**Latent Variable Modeling in Business Research:
A Comparison of Regression Based on
IRT and CTT Scores with Structural Equation Models**

Irene Ruen-Rung Lu

Thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the degree

Doctor of Philosophy in Management

**Eric Sprott School of Business
Faculty of Public Affairs and Management
Carleton University
Ottawa, Ontario, Canada**

**Copyright © Irene Ruen-Rung Lu
June 11, 2004**

Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-97832-X

Our file Notre référence

ISBN: 0-612-97832-X

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

ACKNOWLEDGEMENT

First and foremost, I would like to express my profound gratitude for my mentor and thesis supervisor, Dr. Roland Thomas, not only for his insightful guidance, unwavering support and encouragement, but also for his generosity and inestimable patience throughout my graduate studies. I have always felt fortunate to be his student. I have always had the privilege to fit into his busy schedule. I share with him my ups and downs without any reservations. I believe that without him, I could not have accomplished what I have today. He is one of the most important people in my life. I am enormously indebted to him for all that he has done for me. I respect him so much, he is forever my teacher, and despite what he told me after my defense, I will always call him “Professor Thomas”.

I am extremely thankful to Dr. Bruno Zumbo for the direction, constructive comments, and encouragement. I am also particularly appreciative of Dr. Louise Heslop for giving me valuable comments on the draft of this thesis to help me prepare for my defense. It was a real pleasure to have Dr. George Marcoulides as my external examiner, his questions and comments made the defense challenging but enjoyable. I would also like to thank Dr. Barbara Orser and Dr. Linda Duxbury for the permission to use their data, and along with Dr. Scott Bennett, I thank you all for your kind comments.

I also wish to express my thanks to all the people who have helped me in the past years at Carleton University, and to those that have brought a smile to my face. I gratefully acknowledge financial support from Carleton University throughout my studies, as well as two awards from the Ontario Graduate Scholarship Program. I also wish to thank Statistics Canada, the National Program on Complex Data Structures (NPCDS), and the Mathematics of Information Technology and Complex Systems (MITACS) for financial assistance and the opportunity to work as a Statistics Canada intern during the final stages of my Ph.D. program.

In the end, many heartfelt thanks go to Ernest Kwan for providing me with the strength to complete this endeavor. My deepest appreciation to my siblings for taking care of my parents. I would like to dedicate my thesis to my parents; I am deeply sorry for not being able to fulfill my role as a daughter to them in these past few years.

CONTENTS

Chapter 1 Introduction.....	1
1.1 Research Background.....	1
1.2 Research Goals.....	8
1.3 Organization of the thesis.....	10
Chapter 2 An Overview of Classical Test Theory (CTT).....	11
2.1 Preliminary Definitions and Terminologies.....	12
2.2 Introduction to Classical Test Theory.....	16
2.3 Limitations of Classical Test Theory	29
2.4 Summary	31
Chapter 3 An Overview of Item Response Theory (IRT)	33
3.1 Basic Concepts of IRT.....	34
3.2 IRT Model Formulations	39
3.2.1 Models for Dichotomous Responses.....	40
3.2.2 Illustration of a Dichotomous IRT Model in a Business Context.....	45
3.2.3 Models for Polytomous Responses	47
3.2.4 Illustration of a Polytomous IRT Model in a Business Context.....	49
3.3 IRT Item Parameter Estimation	55
3.4 Advantages and Disadvantages of IRT	59
3.4.1 Advantages of IRT	59
3.4.2 Disadvantages of IRT	65
3.5 Applications of IRT in Business Literature	67
3.6 Summary	72
Chapter 4 IRT- and CTT-based Latent Variable Scores	73
4.1 Methods for Predicting IRT Scores	76
4.1.1 Likelihood Based Prediction Methods.....	77
4.1.2 Prediction Methods Based on the Posterior Distribution	79
4.2 Relative Performance of the IRT- and CTT-based Scores	81
4.3 Summary	85
Chapter 5 The Use of Latent Variable Scores in Statistical Analyses	88
5.1 Problems with Directly Using Latent Variable Scores in Statistical Analyses	90
5.2 Studies on the Use of Latent Variable Scores in Regression	95
5.2.1 Scores from Continuous Scale Items: CTT and FA.....	95
5.2.2 Scores from Discrete Scale Items: IRT	100
5.3 Possible Remedies for the Finite Item Bias Generated by the Use of Scores.....	102
5.3.1 The Case of Continuous Indicator Variables	102
5.3.2 Bias-Corrected IRT Scores: Discrete Indicators.....	104
5.3.3 Plausible Values: Discrete Indicators	105
5.3.4 Structural Equation Modeling (SEM): Simultaneous Estimation of Item and Regression Parameters with Discrete Indicators.....	108
5.3.5 Partial Least Squares: Continuous Indicators	111
5.4 Summary	113

Chapter 6	Structural Equation Modeling (SEM)	118
6.1	SEM-Based Latent Regression Models with Continuous Indicators	120
6.2	SEM-Based Latent Regression Models with Dichotomous Indicators	121
6.2.1	The connection between IRT and SEM for Dichotomous Indicators	123
6.3	SEM-Based Latent Regression Models with Polytomous Indicators	125
6.3.1	The connection between IRT and SEM for Polytomous Indicators	126
6.4	Summary	128
Chapter 7	Research Questions	129
Chapter 8	Study Design and Implementation	134
8.1	Design of the Simulation Study for a Simple Latent Regression Model	139
8.1.1	Simulation Procedures for Assessing Comparability and Accuracy of IRT- and CTT-based Scores	139
8.1.2	Simulation Procedures for the Two-Step Approach	146
8.1.3	Simulation Procedures for the IRT-SEM Approach	147
8.1.4	Simulation Procedures for Partial Least Squares Approach with Discrete Indicators	148
8.2	Design of the Simulation Study for a Multiple Latent Regression Model	150
8.3	Simulation Conditions	153
8.4	Summary	155
Chapter 9	Research Results	156
9.1	Comparability and Accuracy of IRT- and CTT-based Scores	158
9.2	Simulation Results on the Percentage Bias of the Two-Step and IRT-SEM Approaches	168
9.2.1	Results on the Percentage Bias in the Simple Latent Regression Model	169
9.2.2	Results on the Percentage Bias in the Multiple Latent Regression Model	178
9.3	Meta-Models for the Simulation Results from the Two-Step and IRT-SEM Approaches	189
9.3.1	Specifications of the Meta-Models	190
9.3.2	An Application of the Meta-Models	199
9.3.3	Graphical Presentation of the Simulation Results based on the Meta-Models	206
9.4	Simulation Results for Partial Least Squares Approach	217
9.5	Summary	220
Chapter 10	Applications to Two Business Research Topics: Entrepreneurship and Work-Family Conflict	222
10.1	The Entrepreneurship Case Study (Binary Indicators)	225
10.2	The Work-Family Conflict Case Study (Polytomous Indicators)	228
10.3	Summary	233
Chapter 11	Concluding Comments	234
11.1	Review and Major Conclusions	234
11.2	Recommendations	239
11.3	Directions for Future Research	241
11.4	Limitations	243
Appendix A:	Proof That $E(e) = 0$ and $\rho(t, e) = 0$	244
Appendix B:	The Derivation of the Formula for Accuracy	246
Appendix C:	A Mathematical Explanation of the Relationship between the Correlation Coefficient of Exploratory Latent Variables and the Magnitude of the Parameter Bias	250
Bibliography		255

LIST OF TABLES

Table 2.1: Reliability estimation related to different design of reliability experiments.....	32
Table 3.1: Major differences between CTT and IRT.....	64
Table 4.1: Summary of empirical studies on the performance of IRT- and CTT-based.....	87
Table 5.1: Approaches to estimating regression coefficients in latent regression model.....	116
Table 5.2: Summary of the studies on the use of predicted latent variable scores in regression.....	117
Table 8.1: Summary of the simulations for the research questions.....	136
Table 8.2: Summary of the simulation procedures for generating discrete item response data and the simulation conditions in simple latent regression model	149
Table 8.3: Summary of the simulation procedures for generating discrete item response data and the simulation conditions in multiple latent regression model.....	152
Table 9.1: Comparability of IRT- and CTT-based scores with symmetrically distributed items	160
Table 9.2: Comparability of IRT- and CTT-based scores with symmetrically distributed items, where items in the measurement model differed in CD.....	161
Table 9.3: Comparability of IRT- and CTT-based scores with asymmetrically distributed items at skew level of 2.....	162
Table 9.4: Accuracy of IRT- and CTT-based Scores with symmetrically distributed items.....	165
Table 9.5: Accuracy of IRT- and CTT-based Scores of simple latent regression model with symmetrically distributed items, where items in the measurement model differed in CD.....	166
Table 9.6: Accuracy of IRT- and CTT-based scores with asymmetrically distributed items at skew level of 2	167
Table 9.7: Effects of scale Length and number of categories on bias in the simple latent regression.....	171
Table 9.8: Effect of sample size on bias in the simple latent regression.....	172
Table 9.9: Effect of measurement precision (CD) on bias in the simple latent regression.....	172
Table 9.10: Effect of symmetry of item's marginal distribution on bias in the simple latent regression.....	174
Table 9.11: Effect of scale length and number of categories in the simple latent regression under a realistic situation.....	175
Table 9.12: IRT outperforms CTT (1).....	176
Table 9.13: IRT outperforms CTT (2).....	176
Table 9.14: Effects of scale length and number of categories on bias in the multiple latent regression	182
Table 9.15: Effect of sample size on bias in the multiple latent regression	183
Table 9.16: Effect of measurement precision (CD) on bias in the multiple latent regression...	184
Table 9.17: Effect of symmetry of item's marginal distribution on bias in the multiple latent regression.....	185

Table 9.18: Effect of scale length and number of categories on bias in the multiple latent regression under a realistic situation.....	186
Table 9.19: Effect of correlation between two independent latent variables on bias in the multiple latent regression.....	187
Table 9.20: Specifications of the meta-model for the two-step approach.....	192
Table 9.21: Simulation settings of the meta-model for the two-step approach in the simple latent regression.....	193
Table 9.22: Simulation settings of the regression summary for the two-step approach in the multiple latent regression.....	193
Table 9.23: Specification of the meta-model for the simultaneous IRT-SEM approach.....	195
Table 9.24: Simulation settings of the regression summary for the IRT-SEM approach in the simple latent regression.....	196
Table 9.25: Simulation settings of the regression summary for the simultaneous IRT-SEM approach in the multiple latent regression.....	196
Table 9.26: Meta-models for the two-step and simultaneous IRT-SEM approaches.....	198
Table 9.27: Illustrations of the application of meta-models (1).....	203
Table 9.28: Illustrations of the application of meta-models (2).....	204
Table 9.29: Illustrations of the application of meta-models (3).....	205
Table 9.30: Comparison of bias in PLS approach with the two-step and IRT-SEM approaches for the simple latent regression.....	218
Table 9.31: Comparison of bias in PLS approach with the two-step and IRT-SEM approaches for the multiple latent regression	219
Table 10.1: Entrepreneurship case study with binary indicators in the simple latent regression.....	226
Table 10.2: Work-family conflict case study with polytomous indicators in the simple latent regression.....	229
Table 10.3: Work-family conflict case study with polytomous indicators in the multiple latent regression.....	230
Table B.1: Cronbach's alpha vs. accuracy for binary indicators.....	249

LIST OF FIGURES

Figure 2.1: Conceptual measurement model.....	13
Figure 3.1: A 3-parameter normal ogive item characteristic curve.....	44
Figure 3.2: Illustration of a dichotomous IRT in business context.....	47
Figure 3.3: Illustration of a polytomous IRT in business context.....	52
Figure 3.4: An example of an item information function	62
Figure 8.1: Path diagrams of the simulated simple latent regression model.....	137
Figure 8.2: Path diagram of the simulated multiple latent regression model.....	138
Figure 8.3: Marginal frequency distributions of discrete indicator variables.....	142
Figure 8.4: Illustrations for threshold creation and categorization of underlying response variables into discrete item indicators.....	143

LIST OF PLOTS

Plot 9.1: Effect of scale length on R^2 bias in latent variable regression (for the two-step and simultaneous IRT-SEM methods).....	210
Plot 9.2: Effect of number of categories on R^2 bias in latent variable regression (for the two-step and simultaneous IRT-SEM methods).....	211
Plot 9.3: Effect of sample size on R^2 bias in latent variable regression (for the simultaneous IRT-SEM methods).....	212
Plot 9.4: Effect of measurement precision (CD) on R^2 bias in latent variable regression (for the two-step and simultaneous IRT-SEM methods).....	213
Plot 9.5: Effect of symmetry of item's marginal distribution on R^2 bias in latent variable regression (for the two-step method).....	214
Plot 9.6: Effect of population R-square on R^2 bias in latent variable regression (for the two-step methods).....	215
Plot 9.7 Effect of correlation between two exploratory latent variables on R^2 bias in multiple latent regression (for the two-step and simultaneous IRT-SEM methods)	216

Chapter 1

Introduction

1.1 Research Background

Researchers in marketing, organization behaviour, and other fields of business research frequently study variables that are abstract and not directly observable, for example, country image, brand loyalty, service quality, job satisfaction, work stress, managerial competence, etc. To measure such variables, referred to herein as *latent variables*, *latent traits*, or *constructs*, a multiple-item scale of measurement is often used, in which the scale measure is a weighted or unweighted sum of responses to the individual items that form the scale. These items in turn are usually measured as dichotomous or polytomous variables. The prevalent guidelines for scale construction in marketing and business research textbooks are based on *classical test theory* (CTT) (e.g., Aaker, Kumar, & Day, 2004; Churchill & Iacobucci, 2002; Cooper & Schindler, 1998), a measurement technology that has its origin in the field of psychometrics. The recent edition of the *Handbook of Marketing Scales* by Bearden and Netemeyer (1999) describes numerous scales that have been constructed using CTT-based procedures. Examples are the Country Image scale by Martin & Eroglu (1993), the Purchasing Involvement scale by Slama & Tashchian (1985), and the Organizational Commitment scale by Hunt, Chonko, & Wood

(1985). Concepts such as *reliability*, *item-total correlations*, and *Cronbach's alpha* are all based on the CTT model, which posits that the observed scale score is the sum of a true score and an error term.

