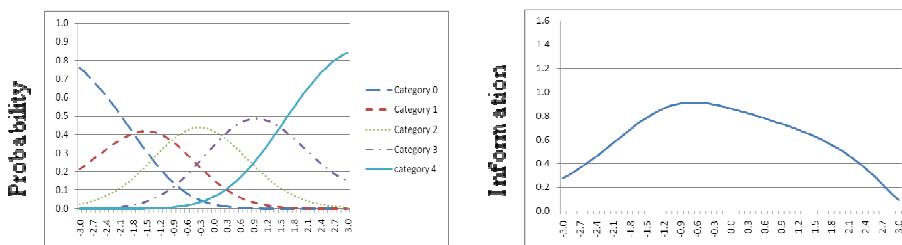


Figure K3. The GGUM mlq26 theta characteristic curves and IIF.

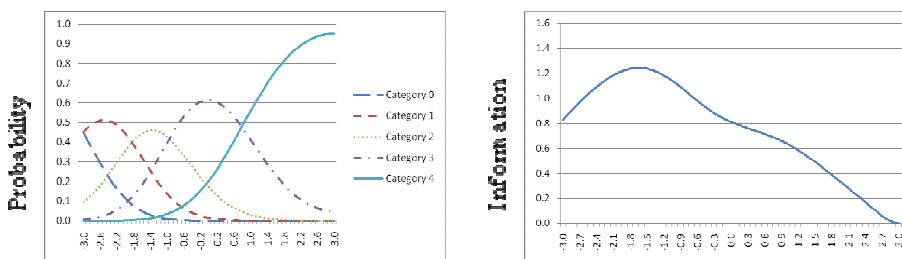


Figure K4. The GGUM mlq36 theta characteristic curves and IIF.

Appendix L: The GGUM graphs for intellectual stimulation items.

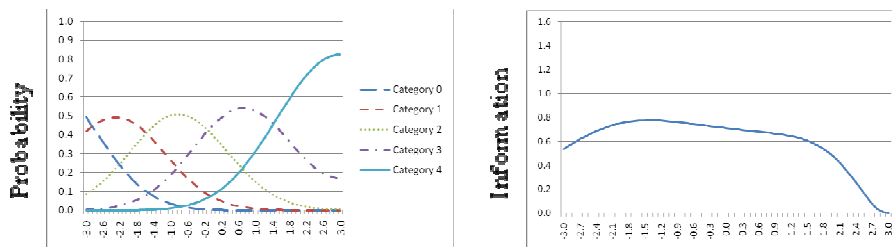


Figure L1. The GGUM mlq2 theta characteristic curves and IIF.

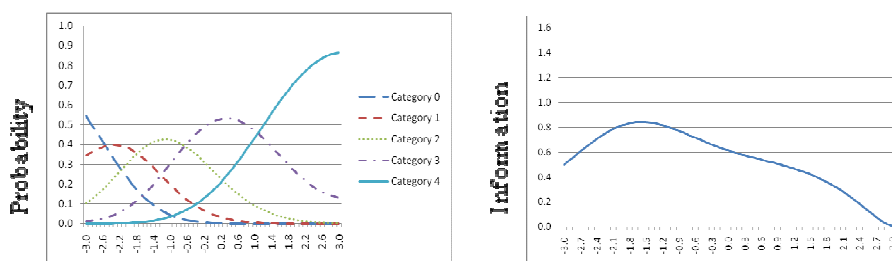


Figure L2. The GGUM mlq8 theta characteristic curves and IIF.

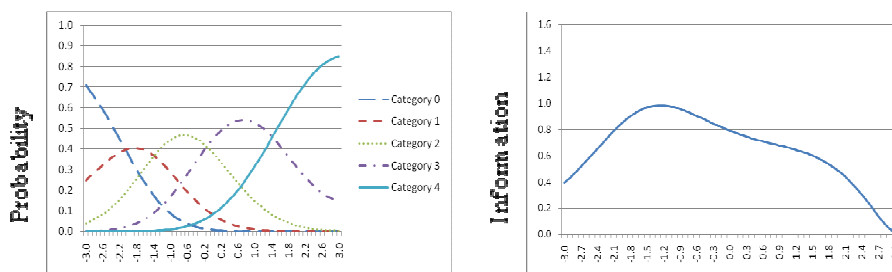


Figure L3. The GGUM mlq30 theta characteristic curves and IIF.