Despite its common acceptance in behavioural research, however, CTT-based procedures of scale construction have many well-documented shortcomings, such as item- or scale-dependent scores, respondent-dependent scale characteristics, and constant standard error of measurement across entire latent variable levels (see, for example, Embretson & Reise, 2000; Hambleton, 1989; Hambleton, Swaminathan, & Rogers, 1991). In an attempt to remedy these shortcomings, psychometricians (a key reference being Lord, 1952) began in the 1950's to develop an alternative model-based approach to measurement, now referred to as *item response theory* (IRT). IRT has existed in parallel with CTT methods for many years, and though both the CTT and IRT approaches to measurement have their proponents, it now appears that IRT is becoming the scale construction method of choice, at least for large measurement projects, such as GMAT testing. Also, a chapter in *Modern Methods for Business Research* edited by Marcoulides (1998) introduces IRT as a "powerful" method for understanding item response behaviour and encourages researchers and practitioners to apply IRT to real-life research issues and problems (McCollam, 1998, p. 44). Most recently, Singh (2004) also suggests that "IRT warrants the serious attention of marketing researchers" (p.205). Given this trend, it is natural to enquire whether or not IRT can profitably be introduced into marketing and other business research, to provide better designed scales with superior measurement properties. Thus the research described in this dissertation began with a detailed reading of the technical literature on IRT, and it was envisaged that some

detailed empirical comparisons between IRT- and CTT-based scales would present evidence to support the introduction of IRT methods into marketing. While the investigation of the literature did reveal numerous advantages of IRT over CTT, which will be described in this thesis, it also revealed some problems that changed the ultimate direction of the research. In fact, the literature review has lead the author to seriously question many of the ways in which scale measures of latent constructs (both IRT- and CTT-based) are used in marketing and other behavioral research, and to explore and develop alternative approaches.

The basic problem relates to the goal of the measurement exercise. In psychometric applications, the goal is often to obtain a measurement on an individual, e.g., the individual's level of depression, or a child's verbal ability. Similar goals may also be encountered in business, e.g., measures of a manager's leadership competency or a particular store's service quality rating may be required. For such individual measurement tasks, it can be argued that from a theoretical perspective, IRT yields better measures than does CTT. For example, specific IRT measures can be obtained that have well defined statistical optimality properties for a single individual subject, or case. However, the typical researcher is frequently interested not in individual measures of a construct but in aggregate measures such as the correlation between two constructs, or the regression coefficients in a linear model relating a dependent construct to a set of independent constructs and other demographic variables. To deal with such research problems, many researchers use a two-step approach (Fox & Jones, 1998; Osgood, McMorris, & Potenza, 2002): (a) obtain measures of the latent variables first, using IRT- or CTT-based latent variable *scores*, and (b) use these scores directly in subsequent

analyses such as regression. Specifically, in four major marketing journals between 1977 and 1994, at least 63 percent of the 103 studies that examined relationships among latent variables in a multi-equation system used composite scores for at least one latent variable in the model (see Baumgartner & Homburg, 1996). In 37 communication-related journals published from 1995-2000, 58 percent of the 118 multi-equation models examining latent variable relationships employed single-indicators or composite scores (see Holbert, 2002).

The problem with the direct use of scale measures in subsequent analyses is now well documented, and stems from the fact that because of measurement error in the predicted scores, the distribution of the predicted scores does not converge to the distribution of the latent variable as the sample size grows, *unless the number of scale items used to derive the scores is very large*. As a result, analyses based on the predicted scores are subject to bias, regardless of sample size (Mislevy, Johnson & Muraki, 1992). Convergence occurs only when the number of scale items goes to infinity, and *finite item bias* is the bias that results from not having a sufficiently large number of items. Nevertheless, given some of the advantages of IRT over CTT, one might expect there to be an appreciable reduction in finite item bias from using IRT-based versus CTT-based scores. Unfortunately, little attention has been focused on this issue in the psychometric or other literature. An investigation of this issue will be described in this thesis.

Another popular approach to studying relationships between latent variables is *structural equation modeling* (SEM). Standard full-information SEM methods implemented in such packages as LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1995), AMOS (Arbuckle & Wothke, 1999), Mplus (Muthén & Muthén, 2001), LINC

(Schoenberg & Arminger, 1990) are different from the aforementioned two-step approach in that they simultaneously estimate the parameters in the latent variable measurement models *and* estimate the relationships between the latent variables, while bypassing the prediction of latent variable scores altogether. Standard full-information SEM methods, which cover a very broad set of models, are designed for use with individual scale items that are continuously measured, and yield parameter estimates that are very close to the true parameter values for suitably large samples (a large sample property referred to as *consistency*). In other words, SEM methods do not suffer from *finite item bias*, i.e., parameter estimates are consistent irrespective of the number of items defining an individual construct. Thus for continuously measured items, standard SEM methods provide an alternative to the two-step method that uses CTT-based scores directly in the regression. It should be noted that SEM methods do suffer from bias whenever the sample size is not large enough for the estimate to get sufficiently close to the true value. In this thesis, this will be referred to as *finite sample bias*.

While full-information SEM methods can be used to avoid finite item bias when individual scale items are continuously measured, or where continuous measurement can be approximated using an interval scale with sufficient measurement points, they cannot be used when items are measured using dichotomous or ordinal polytomous scales having relatively few categories. The approach of item parcelling, in which subsets of dichotomous items are combined to provide a smaller number of approximately continuous items, can be regarded as an application of continuous SEM methods to discrete data. This is a complex issue which is separate from the main topic of this thesis

and will not be considered further. The issues raised by item parcelling can be found in Kim & Hagtvet (2003) and Bandalos & Finney (2001) and the references as cited therein.

IRT-based scores for dichotomously or polytomously measured items can again be used in a two-step procedure, but at the cost of some, as yet unquantified, finite item bias. However, standard SEM methods have now been extended to handle discretely measured items, so it is natural to enquire if discrete SEM methods can be combined with IRT to avoid finite item bias, as in the continuous case. It turns out that this can be done, though the connection between IRT and SEM for discrete items appears not to be widely known. The key reference is Takane and de Leeuw (1987), who show that parameter estimation in the discrete SEM is equivalent to simultaneous estimation of IRT item parameters and the latent regression parameters. In other words, the IRT model is automatically embedded in the SEM model. We call this approach *simultaneous IRT-SEM* estimation, which yields consistent estimates of all parameters, that is, it avoids the problem of finite item bias even when the individual items are measured discretely (Lu, Thomas, & Zumbo, 2003; Lu, Thomas, & Orser, 2004). Another goal of the research will therefore be to see how well the theoretically consistent IRT-SEM estimates perform in practice, i.e., to examine the extent of the finite sample bias and to compare its magnitude to that of the finite item bias of the IRT (and CTT) based two-step methods.

Finally, this thesis will explore another possible approach that can be used whenever the parameters of an IRT scale are known a priori. This situation occurs if an IRT scale has been calibrated using a large normative sample and made available for use in subsequent research. In such a situation, the IRT parameters can be translated into discrete SEM parameters and treated as fixed in the IRT-SEM estimation. It is hoped that

this approach, referred to as *constrained IRT-SEM*, might perform better than the simultaneous IRT-SEM for small samples, since fewer parameters are involved in the estimation.

Another technique known as partial least squares (PLS) is sometimes used by business researchers to estimate the parameters of a structural model, e.g, Arnett, Laverie, & Meiers (2003), Kahai, Sosik, & Avolio (2004), Ravichandran & Rai (2000), Wixon & Watson (2001), and Birkinshaw, Morrison, & Hulland (1998). PLS parameter estimates, however, are not consistent as the PLS technique generates latent variable scores as an intermediate step and so suffers from finite item bias unless the number of items is very large (Chin, 1998; Barclay, Higgins, & Thompson, 1995; Schneeweiss, 1993), as is the case with the two-step method described earlier. Thus, another goal of the research will be to carry out a preliminary comparison of the PLS to the two-step and SEM-based approaches, in order to assess the extent of finite item bias in PLS under realistic model settings. Again, a more detailed examination of this question will be deferred to future research.

1.2 Research Goals

This study has several goals. The first goal is to provide researchers in business with a detailed overview of both the CTT and IRT approaches to measurement theory and scale construction, and to highlight some of the advantages and disadvantages of IRT methods. The second goal is to explore and explain in detail the problems encountered when scale scores (of any type) are subsequently used in statistical modeling and to identify various methods for overcoming these problems. The final and primary goal will be to quantify the impact of the problems identified, focusing in particular on simple and multiple latent variable regressions, and to empirically compare various strategies for bias reduction. More specifically, the following issues will be addressed in the empirical study.

1. How comparable and how accurate are the CTT- and IRT-based scores?
2. Does a two-step approach based on IRT scores yield less finite item bias than a two-step method based on CTT scores?
3. What is the difference in bias from using IRT- and CTT-based scores, versus a “consistent” method like discrete SEM (i.e., simultaneous IRT-SEM) which sidesteps the prediction of the latent variable scores?
4. Does the use of a constrained IRT-SEM analysis reduce the extent of finite sample bias compared to a simultaneous IRT-SEM approach?
5. How does the finite item bias of PLS compare with that of the two-step and IRT-SEM approaches? As noted, this aspect of the research will be preliminary.

Note that although this thesis will utilize structural equation modeling techniques, these will be invoked in a technical context as methods of bias reduction. This thesis is not oriented towards a philosophical discussion of the relative merits of different latent variable modeling techniques, which is a vast topic well outside the scope of the current work. The focus here is much narrower, being a consideration of the properties of various methods of estimating regression coefficients or the corresponding R^2 in a latent structural equation, when the indicators are discrete.

1.3 Organization of the thesis

Following this introduction, Chapters 2 and 3 will provide an overview of CTT and IRT. Chapter 4 describes methods for predicting IRT scores, and examines empirical and theoretical studies on the relative performance of the IRT- and CTT-based scores. This is followed in Chapter 5 by a discussion of the problems associated with using predicted latent variable scores in subsequent statistical analyses. The existing literature on the use of factor scores, CTT-based, and IRT-based scores will be examined and methods of correcting for and avoiding finite item bias will be briefly discussed. In Chapter 6, the continuous SEM and discrete IRT-SEM methods will be described. Chapters 2 through 6 essentially comprise a critical review of a series of methodological issues and techniques connected by their relevance to this investigation of IRT and its applications to latent regression. Specific research questions will be identified as the review proceeds, but these will then be assembled and presented as a connected set of research questions in Chapter 7. The research methodology to be used to empirically explore the research questions will then be described in detail in Chapter 8. Research results will be presented in Chapter 9. The primary research tool used in the empirical investigation is simulation, and Chapter 9 will present results from simulation studies. In Chapter 10, the various estimation methods will be applied to specific research data from an entrepreneurial study and from a large work-family conflict study. These applications will illustrate the finite item bias arising from using CTT-based scores and the PLS approach, and will illustrate the benefits of the IRT-SEM approach. Conclusions and recommendations will be provided in the last chapter.

Chapter 2

An Overview of Classical Test Theory (CTT)

The *classical test theory* (CTT) is the dominant paradigm for scale development in business research. The prevalent guidelines for scale construction in marketing and business research textbooks are based on CTT (e.g., Aaker, et al., 2004; Churchill & Iacobucci, 2002; Cooper & Schindler, 1998), a measurement technology that has its origin in the field of psychometrics. The recent edition of the *Handbook of Marketing Scales* by Bearden and Netemeyer (1999) describes numerous scales that have been constructed using CTT-based procedures. Examples are the Country Image scale by Martin & Eroglu (1993), the Purchasing Involvement scale by Slama & Tashchian (1985), the Organizational Commitment scale by Hunt, et al., (1985), the Salesperson Performance scale by Sujan, Weitz, & Kumar (1994), and the Unethical Behavior scale by Lagace, Ingram, & Boorum (1994).

This chapter will first provide preliminary definitions and terminologies to be used in the thesis. An overview of CTT and a discussion of its limitations will then be presented.

2.1 Preliminary Definitions and Terminologies

Latent Variables. As noted in the introduction, the unobservable underlying characteristics that will be the focus of this thesis are referred to by a variety of terms including: *latent variables*, *latent traits*, *constructs*, *latent constructs*, *factors*, and in the psychometric or education fields by the generic term *abilities*. However, given that this thesis is oriented towards researchers in the business sciences, the more neutral term “latent variables” will be adopted. Thus variables such as country image, brand loyalty, service quality, job satisfaction, managerial competence, etc. will be collectively referred to as latent variables in this thesis, and denoted by the Greek letter η (pronounced “eta”).

Tests or Scales. Latent variables can be measured using *tests* or *scales*. The term “tests” is commonly used in the psychometric or education fields; the term “scales” however is more common in the business sciences and will be adopted in this thesis. A scale consists of a set of items (or questions) whose responses are then aggregated in some way to obtain a composite score (Thomas & Zumbo, 2002). The items we can measure are thought to be related to, or representative of, the latent variable and are usually referred to as *indicators* of the latent variable or as *manifest variables*. Although a latent variable cannot be directly measured, it is assumed that it will take on a specific value under a particular set of conditions. The challenge of measurement theory is to determine how much of a latent trait a responding unit possesses. Please see Figure 2.1 for a conceptual measurement model.

Figure 2.1: Conceptual measurement model

The term *responding unit* (or *respondent*) requires some clarification. It is a generic term that can refer to a person, a store, or some other specific unit for which a measure of the latent variable is required. Also, it can refer to a unit in a sample drawn from a population of units, which is the typical case in survey research.

Dichotomous and Polytomous Scoring. Two major types of scoring can be obtained from items, depending not on what items are, but rather on how items are scored.

(1) *Dichotomous* item responses (also called *binary* item responses) are obtained from (a) items that are dichotomously scored according to a scoring key, for example an attitude scale with “true-false”, “agree-disagree”, or “favor-oppose” response sets, or from (b) a scale with multiple options that are subsequently scored as “correct-incorrect”,

or “endorsed-not endorsed.” For example, in multiple choice achievement scales, only one option is “correct” and the rest of the options are “incorrect”.

(2) *Polytomous* item responses contain more than two scoring options, and can be obtained from (a) *ordinal* (also called *graded response*, *Likert*, *Likert-type*) items, such as a five-point “strongly agree” to “strongly disagree” scale on an attitude measure, or from (b) *nominal* items, such as a three-option “brand A”, “brand B”, and “brand C” scale on a brand preference measure. In this thesis, polytomous item responses will always be of the ordinal type.

Item Score and Latent Variable Score. The term *item response* is often used interchangeably with *item score*. Either way, this refers to a respondent’s answer to one item. In an achievement scale, an item score may reflect whether an answer is “correct” or “incorrect”; in an attitude scale, an item score may reflect the degree of endorsement of a particular statement. A prediction of a respondent’s value on the latent variable of interest can then be obtained based on his/her item scores, and this prediction is referred to as *latent variable score*, and denoted by $\hat{\eta}$. The term “latent variable score” and “predicted value of the latent variable” will be used interchangeably in this thesis.

The use of the term prediction also merits a word of explanation. In this thesis, the term *prediction* is used to distinguish between the generation of scores on random latent variables and the *estimation* of fixed model parameters. In statistical terms, the procedure of assigning a number to a fixed but unknown constant such as a population mean, or a regression coefficient, is referred to *estimation*. However, a latent variable will be modeled as a random variable rather than as a fixed constant, so that the term

estimator is not appropriate. The term *prediction*, however, is commonly used in areas such as time series analysis when a value is to be assigned to a future observation (a random variable). The term prediction will therefore be adopted in this thesis and will describe the process of assigning latent variable scores to individual responding units. In this usage, prediction has no temporal association.

2.2 Introduction to Classical Test Theory

True Score Theory. Classical test theory (CTT) (also called *classical reliability theory*) is primarily concerned with estimating the *reliability* (formal definition given later in this section) of the observed scale score. That is, CTT is concerned with estimating the strength of the relationship between the observed score and the true score. CTT is also called *true score theory* because the fundamental idea of the theory can be expressed by a single equation:

$$\text{Observed score} = \text{True score} + \text{Error}. \quad (2.1)$$

This equation decomposes the *observed score*, i.e., the scale score we observe, into the sum of two components, the *true score* and an *error* (Kaplan & Saccuzzo, 1993; Peter, 1979; Suen, 1990; Traub, 1994; Wainer & Thissen, 2001). The usual example of an observed score is the sum or mean of a set of item scores. However, scale scores constructed in other ways (e.g., IRT scores) can also be considered as observed scores that satisfy Equation (2.1).

The thesis supposes, unless indicated otherwise, that in CTT an observed score (scale score) is formed either as a sum or as a mean of the item scores. For binary items, these are referred to as the *number-right score* and *proportion-right score*, respectively. For polytomous items, where the usual CTT assumption is that ordinal scales can be treated as interval scales, these are referred to as *total scores* or *mean scores*, respectively. It will be further assumed in this discussion that in CTT, all items are of identical importance in assessing the latent variables; that is, the scale scores are

unweighted sums or means. In addition, a standardized form of total scores is also commonly used, where the total scores are standardized to have a sample mean of zero and a sample standard deviation of one.

In reality, the *true score* can never be known and it is generally assumed to be the theoretical average of the observed scores we would obtain if we administered the scale to the same respondent under all possible conditions at different times using different possible items on the same content area. The responses of the respondent can be affected by various factors, such as personal and environmental conditions. In statistical terms, the true score is the expected value (i.e., average) of the observed scores, where the expectation is taken over the distribution of observed scores specific to that individual (Novick, 1966). The observed score is thus an unbiased estimate of the true score for that respondent.

The observed score from any single administration of a scale with a particular sample of items will differ from the true score. The deviation of the observed score from the true score is called *random error*, *measurement error*, or simply *error*, which, for good measurement, we would intuitively like to be as small as possible. By definition, these errors are random.

To summarize the foregoing discussion in mathematical terms, Equation (2.1) for a given respondent p , becomes:

$$x_p = t_p + e_p, \quad (2.2)$$

where x_p is the observed score for respondent p , t_p is respondent p 's true score, and e_p is the error for respondent p . These quantities have the following definitional properties: $E_p(x_p) = t_p$ and $E_p(e_p) = 0$, where $E_p(.)$ is the expected value, or average, of the quantity in parentheses, taken over the distribution of observed scores pertaining to subject p .

To this point, the fundamental equation was constructed by considering only one respondent. However, it can be extended to a population of respondents. The process of measuring the characteristic of a group of respondents can be thought to be: (1) the respondents who are measured are chosen at random from a population and (2) the observed score for a respondent is a random value from this respondent's own distribution since a respondent might give different responses on different occasions. If no repeated measurements are done, this is a conceptual distribution of scores that could be observed on this individual respondent. Then, the analogous equation to Equation 2.2 for a sample of respondents becomes:

$$x = t + e. \quad (2.3)$$

It is easy to show that the expected value of the error e is zero, and the correlation between t and e is also zero (see Appendix A); that is:

$$E(e) = 0 \text{ and } \rho(t, e) = 0, \quad (2.4)$$

where the expectation E is taken over the joint distribution formed by sampling respondents from the population and sampling individual scores from each selected

respondent. Application of Equation (2.4) yields the following results: the expected value of the observed score x and the true score t are equal:

$$E(x) = E(t), \quad (2.5)$$

and the observed score variance is the sum of the true score variance and the error variance, i.e.,:

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2. \quad (2.6)$$

In Equations (2.5) and (2.6), the required expectations are taken over the joint distribution of scores within individuals and individuals within the population.

The fundamental equations of CTT, namely Equations (2.3) through (2.6), are not assumed properties, but are constructions based on very elementary axioms, as described by Novick (1966). He states that “CTT is the simplest case of weak true score theory, by which we mean that collection of models that make no specific assumptions concerning the functional form of observed score, true score, or error score distributions (Novick, 1966, p.2)”. It is important to note that the true score (t), in fact, is not a “Platonic true score”, rather an expected value of a respondent’s observed score under repeated measures (Traub, 1994, p.22). The true latent variable value η , however, is the level of the construct of interest that the respondent really holds on the latent variable being measured and as will be seen later, this is not necessarily equal to the “true score” of CTT theory.

Reliability. Scale reliability is the strength of relationship between the observed score and the true score. (Note that reliability is defined for an entire scale, i.e., for an

entire test.) The reliability coefficient is defined as the ratio of true score variance to observed score variance, or equivalently as the squared correlation between observed score and true score:

$$\rho_{xx}^2 = \sigma_t^2 / \sigma_x^2. \quad (2.7)$$

Note that, in view of the statement following Equation (2.6), reliability is a population property of a scale. The true score variance and error score variance are nonnegative, and it is assumed that at least one of these variances is greater than zero, which implies that the observed score variance must be greater than zero. The reliability coefficient will be near its lower bound of zero, when the observed score variance is almost totally due to error variance and the true scores are nearly all the same. The coefficient will be near its upper bound of one when the observed score variance is almost completely due to differences among true scores, i.e., there is very little error variance (Dunbar, 1998; Peter, 1979; Traub, 1994). A scale with a reliability coefficient near one measures the true score very well, while a scale with a near zero coefficient measures the true score very poorly.

The definition of reliability is intuitively appealing; unfortunately the coefficient in Equation (2.7) cannot be directly computed from observed data because the true scores are unknown. However, it is possible to estimate the reliability coefficient when a particular set of assumptions are met. There are several methods for estimating reliability, but they are derived from definition (2.7) above. How one conceptualizes and operationalizes reliability, however, differs according to the computational method used.

One approach to estimating reliability involves carrying out measurement on respondents more than once, each time with a complete scale; either the same scale is administered, or a different scale is used each time. If the same scale is used, the resulting reliability estimate is called a *test-retest* coefficient of reliability or the *coefficient of stability*. If different scales are administered each time, the resulting reliability estimate is referred to as an *alternate-forms* coefficient of reliability or the *coefficient of equivalence*. The simplest way to obtain an estimate of reliability under this approach is to calculate the correlation between the scale scores obtained on the two occasions or to conduct an analysis of variance (ANOVA) of the scores obtained from two occasions. In the ANOVA analysis, the estimates of the variances of the true score and the error can be derived from the mean squares of a one-way ANOVA (for example, see Traub, 1994).

Another approach to estimating reliability involves recruiting a group of subjects and carrying out repeated measurements with just one scale. This approach requires that the scale be divided into distinct parts, where each part may be viewed as a separate measurement. Estimating the reliability coefficient from data provided by a scale administered only once is not straightforward, but it is easier when we have a scale administered more than once.

Let us first consider the case in which the scale is split into two parts, where ideally, these two parts are *parallel*. The two parts are considered parallel when the observed score for a respondent on both parts have the same expected true score and the same variance. An estimate of the reliability of either half-scale can be obtained from correlating the results of the two halves, an estimate that is referred to as *split-half*

reliability and is one form of the *coefficient of internal consistency* (Suen, 1990; Traub, 1994; Wainer & Thissen, 2002).

There are three problems with using split-half methods. First, the more items a scale has the more reliable it is, which means that split-half reliability is always an underestimation of the true reliability (Kaplan & Saccuzzo, 1993; Kline, 1993; Suen, 1990; Wainer & Thissen, 2001). Second, what we want is the reliability of the entire scale, not of the half scale. Third, different results may be obtained depending on how the items are split in half (Kaplan & Saccuzzo, 1993; Kline, 1993; Peter, 1979; Tull & Hawkins, 1993). Thus, the researcher is faced with the question of what is the ‘real’ reliability coefficient. Let us look at some traditional solutions to these problems. Alternative approaches provided by generalizability theory (see Marcoulides, 1998, 1999) will not be pursued in the thesis.