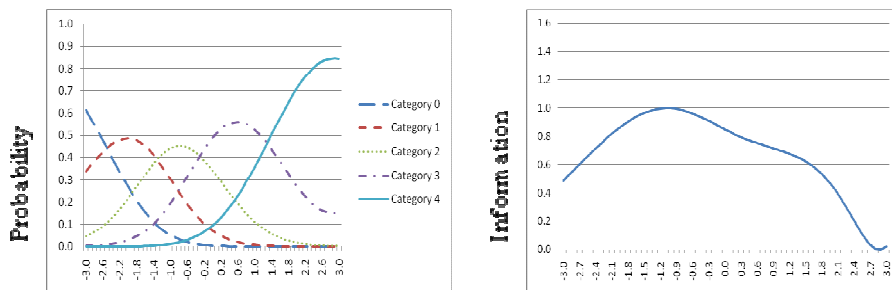


Figure L4. The GGUM mlq32 theta characteristic curves and IIF.

Appendix M: The GGUM graphs for individual consideration items.

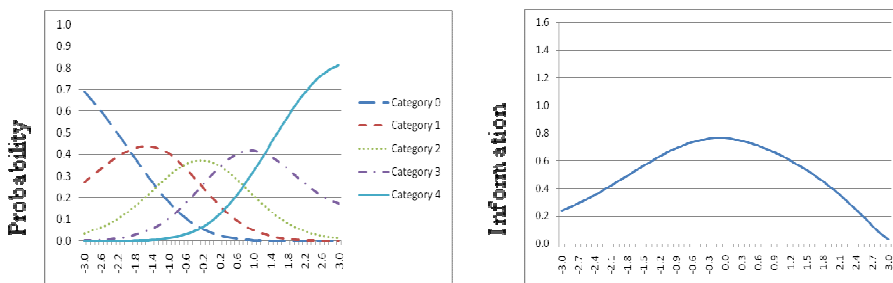


Figure M1. The GGUM mlq15 theta characteristic curves and IIF.

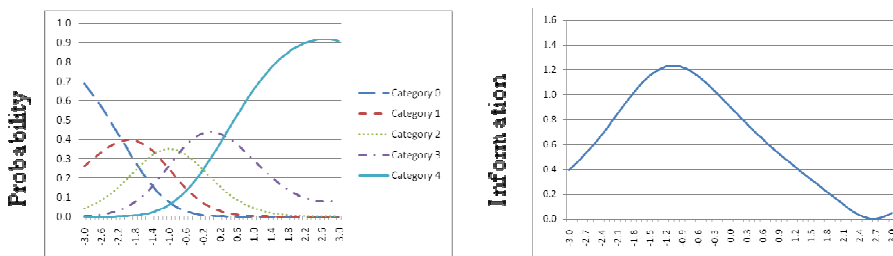


Figure M2. The GGUM mlq19 theta characteristic curves and IIF.

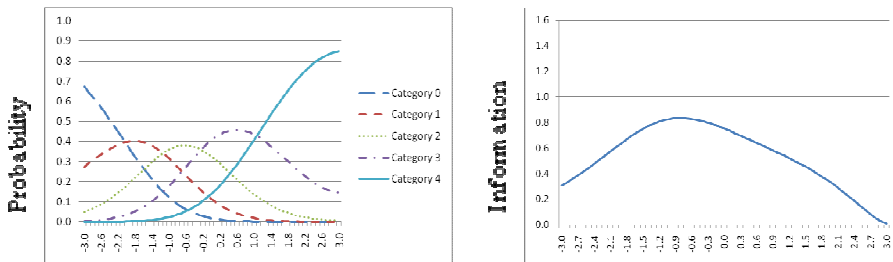


Figure M3. The GGUM mlq29 theta characteristic curves and IIF.