The solution to the first two problems is to use the Spearman-Brown prophecy formula to correct the underestimation in the correlation between scores on the half-scales (Kaplan & Saccuzzo, 1993; Suen, 1990; Traub, 1994; Wainer & Thissen, 2001). In the situation where the parts are not parallel, Rulon’s formula can be used to provide a lower bound to the reliability coefficient (Traub, 1994).

In the measurement literature, the solution to all three problems is assumed to be provided by *Cronbach’s alpha* (also called *Coefficient alpha*) which computes the mean for all possible split-half coefficients, when the relevant variables are continuous, so that a true score model of the form (2.3) applies. Its formula is expressed as follows:

$$\alpha = \frac{n}{(n-1)} \left[1 - \frac{\sum_{i=1}^n \sigma_{x_i}^2}{\sigma_x^2} \right], \quad (2.8)$$

where n is the number of items in the scale; $\sigma_{x_i}^2$ is the variance of the observed item scores for the i -th item, and σ_x^2 is the variance of the observed scale scores. Coefficient alpha provides a lower bound to reliability (Kaplan & Saccuzzo, 1993; Kline, 1993; Suen, 1990; Traub, 1994, Wainer & Thissen, 2001) for scales with unequally weighted items (McDonald, 1999). If alpha is very high, we can assume that the reliability of the scale is acceptable. This means that the lower boundary of reliability is still high, i.e., the reliability will not drop below alpha. On the other hand, if alpha is low, it is not very informative because the actual reliability may be low or high. Coefficient alpha is by far the most commonly used reliability estimate of a multi-item scale in marketing and other social science research, as seen from *Handbook of Marketing Scales* by Bearden and Netemeyer (1999).

True Score and Latent Variable. Note that CTT does in fact apply irrespective of the distribution of observed scores. In particular, it applies to discrete items. Lord (1980) discusses the relationship of classical test theory to a simple one-factor latent variable model that underlies the discrete observed data, a model that will be extensively used later in this thesis. In this section, the relationship of this model to CTT will be described. In particular, the “true score” for a scale based on binary indicators will be derived, and it will be seen that for binary data, the true score is not same as the underlying latent variable in the one-factor model. This result holds for polytomous data

also, but the relationships are most easily illustrated in the binary case. As a corollary, it will be seen that when the items are continuously measured, the true score as defined by Novick (1966) is precisely equivalent to the latent variable.

The one-factor model for a binary indicator can be expressed as:

$$x^* = \lambda\eta + \varepsilon, \quad (2.9)$$

where η is the latent variable and ε is an uncorrelated respondent-specific error with variance θ^2 . We do not observe x^* , a continuous indicator of η ; instead, we observe the binary variable x . The observable discrete indicator variable for respondent p is then modeled as

$$\begin{aligned} x_p &= 1 \text{ if } x^* \geq \tau, \\ x_p &= 0 \text{ otherwise.} \end{aligned} \quad (2.10)$$

The key connection with Novick (1966) is that the distribution of observed scores specific to an individual is determined by the distribution of the specific error ε , and that the focus on a specific individual, p , is equivalent to fixing the value of η . Thus Novick's "true score" for an observed variable x_p is given by

$$t_p = E(x_p \mid \eta). \quad (2.11)$$

It can be shown that (Lord, 1980)

$$t_p(\eta) = E(x_p \mid \eta) = \Phi[(\lambda\eta - \tau)/\theta], \quad (2.12)$$

where Φ is the cumulative standard normal density. It is then clear that the true score is not equal to the latent variable, i.e., $t \neq \eta$. In fact, the true score $t_p(\eta)$ is measured on a scale that goes from zero to one, and is a non-linear (ogive) function of the latent variable η . The significance of Equation (2.12) to the test characteristic functions in item response theory will become evident later. It follows from the above discussion that, without the discretization process (i.e., in continuous case), $t_p(\eta) = E(x^* | \eta) = \lambda\eta$, i.e., the true score is a linear function of the latent variable.

Reliability and Discrete Items. An important consequence of the relationship (Equation 2.12) of true score (t) to the latent variable (η) for discrete variables is that Cronbach's alpha will not be the most appropriate measure of "reliability" for assessing the various latent variable prediction methods to be explored in this thesis. In fact, for continuous items, Cronbach's alpha measures $\rho^2(x, \eta)$, where $t = \lambda\eta$, while for discrete items, it measures $\rho^2(x, t(\eta))$, where $t = \Phi[(\lambda\eta - \tau)/\theta] \neq \eta$.

The quantity $\rho^2(x, \eta)$ indicates how much variance the observed scores and the "true" latent variable share in common. This measure could be considered as alternative reliability index. However, to avoid confusion it will be referred to only as "*accuracy*" or "*shared variance*" in this thesis. Though such a measure of "reliability" as scale accuracy has previously been used in the literature (e.g., MacDonald and Paunonen, 2002), the relationship of this measure to Cronbach's alpha appears not to have been previously noted. It will be seen in Chapter 9 and Appendix B that Cronbach's alpha is

not a lower bound for accuracy; on the contrary; it may provide an optimistic estimate of accuracy, i.e., Cronbach's $\alpha > \rho^2(x, \eta)$, when the scale items are discrete.

Standard Error of Measurement. An important consideration of any measurement we perform is the size of the measurement error. The smaller the error, the more precise we consider the measurement to be. We cannot know the size of the error in a particular measurement, but we can estimate it. Recall that reliability is the proportion of observed variance that is actually true score variance. When reliability is known (and given the assumption of independence), it is possible to estimate error variance as follows:

$$\text{Error variance} = \text{Observed variance} \times (1 - \text{Reliability}). \quad (2.13)$$

The square root of the error variance is the *standard error of measurement* which can be used to establish a confidence interval for the true score. This confidence interval is an indicator of the size of measurement errors, on average, for the observed score under study (Suen, 1990; Traub, 1994).

Item Difficulty and Item Discrimination. CTT can provide us with information to guide selection from a pool of items so that the reliability of the total score can be enhanced. The reliability of the scale score is related to characteristics of the items, namely, the item *difficulty index*, and the *discrimination index*.

An item's *difficulty index* is the mean value or expected value of its observed score. In aptitude and achievement measurements, it is estimated by the proportion of respondents who answered the item correctly; this proportion is referred to as the *item*

mean, *item difficulty*, or *item p-value*, where a value near one indicates an easy item and a value near zero indicates a difficult item (MacDonald & Paunonen, 2002, McDonald, 1999; Suen, 1990; Traub, 1994). Although the concept of item difficulty is primarily used in aptitude and achievement measurements, it can also be applied to measures that do not assume right or wrong answers. Suppose we use a 5-point Likert scale, with options from strongly disagree to strongly agree. The average of all responses to that item across all respondents can be seen as an item difficulty index in the CTT framework (Walsh & Betz, 2000). To maximize reliability of a scale score, it is recommended that the items chosen for a scale should be of medium difficulty, but note that this is not sufficient to ensure that scores on the scale will be highly reliable (Traub, 1994).

One purpose of measurement is to differentiate among respondents with different true scores. A good item should be able to discriminate those with high true scores from those with low true scores. For example, in aptitude and achievement scales, an item is considered effective if those with high true scores tend to answer it correctly and those with low true scores tend to answer it incorrectly. The reliability of the total score is maximized when all items in the scale have high discrimination power. The *discrimination index* for an item is defined to be the correlation between the observed score for the item and the total scale score; this correlation is also referred to as the *item-total coefficient of correlation*. This index reflects the extent to which the item score (for all respondents) places respondents in the same relative order as does the total scale score, that is, it reflects the ability of an item to differentiate respondents of different levels of true score in a way that is congruent with the way respondents are differentiated by their total scale scores. When the item score is a part of the total score, however, the

item-total correlation is inflated. An alternative way to calculate item discrimination (i.e., the item-total correlation), therefore, is to compute total score excluding the item score on a given item being analyzed to correct for the inflation.

A large item discrimination value indicates that the item effectively differentiates between respondents with high and low true scores, whereas a near-zero or negative item discrimination value indicates poor differentiation (MacDonald & Paunonen, 2002, Suen, 1990; Traub, 1994).

2.3 Limitations of Classical Test Theory

Limitations of CTT have been noted by Lord (1953) and more recently by several other researchers (Balasubramanian & Kamakura, 1989; Embretson, 1996, 1999; Hambleton, 1989; Hambleton, et al., 1991; Nunnally & Bernstein, 1994). Three main lines of criticism can be identified in the literature.

The first line of criticism points out that respondent and item characteristics are not distinguished in CTT. The observed score for a respondent is dependent on the particular scale used, where we say the scale scores are *item- or scale-dependent*. To make this idea concrete, we consider the case of achievement or aptitude measurements. When the scale is difficult, the respondent will appear to have low true score; when the scale is easy, the respondent will appear to have high true score. That is, respondents' predicted true scores change as the scale context changes. However, whether a respondent is evaluated to have low or high true score should not be dependent on the scale used. On the other hand, item or scale characteristics are dependent on a particular group of the respondents measured, where we say the item or scale characteristics are *respondent-dependent*. When the respondents with high true scores are measured, the scale will appear to be relatively easy. In other words, item or scale characteristics change as the respondents change. However, whether an item or scale is judged to be difficult or easy should not depend on the true scores of the respondents.

A consequence (i.e., problem) of scale-dependent scores is that it is difficult to compare respondents who are measured by using different scales. Respondent-dependent

item and scale indices are of limited use when scales are constructed from one population of respondents but will subsequently be used by a dissimilar population of respondents. Moreover, item and scale indices cannot be compared when they are obtained from different populations of respondents.

The second line of criticism focuses on the standard error of measurement. The standard error of measurement is a function of scale reliability and variance, as defined in Equation (2.13), so that the same standard error is assumed for all respondents because there is only one reliability coefficient. In other words, scale scores are assumed to be equally precise measures for respondents with different true scores. This is an unrealistic assumption because the standard error of measurement should be different for respondents with different true scores. To better understand this argument, consider two extreme cases in the context of an achievement test: (a) a respondent received a scale score of zero by responding incorrectly to all the items and (b) a respondent obtained a full scale score by answering all the items correctly. These scores provide no information about how low or how high their true scores are. On the other hand, when a respondent get some items right and some items wrong, the observed scale score then contains more information about this individual's true score. Therefore, it is expected that the standard error of measurement should be greater for the respondents at the two ends of the true latent variable continuum but lower in the middle.

The final criticism of CTT is that it is *scale-oriented* rather than *item-oriented*. The classical true score model provides no consideration of how respondents respond to a given item, and hence no basis exists for determining how a particular respondent (or sub-group of respondents) might respond to a certain item (Hambleton et al., 1991).

2.4 Summary

Classical test theory (CTT) is primarily concerned with estimating the reliability of the observed scale score, defined as the ratio of the true score variance to observed score variance. There are various methods for computing reliability, which have different utility in different situations. A summary of the various methods is provided in Table 2.1.

CTT maintains a strong influence among business and other behavioural researches. Nevertheless, its methods of item analysis and true score prediction are perceived to be limited by many researchers in the fields. Arguably the most telling criticism of CTT is that respondent statistics (i.e., latent variable predictors) and item statistics (i.e., item difficulty and item discrimination) cannot be completely separated, a shortcoming that ultimately limits the utility of the respondent and item statistics in practical scale construction and application.

The following chapter describes the technique of item response theory (IRT) which was designed to address the various perceived shortcomings of the classical CTT approach.

Table 2.1: Reliability estimation related to different design of reliability experiments

Design	Estimation of Reliability
Type I: More Than One Administration	
Test-retest	<ul style="list-style-type: none"> - Correlation coefficient of the two scales - An analysis of variance
Alternate forms	<ul style="list-style-type: none"> - Correlation coefficient of the two scales - An analysis of variance
Type II: Single Administration	
Parallel forms	Spearman-Brown Formula
Non-parallel forms	Rulon's formula
Single form	Coefficient Alpha

Chapter 3

An Overview of Item Response Theory (IRT)

Item response theory (IRT) is derived from what is called “modern test theory” and is one of the methodologies that have resulted from what Embretson (1999) and Embretson & Reise (2000) call “the new rules of measurement.” IRT has been also called "latent trait theory" or "item response curve (ICC)" theory. Compared to CTT-based methods of test development and analysis, IRT offers many advantages, even though most of the scales used by marketing and organization behaviour research have been developed using CTT. This chapter will provide an overview of IRT as well as a discussion of some of its advantages over CTT and its limitations. Illustrations of applications of dichotomous and polytomous IRT in a business context will also be provided. Last, the applications of IRT in the business literature will be reviewed.

3.1 Basic Concepts of IRT

IRT is built on mathematical models that relate observable item responses to a latent variable. This latent variable is a hypothetical variable that takes values on a scale that is continuous, with equal intervals, and is free from measurement error (Osgood, et al., 2002). The latent variable is assumed to underlie the observed responses and is generically referred to in the education and psychometric literature as *ability*. As noted earlier, we shall retain the more neutral term *latent variable* in our discussion. Each item can be described using one or more of three properties: item difficulty, item discrimination, and the probability of a correct response by guessing (details later). IRT incorporates these individual item properties into the response model as parameters, which are then used to guide scale construction.

An IRT model specifies the probability that a respondent will provide a particular response to a particular item as a function of that respondent's level on the underlying latent variable in question, with the aforementioned properties as model parameters. The model allows a knowledge of these parameters to be combined with a knowledge of a respondent's item responses to *predict* the level of the latent variable for that respondent.

To illustrate the structure of an IRT model, consider a set of n dichotomous scale items, administered to a sample of N respondents, with the aim of measuring an underlying latent variable or construct. Let us denote the respondent's latent variable value by η_i , and the respondent's measured item score (or response) as x_{ij} where

$$x_{ij} = \begin{cases} 1 & \text{when respondent } i \text{ answers item } j \text{ positively} \\ 0 & \text{when respondent } i \text{ answers item } j \text{ negatively} \end{cases}.$$

The probability that respondent i will respond positively to item j is denoted as $P(x_{ij} = 1 | \eta_i)$ or $P_j(\eta_i)$, where $P_j(\eta_i)$ contains up to three parameters representing difficulty, discrimination, and propensity to guess for each scale item j . Strictly speaking the response probability $P_j(\eta_i)$ should be denoted $P_j(\eta_i; a_j, b_j, c_j)$ where a_j , b_j , and c_j denote difficulty, discrimination, and “guessing” parameters for item j . However, these parameters will be suppressed to simplify the notation. Note also that the terms “positively” and “negatively” are used in a generic sense. Answering an item positively simply means endorsing a particular item category, while answering negatively means failure to endorse.

The basic idea of IRT is that if the relationship between the latent variable η_i and the response probability $P_j(\eta_i)$ is known for each item in the scale (i.e., all the item parameters are known), then the value of the respondent’s latent variable can be predicted, along with the associated measurement error (Suen, 1990).

IRT Model Assumptions

The commonly used IRT models involve three basic assumptions, namely *monotonicity* (M), *local independence* (LI), and *unidimensionality* (UD). Assumption M is met when a respondent’s probability of responding positively to an item increases as the respondent’s latent variable value increases. Assumption UD is met if all the items in a scale are measuring a unidimensional (as opposed to multidimensional) construct, i.e.,

if a construct of interest can be represented by a single latent variable (Nandakumar & Stout, 1993). Assumption LI is met if for a fixed level of the latent variable, the responses to different items are independent of each other. These basic assumptions are discussed in greater detail below.

Unidimensionality. It is noteworthy that in a strict sense the issue of dimensionality refers to the $N \times n$ data matrix consisting of the responses of each of the N respondents to the n scale items. Dimensionality refers only to that matrix, not to the items by themselves, not to the scale, nor to the respondents. Thus, a set of items (i.e., a scale) cannot be described as unidimensional, only the item response data matrix can be deemed unidimensional, or not unidimensional (Nandakumar & Stout, 1993; Reckase, 1994). To illustrate this notion, consider a single response to a scale consisting of two items. This can be represented as a point in a two-dimensional (x-y) graph. Two responses can be plotted as two points on the graph, and these two points can be connected by a straight line. Only when three responses to the two items are plotted is it possible to determine whether one or two-dimensions are needed. The three points could all be on the same line, representing a unidimensional scale, or they might not, in which case both dimensions of the two-dimension graph are needed to represent a two-dimensional scale.

That a set of answered items in a scale all measure just one thing in common is a very critical and fundamental assumption in measurement theory (Hattie, 1984). Besides being the basis of most mathematical measurement models, there are at least two reasons why it is essential that a scale be unidimensional.

First, it is vital that a scale that claims to measure the level of a certain latent variable or construct is in fact not contaminated by varying levels of one or more other latent variables or constructs pertaining to the respondent (Stout, 1987). A common example given in the educational measurement context is that a test of mathematical ability should not be contaminated by varying levels of verbal ability of the examinees. In the organizational behaviour context, one would not want a scale measuring work to family interference to be contaminated by varying levels of family to work interference.

Second, it is difficult to make any claims for using a total score if the scale has items presumably measuring different constructs (Hattie, 1984; Hattie, Krakowski, Rogers, & Swaminathan, 1996). For example, suppose a scale of service quality consists of two general sorts of items, one concerned with facility and equipment and the other with speed of service issues. Two individuals could thus arrive at the same scale score even with very different opinions on the facility and the speed of service; the meaning of a scale score here then is ambiguous. From this viewpoint, even if a two dimensional set of items is appropriate to use in a scaling situation, the items should be analyzed and interpreted as two scales (Stout, 1987).

Local Independence. LI means that when the latent variable influencing a respondent's responses on a scale is held constant, then a respondent's responses to any pair of items are statistically independent. In other words, after taking the respondent's latent variable into account, no relationship is thought to exist between the respondent's responses to different items. In other words, the latent variables specified in the model are the only factors believed to be influencing a respondent's responses to scale items (Hambleton, et al., 1991).

When LI is met, it implies that for a given respondent (or for all respondents at a given level of the latent variable) the probability of a response pattern on a set of items, denoted $P(x_1, x_2, \dots, x_n | \eta_i)$, is simply equal to the product of the probabilities of the respondent's responses to each individual item. For n items, local independence then can be stated mathematically as:

$$P(x_1, x_2, \dots, x_n | \eta_i) = \prod_{j=1}^n P(x_j | \eta_i). \quad (3.1)$$

Although the UD assumption is hard to meet, Stout (1990) argues that UD and LI can be replaced by the weaker and more realistic assumptions of *essential unidimensionality* and *essential independence*, respectively. Theoretically, it allows for the presence of minor nuisance dimensions. Junker (1988) has proved mathematically the connection between LI and essential independence and has also derived statistical consistency results for maximum likelihood estimates of a latent variable under the assumption of essential unidimensionality (Nandakumar, 1991). Nevertheless, IRT researchers have also developed models to account for multidimensional constructs (see, for example, MacDonald, 1999; Reckase, 1997). This thesis, however, will be concerned solely with unidimensional constructs / latent variables.

3.2 IRT Model Formulations

Based on the previously described assumptions, several popular IRT models have been defined, and these models can be distinguished in several ways. Sijtsma & Hemker (2000) and Thissen & Steinberg (1987) offered a detailed taxonomy of IRT models, but here we consider only two aspects for distinguishing IRT models. The first aspect is the number of item parameters used to capture the characteristics of an item. The second aspect is concerned with whether item responses are scored dichotomously or polytomously.

This section will first introduce the three most widely used IRT models for dichotomous responses. These three models differ with respect to the number of item parameters. Dichotomous IRT models provide the foundation for understanding more complex models, such as polytomous response models, of which only the most commonly used, Samejima's (1969) graded response model, will be introduced. The methods for item parameter estimation will also be briefly described.

3.2.1 Models for Dichotomous Responses

An important aspect of an IRT model is the relationship between the value of the latent variable and the probability of a positive response to an item. This relationship can be modelled by a nonlinear regression, which expresses the probability $P_j(\eta_i)$ that a respondent will endorse an item, as a function of the underlying latent variable. This fitted regression line is referred to as the *item characteristic curve* (ICC). In the regression, $P_j(\eta_i)$ can be modeled using any kind of link function, the most common being the *logit* link which uses the log-odds transformation, and the *probit* link (models with this link are commonly called *normal ogive* models) which employs the standard normal cumulative distribution function (Thomas & Zumbo, 2002). Note that the link is the function that relates the probability of $P_j(\eta_i)$ to the linear predictor based on η_i , as illustrated in what follows. The number of item parameters modeled gives rise to the name of the model, e.g., the 1-, 2-, and 3-parameter models (Lord, 1952, 1980).

Normal Ogive Models. Figure 3.1 shows an ICC for the 3-parameter normal ogive model. ICC's are typically graphed over the domain of ± 3 standard deviations on the latent variable continuum. In mathematical form, the 3-parameter model with probit link is given by

$$P_j(\eta) = c_j + (1 - c_j)\Phi[a_j(\eta - b_j)] \quad (3.2)$$

for $j = 1, 2, \dots, n$, where

$$P_j(\eta) = \begin{array}{l} \text{probability that a randomly chosen respondent with latent variable} \\ \text{value } \eta \text{ answers item } j \text{ positively} \end{array}$$

- $\Phi(\cdot)$ = normal cumulative distribution function
- b_j = difficulty parameter for item j
- a_j = discrimination parameter for item j
- c_j = guessing or pseudo-chance level parameter for item j
- n = number of items in the test.

Note that $\Phi(x)$, for any real number x , denotes the probability that a random variable X drawn from a standard normal distribution (i.e., a normal distribution with mean 0 and variance 1) will not exceed x , i.e., $X \leq x$. Thus, $\Phi(x)$ tends to one as x tends to infinity and $\Phi(x)$ tends to zero as x tends to negative infinity, yielding the familiar s-shape ogive function shown in Figure 3.1. A simpler normal ogive model can be obtained by setting $c_j = 0$ (leading to a 2-parameter model) or by setting $c_j = 0$ and $a_j = 1$ (leading to a 1-parameter model). The latent variable η is theoretically defined to be in the range $(-\infty, \infty)$. In many IRT software programs it is scaled to have a mean of zero and a standard deviation of one. Parameter c in the model is the height of the lower asymptote of the ICC and in an achievement testing context represents the probability of respondents with low ability answering the item correctly. In the business research context, the parameter c does not have such an intuitive meaning. Thus, the work described in this thesis will focus on models for which parameter c is set as zero, i.e., the 2-parameter model. Parameter b is the point on the latent variable scale where a respondent has $(1 + c)/2$ probability of a positive response (Baker, 2001; Hambleton & Jones, 1993). Thus, in the 2-parameter model, the difficulty or affectivity parameter, b , is used to generically refer to the point on the latent variable scale where respondent has a

50-50 chance of a positive response to the item. The discrimination or sensitivity parameter, a , is proportional to the slope of the ICC at point b of the latent variable scale (see Figure 3.1). Fraley, Brennan, & Waller (2000) report that in the assessment of personality, attitudes, and interpersonal behaviour, item discriminations often fall between 0.5 and 2.5.

Logistic Models. The next models to be discussed are the well-known 1-parameter logistic models, also referred to as the *Rasch model* (Rasch, 1960), and the 2-, and 3-parameter logistic models (Birnbaum, 1968). These three models have been more widely used than the normal ogive model due to their mathematically more convenient properties. The ICC of the normal ogive models represents an incomplete integral of the normal density function, which can only be numerically evaluated. On the other hand, the ICC of the logistic model (corresponding to the cumulative distribution function of the logistic density) can be expressed in closed form, as in equation (3.3). Although mathematically different, the normal ogive ICC's and the logistic ICC's almost have the same shape. The ICC for the 3-parameter logistic model takes the following form:

$$P_j(\eta) = c_j + (1 - c_j) \frac{e^{Da_j(\eta - b_j)}}{1 + e^{Da_j(\eta - b_j)}}, \quad (3.3)$$

for $j = 1, 2, \dots, n$, where $P_j(\eta)$, a_j , b_j , and c_j are defined as for the 3-parameter ogive model (3.2). The additional parameter in this model is a scaling factor. D is often set as 1.7 to make the logistic ICC as close as possible to the normal ogive ICC and the correspondence is very close in most cases; this is an analogy to the similarity between logistic regression and probit regression for binary regression analysis.

The choice among these three models (with 1-, 2-, or 3-parameters) depends on the item parameters needed to describe the response, although it is also possible to take a data-driven approach since one can compare the fit of each model to a given set of data. If the items are expected to vary in difficulty but to have similar discriminating power, the 1-parameter model is preferred. On the other hand, if items are also expected to vary in their discriminating power, then a 2-parameter model is appropriate. In cases where respondents with low values of the latent variable might nevertheless respond positively, then a 3-parameter should be considered (Harwell & Gatti, 2001). As mentioned, this thesis will focus only on 2-parameter models.