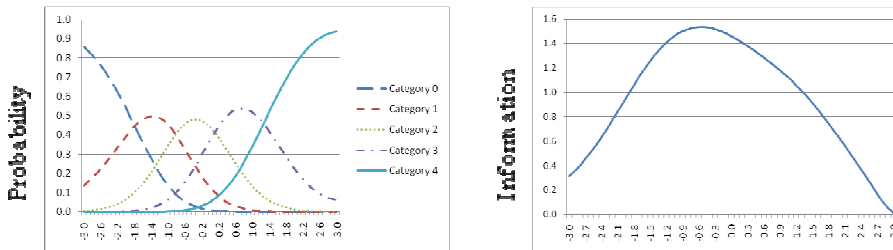
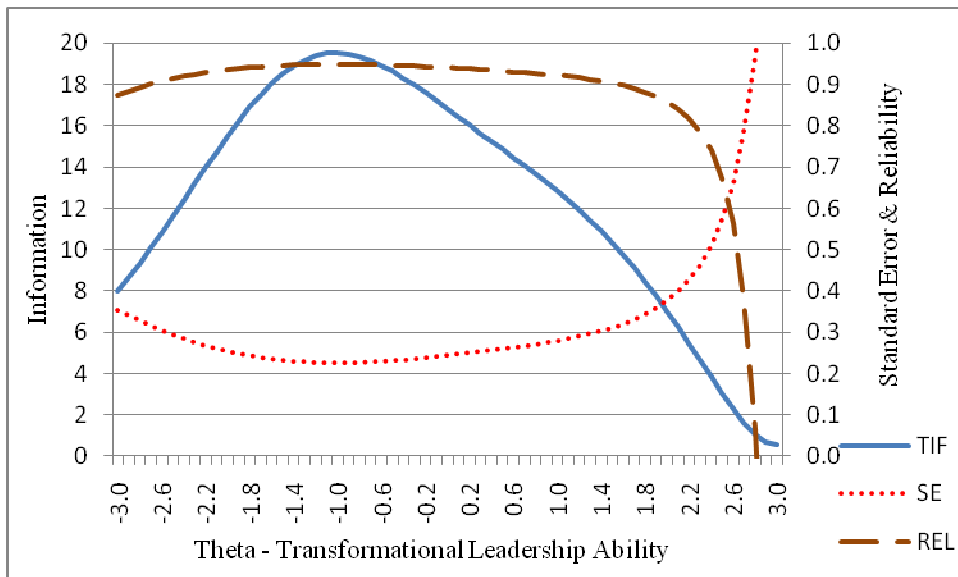


Figure M4. The GGUM mlq31 theta characteristic curves and IIF.

Appendix N: The GGUM test information function, standard error, and reliability.



Dale F. H. Martin

Curriculum Vitae

ACADEMIC EXPERIENCE

- 2005-Present* Candidate for Doctor of Philosophy – Organizational Psychology, **Walden University**, Minneapolis, Minnesota
- 1984-1986* Master of Business Administration, **University of St. Thomas**, St. Paul, Minnesota
- 1981-1982* Bachelor of Science – Electrical Engineering, **Washington University**, St. Louis, Missouri
- 1978-1981* Bachelor of Arts – Physics, **Gustavus Adolphus College**, St. Peter, Minnesota

PROFESSIONAL EXPERIENCE

- 2010-Present* **Managing Director for USA & Canada**
ROS Systems USA, LLC; Naperville, Illinois
- ROS, a worldwide leader in hydraulic reconstruction, using a patented repair process. Responsibilities include expanding industrial usage by inspiring reclamation and restoration of existing materials.*
- 2006-Present* **President**
TFC Info, LLC; Tampa, Florida
- TFC Info is a leading provider of audio and visual market research. Responsibilities include providing the best research based insight for product development business decisions.*
- 2000– 2005* **Executive Vice President**
Cargill Investor Service, Chicago, Illinois

Responsibilities included managing a leading financial services firm with international operations and clients. Creative teams implemented significant new product and service offerings across the globe.

1990-2000

Various Leadership Positions

Cargill, Inc., Minneapolis, Minnesota

Responsibilities included increasing client impact around the world.

1982 – 1990

Engineer & Program Manager

Sperry Rand, St. Paul, Minnesota

Responsibilities included quality control and program management for advanced technology projects.

COMMUNITY SERVICE AND CONSULTING EXPERIENCE

2005-2010

Wayside Cross Ministries, Stephen Ministries, Compassion International, Philippine Frontline Ministries

*Advisor to organizational leaders
Chicago metropolitan area, Illinois*