Figure 3.1: A 3-parameter normal ogive item characteristic curve

3.2.2 Illustration of a Dichotomous IRT Model in a Business Context

Suppose we are interested in measuring the latent variable or construct "consumer self-actualization" (Brooker, 1975), using a scale containing multiple items. Two of the items might be:

Item 1: *When I am shopping for myself, I make decisions without help from others (1) vs. I often look to people I know for help when buying something for myself (0).*

Item 2: *It's true, money can't buy happiness (1) vs. Money may not be everything, but it's got a big lead over whatever is second (0).*

As an example, Figure 3.2 shows two possible ICCs for the 2-parameter model of these two items. The latent variable "consumer self-actualization" is represented on the abscissa (x-axis) and the probability of responding to an item positively (answering "1") is on the ordinate (y-axis). Notice that for each item the probability of a consumer answering "1" increases as his/her self-actualization increases.

An ICC has two important features: its location relative to the latent variable continuum and its slope. The first feature reflects the item's difficulty. The difficulty of an item is defined as the point on the continuum where a respondent has a 50-50 chance of responding positively. In this example, Figure 3.2A shows that the difficulty parameter of item 1 is the level of self-actualization where a respondent is required to have for a 50-50 chance of answering "1", shown as the letter b_1 ; the difficulty parameter for item 2 is b_2 . It is also shown that item 2 requires a higher level of self-actualization to

achieve a 50% chance of endorsing “1”. Given a particular level of self-actualization, say -1, the consumer has higher probability of answering “1” to item 1 than item 2.

The second important feature of an ICC is item discrimination. It is the rate at which the probability of responding positively to an item changes with latent variable level, i.e., the slope of an ICC. The item discrimination parameter indicates how well an item can differentiate those with a higher level of latent variable value from those with a lower level. The steeper the ICC, the more sharply the item differentiates between consumers at different levels of self-actualization. Figure 3.2B shows that two items have the same difficulty but item 1 has greater discrimination than item 2, especially at the range of self-actualization from -2 to 1.

Figure 3.2: Illustration of a dichotomous IRT in business context

A. Two ICC's for two items with different difficulty, same discrimination

B. Two ICC's for two items with same difficulty, different discrimination

3.2.3 Models for Polytomous Responses

In the business sciences in particular, scales with polytomous items are often preferred over those with dichotomous items for the following reasons: (a) polytomous indicators contain more information than dichotomous indicators, resulting in reduced standard

errors in latent variable prediction; in other words, fewer polytomous items are typically needed to obtain the same degree of reliability; (b) some latent variables are more easily measured on polytomous scales (Van Der Ark, 2001; Zickar, 2002).

Several polytomous item response models have been developed as extensions of dichotomous models, each having special features for different types of data. If the data are ordered categories, i.e., a Likert scale, a graded response model (Samejima, 1969) or a Rasch family of polytomous item response models – including the partial credit model (Masters, 1982) and the rating-scale model (Andrich, 1978) – are appropriate. If the responses are scored in two or more nominal categories, then Bock's (1972) nominal response model is applicable.

Of these polytomous IRT models, the author will limit her attention to the most flexible and best known, which is Samejima's (1969) graded response model (GRM), as:

- (a) it is very useful for any ordered categorical data common for attitude, perception, and social / behaviour scales used in marketing and organizational behaviour research, and
- (b) it is equivalent to a factor analytic model of ordered categorical data (an important property that will be used in Chapter 6).

The GRM is an extension of the dichotomous 2-parameter normal ogive and logistic models described in the previous section. In addition to the assumptions previously discussed, the GRM further assumes ordered response categories, such as a 5-point Likert scale, or scales for levels of accomplishment or partial credit.

3.2.4 Illustration of a Polytomous IRT Model in a Business Context

We will start with an example to illustrate the GRM, followed by an introduction to its technical aspects. Suppose we are interested in measuring “buying impulsiveness” (Rook and Fisher, 1995), using a 5-point Likert scale with multiple items. One of the indicator statements is: *I often buy things spontaneously*. The five possible response options (from “strongly disagree” to “strongly agree”) to this indicator statement can be presented on a continuum of buying impulsiveness with four thresholds. A value on the continuum that falls between the first and second thresholds, for example, will correspond to a response in the second of the Likert categories. The values of the thresholds for each variable can be estimated from the data as part of the GRM estimation process.

One of the important goals of fitting a GRM is to determine the probability that a particular option will be endorsed by a randomly selected respondent. The model for representing this probability is referred to as *item category response function* (ICRF), and can be constructed by decomposing the ordinal responses into a series of binary comparisons, i.e., 0 vs. 1, 2, 3, 4; 0, 1 vs. 2, 3, 4; 0, 1, 2 vs. 3, 4; 0, 1, 2, 3 vs. 4, in this example (see, for example, Embretson & Reise, 2000). In other words, the five categories can be reduced to a set of four binary comparisons, each of which can be modelled by an ICC as described in the previous section. Each ICC can then be interpreted as the probability of a respondent's response falling in or above a given category. Figure 3.3A presents four hypothetical ICCs for this 5-point Likert scale. ICC_1 is the probability of choosing "1: disagree" or a higher category over the "0: strongly

disagree"; ICC_2 is the probability of answering "2: uncertain" or a higher category; ICC_3 is the probability of choosing "3: agree" or higher; and ICC_4 is then the probability of selecting "4: strongly agree". Note that the probability of choosing "0: strongly disagree" or higher is one. Thus, an ICC representing a lower level of buying impulsiveness will be to the left of an ICC representing a higher level.

In keeping with this present example, for a scale with five categories we need to estimate four thresholds. The threshold b_3 represents the point on the buying impulsiveness continuum where a respondent has a 50% chance of selecting "3: agree" or "4: strongly agree". In other words, the threshold b_3 indicates that level of buying impulsiveness possessed by a respondent who is likely to choose "3: agree" or "4: strongly agree" with a 50-50 chance.

Once the ICCs are known, the ICRF can be obtained:

$$\begin{aligned}
 \text{Probability [choosing "0: strongly disagree"]} &= 1.0 - ICC_1 \\
 \text{Probability [choosing "1: disagree"]} &= ICC_1 - ICC_2 \\
 \text{Probability [choosing "2: uncertain"]} &= ICC_2 - ICC_3 \\
 \text{Probability [choosing "3: agree"]} &= ICC_3 - ICC_4 \\
 \text{Probability [choosing "4: strongly agree"]} &= ICC_4 - 0.
 \end{aligned}$$

Each ICRF represents the probability of choosing a particular category ("0: strongly disagree" to "4: strongly agree") given a respondent's level of buying impulsiveness. Figure 3.3B shows that the less impulsive respondents at the lower end of the buying impulsiveness continuum have a higher probability of selecting the option "0: strongly disagree", and the strongly impulsive respondents at the high end of the continuum are most likely to endorse "4: strongly agree". It is important to note that the sum of the item category response probabilities is one for any fixed level of buying impulsiveness.

Readers who are only interested in the GRM from a practical point of view can stop reading and proceed to Section 3.3. The remainder of this section will describe the GRM in more mathematical terms.

Figure 3.3: Illustration of a polytomous IRT in business context

A. Item characteristic curves (ICCs) for five ordinal categories

B. Item category response functions (ICRFs) for five ordinal categories

The Mathematical Formulation of the GRM

In the GRM, a response to an item can be categorized into one of $(m + 1)$ ordered categories. Let x_j be the response for a randomly selected respondent on item j , and $x_j = 0, 1, \dots, m$. For example, for a five category scale, a response in the first category will be denoted by $x_j = 0$, the second by $x_j = 1$, and so on.

A suitable model can be constructed first considering a set of m ICCs, where each 2-parameter ICC represents the probability of a respondent endorsing a particular option, or higher. As described previously, the first ICC will treat the first category as negative endorsement and any of other categories as positive; the second ICC will treat the first two categories as negative endorsement and any other categories as positive, and so on. In general then, under an ogive model, each of these ICCs takes the form:

$$P_{x_j}(\eta) = \Phi[a_j(\eta - b_{x_j})]. \quad (3.4)$$

Under a logistic model, we would then have

$$P_{x_j}(\eta) = \frac{e^{Da_j(\eta - b_{x_j})}}{1 + e^{Da_j(\eta - b_{x_j})}}. \quad (3.5)$$

Note that b_{x_j} is the category “difficulty” representing a between-category “threshold” parameter which specifies the point on the latent variable continuum where there is a 50% chance of endorsing category x_j or higher. Other parameters in the model are defined the same way as before. With $(m + 1)$ categories, m thresholds and one item discrimination parameter need to be estimated for each item.

Once the 2-parameter ICC for each category is known, the probability of a respondent endorsing category x_j , then can be obtained by

$$P_{x_j}^*(\eta) = P_{x_j}(\eta) - P_{(x+1)_j}(\eta), \quad (3.6)$$

where $P_{x_j}(\eta)$ is the probability of selecting category x or higher and $P_{(x+1)_j}(\eta)$ is the probability of endorsing the next highest category or higher. We call these $P_{x_j}^*(\eta)$ *item category response functions* (ICRFs), and they represent the probability of a respondent responding in a particular category conditional on the latent variable level.

3.3 IRT Item Parameter Estimation

In the psychometric testing context, it is typical that a subject is administered by a number of test items, and the item responses are subsequently transformed into an ability or latent variable score for that individual, using a previously calibrated IRT model. In the survey research context, the test items are administered to all survey respondents and predictions of individual latent variable levels and IRT model parameters are obtained from the same survey dataset. In both contexts, the key feature that separates IRT from CTT methods is a model that links the characteristics of a given scale item and the true value of a respondent's latent variable level, to the probability that the respondent will respond positively to that scale item. Prediction of individual latent variable values will be discussed in the next chapter, under the assumption that the IRT parameters are known, or can be treated as fixed, having previously been estimated. The actual parameter estimation process will be reviewed briefly in this section. This estimation process and the latent variable prediction process to be described in Chapter 4 both rely heavily on the principle of likelihood. Since this may not be familiar to all readers, it will be described briefly prior to the description of IRT parameter estimation.

Likelihood-Based Estimation. To many applied researchers, the notion of parameter estimation is synonymous with the principle of least squares, as used in multiple regression analysis. However, for more complicated and non-linear models, featuring discrete rather than continuous data, least squares is difficult if not impossible to apply. The likelihood approach to estimation and inference provides a very general approach to parameter estimation, and can be used with a wide range of statistical models

whenever the distribution of the sample data can be defined. The basic idea starts with the specification of the probability distribution of the data, defined in terms of the unknown parameters of the structural model and the parameters of the data distribution. We will write this in general term as $P(X;\theta)$ where X denotes the random variables representing the data to be drawn, and θ represents the vector of parameters to be estimated. Once the data have been drawn, i.e., observed, the function P no longer represents a probability. At this point, the data, now represented as x , are known and there is no uncertainty. Thus the function $P(x;\theta)$ is called a *likelihood*, usually denoted by $L(\theta)$, and represents a function only of the unknown parameters, given known data. Estimation of the parameters θ now consists of finding a set of parameter values, $\hat{\theta}$, that maximize the likelihood $L(\theta)$. These estimates are called *maximum likelihood estimates* (MLE's).

Intuitively, the idea behind these MLE's can be stated as: Data are generated by the probability distribution $P(X;\theta)$. The data x were those actually observed, and will therefore be regarded as corresponding to the random variable values with highest probability of being observed. Maximization of the likelihood $L(\theta)$ for the observed x will therefore provide estimates of the parameters of the distribution that makes $X = x$ have highest probability. The maximization process is often quite difficult to implement, featuring highly non-linear functions. The MLE process for IRT models will be described below.

The IRT likelihood. Under the assumption of local independence, the conditional probability (given η) of observing an outcome vector of responses on n items, $\mathbf{x}$, is given by

$$P(\mathbf{x} | \eta) = \prod_{j=1}^n [P_j(\eta)]^{x_j} [1 - P_j(\eta)]^{1-x_j}, \quad (3.7)$$

where the dependence of the item response function P_j on the item parameters has been suppressed to simplified notation. The joint probability for the random data $\mathbf{x}$ and the latent variable η is obtained as the product of the conditional probability $P(\mathbf{x} | \eta)$ and the population distribution of η , whose form (at this stage) is assumed to be known, and denoted by $g(\eta)$. The marginal probability of $\mathbf{x}$ can then be obtained by averaging the joint probability $P(\mathbf{x} | \eta) g(\eta)$ over η , i.e., in mathematical terms by integrating out the dependence on η . The marginal likelihood for an observation $\mathbf{x}$ can thus be specified as

$$L_i(\boldsymbol{\beta}) = P(\mathbf{x}; \boldsymbol{\beta}) = \int P(\mathbf{x} | \eta) g(\eta) d\eta, \quad (3.8)$$

where the dependence on the unknown IRT parameters is now made explicit. Note that the parameter vector $\boldsymbol{\beta}$ may also include parameters of the population distribution $g(\eta)$ of the latent variable. The marginal likelihood for all independent observations on the subjects is then given by the product of all sample respondents' likelihoods, i.e.,

$$L_N(\boldsymbol{\beta}) = \prod_{i=1}^N P(\mathbf{x}_i). \quad (3.9)$$

The above IRT likelihood is used for binary indicator variables. However, its generalization to the multi-category case is straightforward (see, for example, Bock, 1972).

Application of the principle of maximum likelihood now leads to the well known *maximum marginal likelihood* (MML) approach to IRT parameter estimation developed by Bock and Lieberman (1970), and later refined by Bock and Aitkin (1981). In the Bock and Aitkin version, the averaging or marginalization process over η is facilitated by treating $g(\eta)$ as discrete, and the resulting likelihood equations are solved using an EM (Expectation-maximization; Dempster, Laird, & Rubin, 1977), scheme. Under all versions of the MML algorithm for identically and independently distributed (*i.i.d.*) observations, MML item parameter estimates are obtained by maximizing the marginal likelihood (3.9) with respect to the set of item parameters. The MML estimation procedure yields consistent estimates of the item parameters for scales with finite length, under the assumption that the respondents are a simple random sample from a large population (Thomas, 2001). Alternative approaches include the use of collateral information in the estimation (Mislevy, 1987) as well as Bayesian Markov Chain Monte Carlo (MCMC) methods for simultaneously estimating item parameters and predicting latent variables (Patz & Junker, 1999).

3.4 Advantages and Disadvantages of IRT

3.4.1 Advantages of IRT

The main advantages of IRT models stem from their key feature, namely an explicit measurement model that characterizes the relationship between a latent variable and the observed item responses. In other words, IRT is a model-based approach to scale development (Embretson, 1996; Embretson & Hershberger, 1999; Zickar, 1998), and because of this feature, IRT offers a number of advantages over CTT. Such advantages will be discussed in the following sections (please see Table 3.1 for a summary).

Item-level oriented. As its name implies, IRT primarily focuses on item-level information. In contrast, CTT focuses more at the overall test-level, e.g., calculation of an overall reliability coefficient, and the standard error of an entire scale. The classical true score model provides no consideration of how respondents respond to a given item (Hambleton et al., 1991). While CTT can quantify the item difficulty (i.e., item p-value) and item discrimination (i.e., item-total coefficient correlation) for an item, it does not incorporate such information into the measurement model. Because IRT integrates the item properties into an item response characteristic function to assess item response patterns, it can yield greater information relevant to scoring.

Invariance of item and respondent statistics. CTT measurement results are scale-dependent and respondent-dependent. IRT results however have the property of *invariance* (within a linear transformation). First, item statistics are independent of the sample of respondents from the population of respondents for whom the scale was

designed, a property referred to as *item invariance*. Second, respondent statistics (i.e., latent variable scores) are independent of the particular choice of scale items taken from the population of items which were calibrated, a property that is called *person invariance* (Embretson, 1996; McCollam, 1998). Note that these strong properties derive from a strong assumption that is not always stressed in the IRT literature, namely: invariance depends on the assumption that the IRT model is correct. For an example when this assumption breaks down, see Thomas & Cyr (2002).

Well-defined standard error of measurement. In IRT, latent variable scores have well-defined estimates for the standard error of measurement (Lord, 1980; Lord & Novick, 1968). The standard error of measurement indicates the precision with which each respondent's latent variable level is predicted. This is expressed as a function of latent variable level and can therefore vary from one respondent to another. The standard error of estimates is therefore reported for each level of latent variable, rather than for an entire scale as in CTT (Fraley et al., 2000; Hambleton & Swaminathan, 1985). In CTT, it is assumed that the standard error of measurement is constant across the entire range of the latent variable levels.

Placing items and respondents on a common scale. CTT does not place items on the same scale as the latent variable. The true score model does not include item parameters so that item difficulties are not directly linked to latent variable levels. For IRT models, on the other hand, item difficulty and latent variable predictions are on a common scale with additive effect on item response probabilities. In other words, increasing the latent variable level has the same effect as decreasing item difficulty by the

same amount (McCollam, 1998), as can be seen from the argument of the ogive and logistic response functions, namely $a_j(b_j - \eta_i)$.

Interval level measurement. CTT-based total scores allow for ordinal rather than interval scale interpretations partly because the latent variable scores are not invariant across items. In CTT, at best ordinal measurement is achieved because the relative distance between respondents on total scale scores depend on item characteristics. It should be noted, however, that CTT scales are almost always treated as interval scales by researchers, and that in most cases this is a good approximation. In contrast, the latent variable scale produced from IRT models does not depend on item difficulty characteristics because item parameters are included within the model and latent variable scores can be adjusted accordingly (Embretson, 1996; McCollam, 1998).

Harwell & Gatti (2001) refer to Fischer's (1995) proof that the latent variable predictions possess an interval scale under the Rasch model for dichotomous responses, but they also note that for other IRT models, such as 2-parameter and 3-parameter models, interval scales have not been established. However, their preliminary simulation results suggest that the latent variable predictions under Samejima's polytomous GRM possess an interval scale.

Item information function. An *item information function* is an index for an item, which specifies how much a particular item contributes to the prediction of a particular level of a latent variable. That is, an information function tells us which latent variable level will be most accurately assessed by the item. This index is useful for item evaluation and scale development. Item information functions also have an additive

property, which means they can be added into or deleted from a *scale information function* (an aggregate information function for all items in a scale). Thus, IRT allows scale developers to easily determine the effect of adding, or deleting, a given item or set of items from a scale. Using such information allows us to assemble a set of items that provide acceptable levels of reliability across a range of latent variable levels of interest, and also enable us to know the variation in accuracy of the item or scale as a function of latent variable level.

Figure 3.4 illustrates an item information function, for which the amount of information has a maximum at a latent variable level of zero. Within the range of -1.5 and 1.5, the latent variable level is predicted with some precision; outside this range, the precision (an inverse function of the information) decreases rapidly, and the latent variable levels near both ends of the latent variable scale are not predicted very well. Thus, the information function tells us how well each latent variable level is being predicted. In other words, the precision of a scale depends upon where the respondent's latent variable level is located on the latent variable scale (Baker, 2001).

Figure 3.4: An example of an item information function

Differential item functioning. IRT provides a means to detect an important form of item bias, known as *differential item functioning* (DIF). DIF occurs when one or more items (and ultimately the scale they constitute) function differently across certain subgroups of respondents. In a broad sense, this is a failure of measurement invariance, because the scale is not performing in the same manner for different groups. It is important that scales used by researchers be free of any systematic bias for demographic subgroups, an issue that has been a focus of legal action in the US.

In contrast, CTT-based methods of assessing bias are limited, especially methods that try to assess bias through group mean differences in total scores across demographic groups. Such approaches cannot distinguish between situations where group means differ because of an actual scaling bias (DIF), versus situations where group means differ because latent variable levels are genuinely different on average between the groups (Harvey & Hammer, 1999).

Table 3.1: Major differences between CTT and IRT

Area	Classical test theory	Item response theory
Model	Linear function of true score and random error	Non-linear function of item parameters and latent variable
Level of analysis	Scale level oriented	Item level oriented
Assumptions	Weak (i.e., easy to meet with measurement data)	Strong (i.e., more difficult to meet with measurement data)
Item statistics	Item difficulty (item p-value) and item discrimination (item-total correlation)	Item difficulty, discrimination, guessing parameters (for 3-parameter model)
Invariance of item and respondent statistics	No	Yes, if model fits the measurement data
Measurement precision	Equal measurement error for all respondents	Measurement error for each latent variable level
Detection of measurement invariance	Limited	Differential item functioning technique
Item and respondent characteristics	Not on a common scale	On a common scale
Interval level measurement	No	Yes, proved for Rasch model
Item information function	No	Yes

3.4.2 Disadvantages of IRT

Having highlighted several key features of IRT, the limitations of this modern test theory will be discussed next.

First, IRT requires the assumption of unidimensionality. However, scales used in the social sciences have not been adequately demonstrated to be uni-dimensional (Shevlin, Miles, & Bunting, 1997). For example, the Cognitive Dissonance After Purchase scale developed by Sweeney, Hausknecht, & Soutar (2000) is multidimensional. Recently, however, the development of multidimensional IRT (McDonald, 1999; Reckase, 1997) has provided IRT users with a potential solution to this difficulty.

Second, the IRT model is based on strong assumptions. Although IRT provides techniques for solving some measurement problems that CTT cannot deal with, the validity of IRT techniques may be subject to model failure, i.e., data do not meet the model assumptions (Oshima & Miller, 1990), as is the case with any model-based procedure. Also, it is still not entirely clear what the consequences of violating the model assumptions are (Embretson & Reise, 2000; Reise, 1999).

Another disadvantage is related to the use of the IRT scores. It will be discussed and demonstrated in the later chapters that the use of the IRT scores in subsequent analysis is as problematic as using CTT scores. In other words, though IRT provides many advantages in a measurement sense, it does not resolve the problem of using latent

variable scores in subsequent analyses. This finding is somewhat of a surprise and is a major contribution of this thesis.

Finally, from a practical point of view, it is evident that the cost of using IRT entails learning a mathematically complex theory, which is not trivial, at least to some business researchers.

3.5 Applications of IRT in Business Literature

Although the development of IRT began in the 1950's and although its development has been exponential since then, it is surprising to note that it has been slow to diffuse into the business literature. Published introductions to and applications of IRT in the business literature are largely conspicuous by their absence, and most of the existing examples appeared in the last five years. The earliest work on IRT in the business literature was found to be Kamakura & Srivastava's (1982) article published in *Advances in Consumer Research*. They demonstrated the usefulness of IRT for attitude scaling with a focus on item selection. Not until 1998 was IRT introduced to business researchers and practitioners, when McCollam described it as a "powerful" method for understanding item response behavior in *Modern Methods for Business Research*, edited by Marcoulides (1998). Recently, Singh (2004) demonstrated IRT's potential over CTT by using Role Conflict and Role Ambiguity scales, and suggested that "IRT warrants the serious attention of marketing researchers" (p.205).

Other researchers have used IRT to study the psychometric properties of particular scales and to establish the applicability of IRT models to their data, as well as exploring the general benefits of IRT application. Most of these applications of IRT have focused on previously created scales. To the author's knowledge, IRT models have not yet been used to develop any new scales in business measurement with the exception of the Job Security scale constructed by Probst (2003), where the IRT technique was used for examining items' discrimination and difficulty properties. In what follows, some of these applications will be highlighted.

Personnel Selection. Robie, Zickar, & Schmit (2001) used IRT to evaluate the validity of a variety of personality scales used in personnel selection by examining the measurement equivalence between applicant and incumbent groups. Six personality scales were tested between a group applying for jobs as sales managers in a large retail organization and a group of sales managers currently employed in that organization. Samejima's graded IRT model (see Section 3.2.3 and 3.2.4) was fitted to the personality scales in each group. They found that the personality test items and the scales of interest functioned similarly across the applicant and incumbent groups.

Although not seen in the business literature, other applications of IRT in personnel selection can be to evaluate test/scale equivalence among race, sex, ethnic groups, and to predict personality scores with well defined statistical properties for applicants, to help HR professionals make better hiring decision.

International Research. One of the most critical issues in cross-culture research is whether or not the individual items or the complete scale functions in the same way from one culture to another. Two recent works by Craig & Douglas (2000) and Maheswaran & Shavitt (2000) suggest that IRT can be used to assess culture and translation equivalence in international research. However, this application of IRT has not yet been used by other business researchers in their international studies.

Longitudinal Research. Meade, Hecht, Lautenschlager, Barroso, & Stokes (2000) address the importance of measurement equivalence over time when examining the results of longitudinal studies. If measurement equivalence does not hold over two or more measurement occasions, the observed differences in responses over the different

measurement occasions will not be interpretable. Meade, et al. (2000) demonstrate how IRT can be used to examine measurement equivalence over time in a longitudinal assessment of job satisfaction.

Adaptive Testing. One of the implications of IRT, in contrast to CTT, is that it enables the use of computerized adaptive testing. With scales developed under the CTT framework, respondents can be compared if and only if they are measured using the exact same set of items. However, under the IRT framework, respondents' scale scores are independent of the particular items used in the measurement, provided some items overlap across respondents, and thus are comparable even if they are based on different items. This approach enables researchers to pose to the respondent only those items that can provide useful information for predicting the value of his or her latent variable of interest. Balasubramanian and Kamakura (1989) have successfully applied IRT for binary data to adaptive testing for consumer alienation and discontent measurement. They state that the adaptive testing approach yields cost/time savings without significant sacrifices in measurement precision and recommend its widespread applications for marketing researchers. Singh, Howell & Rhoads (1990) used Monte Carlo simulations to study adaptive designs with the 12 items from the consumer discontent Likert scale (Lundstrom and Lamont, 1976). Their findings suggest that adaptive designs for Likert-type data can yield measurement efficiency and precision with a substantially smaller item pool.

Coupon Redemption Prediction. Drawn from the two-parameter (2-PL) IRT model, Bawa, Srinivasan, & Srivastava (1997) proposed a model of coupon redemption that can help managers predict how a given consumer will respond to a specific coupon

so as to effectively design coupon promotions. In their model, the probability of a consumer redeeming a coupon j is a function of the consumer's coupon proneness and coupon characteristics such that:

$$P_j(\eta) = \frac{e^{Da_j(\eta-b_j)}}{1 + e^{Da_j(\eta-b_j)}},$$

where

$P_j(\eta)$ = probability that a randomly chosen consumer will redeem coupon j

η = the coupon proneness of a consumer

b_j = the attractiveness of coupon j

a_j = the ability of coupon j to discriminate among consumers with different levels of coupon proneness.

D = a constant of 1.7 to make the logistic ICC as close as possible to the normal ogive ICC.

They found that the model correctly predicts redemption intentions for nearly 90% of consumers in a holdout sample and significantly outperforms the traditional coupon redemption model.

Risky Choice Prediction. Zickar & Highhouse (1998) fitted the 2-PL IRT model to four risky-choice items under either a gain or loss frame. A chance to avert losses or a chance to realize gains is the framing of a dilemma that has influence on people's risky choices. For example, when people are given a choice between a sure loss versus a small chance of no loss, they tend to choose the latter option. However, when the same dilemma is framed as an option between a sure partial gain versus a small chance of a complete gain, people prefer the former option (Zickar & Highhouse, 1998). They used an IRT model to relate the probability of choosing a risky option as function of an individual's preference for risk. Their results supported the typical framing finding of

risk-aversion for gains, and risk-seeking for losses, but also found that the semantic framing effects are most likely to be observed for people having moderate risk attitudes. This implies that to understand how framing affects risky choice requires a better understanding of the decision maker's risk attitude

Two recently developed IRT techniques, mixed-model IRT (Rost, 1990, 1991) and non-parametric IRT (Ramsay, 1990) have also been used by business researchers, e.g., Zikar, Gibby, & Robie (2004) and Laroche, Kim, & Tomiuk (1998, 1999).

3.6 Summary

IRT models have been developed to deal with responses that are scored in either a dichotomous (e.g., “true” and “false”) or a polytomous (e.g., Likert-type scales, or nominal scales) fashion. IRT methods are applicable to virtually any type of standardized psychological assessment instrument (Harvey & Hammer, 1999) and have been recommended for use in business by McCollam (1998) and Singh (2004).

IRT-based methods offer some of the best alternatives for devising and optimizing scales and tests, and for performing item analysis; however, it is clearly not the panacea to all the measurement problems. The validity of IRT techniques may be limited by the degree to which data meet the assumptions of the model, as is the case with any model-based procedure.

The applications of IRT in business are still relatively scarce. It is hoped that the current and previous chapters can provide researchers and practitioners in marketing and other social sciences with an awareness of the potential benefits offered by IRT as well as some of its shortcomings, and will encourage them to investigate the use of an IRT framework for developing or analysing their scales.

The remaining chapters in this thesis will focus on the issue of IRT scores and their use in subsequent statistical analysis. It will be seen that the findings will provide readers with cautions regarding this application of IRT. Alternatives will also be discussed.

Chapter 4

IRT- and CTT-based Latent Variable Scores

The preceding chapters provided a review of CTT and IRT. This chapter focuses on the predicted latent variable scores derived from both theories. As noted in the introduction, a central theme of this thesis involves the latent variable scores (i.e., predicted values) generated via the two approaches. In particular, how do they compare, and how well do they perform (in the sense of bias) when used in a two-step latent regression analysis. As described previously, the scoring process for CTT is very simple, consisting of totals, or means, or standardized totals or means of scale items. For IRT, however, prediction is not as simple. The aim of this chapter will be to present a short overview of the various methodologies available for predicting IRT scores. In this section, the model parameters for each item (discrimination and difficulty) will be treated as known. In practice, of course, all parameters must be estimated, but it is standard practice for estimated parameters to be treated as fixed during the estimation.

Given the advantages of IRT over CTT described earlier, one might expect there to be an appreciable difference between the IRT- and CTT- based scores derived from a set of item responses. From a theoretical point of view, the comparison of IRT and CTT

scores possess some difficult mathematical challenges, with the result that the comparative information is fragmentary. A summary of what is known follows:

First, in the one-parameter logistic model with binary items, the Rasch (1960) model, the unweighted total score is a *sufficient statistic* for predicting the latent variable, η , i.e., the pattern of the individual item response provides no information over the sum of the item responses (McDonald, 1999; Nunnally & Bernstein, 1994; Sijtsma & Hemker, 2000; Thissen & Orlando, 2001). In other words, the Rasch model implies that all respondents with the same total scores have the same IRT score when that predicted score is obtained using the method of maximum likelihood (ML), described later in this chapter. Among polytomous IRT models, only the Rasch family -- the partial credit model (Masters, 1982) and the rating-scale model (Andrich, 1978), have sufficient statistics of raw total scores (Embretson & Reise, 2000).

Second, in the dichotomous two-parameter logistic model with known parameters, a weighted total score (with each item weighted by its discrimination parameter) is the sufficient statistic for predicting η (Bond & Fox, 2001; McDonald, 1999; Nunnally & Bernstein, 1994; Sijtsma & Hemker, 2000). In other words, weighted total scores contain all the information for predicting IRT scores using the ML techniques. A respondent who answers a given number of more discriminating items will have higher IRT ML-based scores than a respondent who answers the same number of less discriminating items.

Lastly, Sijtsma & Hemker (2000) proved a *stochastic ordering* property, which can be described mathematically by the probability statement:

$$P(\eta \geq \eta_0 \mid X_+ = x_1) \leq P(\eta \geq \eta_0 \mid X_+ = x_2), \quad (4.1)$$

where X_+ is the raw total score, X_1 and X_2 are two observed total scores with $x_1 < x_2$. It implies that, in probability, the respondents' ordering on the total score will be the same as their ordering on η . This property holds for all dichotomous IRT models and for two polytomous IRT models – the partial credit model (Masters, 1982) and the rating-scale model (Andrich, 1978). This is a useful property if the total score is to be used in place of the IRT scores for ranking individual respondents.

In the following, methods for computing IRT predicted scores, together with empirical results on the relative performance of IRT-based scores versus CTT total scores, will be discussed.

4.1 Methods for Predicting IRT Scores

This section describes how IRT models are used to score respondents on a latent variable η . Although the discussion is limited to dichotomous item models, most of the methodologies can be extended to polytomous item response models. In IRT a respondent's predicted score, $\hat{\eta}$, on the latent variable is not the same as the total score calculated in CTT. Unlike the latter, an IRT-score is a non-linear function of the observed item responses. A variety of methods have been proposed for predicting the value of the latent variable for a specific respondent, given the respondent's measurement outcomes, denoted by the vector $\mathbf{x} = (x_1, \dots, x_n)$. These prediction methods generally fall into two broad classes: those based on the likelihood function, and those based on the posterior distribution of η .

4.1.1 Likelihood Based Prediction Methods

Maximum Likelihood Estimates (MLE). MLE's of respondents' latent variable levels are based on the idea that a single respondent's latent variable level is a fixed parameter, η_i , and that the respondent's latent variable score is a random draw from a population of respondents having that particular latent variable level. The appropriate likelihood for ML estimation is then given by equation (3.7), with item parameters fixed at their MML estimated values. Maximization of this likelihood yields the predicted latent variable score $\hat{\eta}_i^{ML}$, which is a function of that respondent's response outcome vector, which we can express as $\hat{\eta}_i^{ML}(\mathbf{x}_i)$.

Lord (1983) showed that the asymptotic conditional bias of $\hat{\eta}_i^{ML}$, namely the bias in the estimate at a particular value of the latent variable, denoted $E(\hat{\eta}_i^{ML} - \eta | \eta)$, is of order $O(n^{-1})$, where n is the number of scale items. Note that a bias of order $O(n^{-1})$ means that the bias goes to zero at a rate proportional to n^{-1} as n goes to infinity, or, for large n , the bias is equal to some constant multiplied by n^{-1} . He also gave an expression for the conditional mean squared error (MSE) of $\hat{\eta}_i^{ML}$, with error term of order $o(n^{-1})$, i.e., the error goes to zero faster than n^{-1} , and further showed that for identically and independently distributed (*i.i.d.*) samples, the population variance of the latent variable η differs from the population variance of the ML estimate $\hat{\eta}_i^{ML}$, by terms of order $O(n^{-1})$. Lord (1983) also noted that even for long scales (of 50 to 100 items), order n^{-1} adjustments to ML estimates designed to yield unbiased estimates of η and of the

population variance of η (correct to order $O(n^{-1})$) may still be inadequate if the sample contains respondents with very low levels of the latent variable, for which predictions of latent variable scores are particularly inaccurate. A notable property of an MLE is that no estimates can be computed when all questions are answered positively, or when all questions are answered negatively.

Weighted Likelihood Estimates. Warm (1989) proposed an estimator that is closely related to the MLE. It has the same asymptotic variance as the MLE, but a smaller conditional bias that is $o(n^{-1})$ as $n \rightarrow \infty$, compared to $O(n^{-1})$ for the MLE (Lord, 1983). This estimator is calculated by maximizing a weighted likelihood obtained by multiplying the likelihood (3.7) by a function $f(\eta)$ designed to remove the first order bias term. Though this weighted likelihood estimator (WLE) has the form of a Bayes modal predictor (please see below), it is likelihood-based since it makes no assumptions about the prior distribution, $g(\eta)$, of the latent variable.

4.1.2 Prediction Methods Based on the Posterior Distribution

As noted, estimation of IRT parameters are most commonly derived using marginal maximum likelihood methods. ML and its variants like WML are also frequently used for predicting / estimating respondent latent variable levels. However, for prediction purposes, Bayesian methods are often used. These are based on the notion that the latent variable, η , and the vector of response outcomes, $\mathbf{x}$, for a respondent are jointly distributed in the population according to a joint distribution given by $P(\mathbf{x} | \eta) g(\eta)$. Instead of averaging over η to get a marginal likelihood as in equation (3.8), the Bayesian approach is to convert the former distribution into a conditional distribution $P(\eta | \mathbf{x})$ by dividing the joint distribution by the marginal distribution of $\mathbf{x}$. This leads to the expression referred to in Bayesian terminology as the *posterior distribution*:

$$P(\eta | \mathbf{x}_i) = \frac{P(\mathbf{x}_i | \eta) g(\eta)}{\int P(\mathbf{x}_i | \eta) g(\eta) d\eta}. \quad (4.2)$$

The dependence of the posterior distribution on the item parameters has been suppressed to simplify the notation. It should be recalled that the usual procedure is to fix the item parameters contained in the posterior distribution and the parameters of the latent distribution $g(\eta)$ (if any) at their estimated values. Predictors of this type are usually referred to as *empirical Bayes predictors*.

Bayes Modal Predictor. The Bayes modal predictor is also called the *maximum a posteriori* (MAP) estimator by psychometricians. The MAP is obtained by maximizing

equation (4.2) for each subject. Results on the asymptotic bias and variance of the MAP estimator for known item parameters are reported by Warm (1989) and Hoijsink & Boomsma (1996) based on some earlier results from Lord. As with the MLE, the conditional and unconditional bias of the MAP is of order $O(n^{-1})$.

Bayes Posterior Mean. Called the *expected a posteriori* (EAP) estimator by psychometricians, this estimator is the mean or expected value of the posterior distribution (4.2). It is usually evaluated using Gauss-Hermite numerical integration of the posterior latent distribution (see Stroud and Secrest, 1966). Although theoretically the conditional bias of the EAP estimator, $E(\hat{\eta}^{EAP} - \eta | \eta)$, is not available, Kim & Nicewander's empirical results (1993) show that it is close to that of MAP, namely $O(n^{-1})$. However, Thomas (2001) proved that the unconditional bias of EAP is zero and its unconditional variance underestimates the unconditional variance of the latent variable distribution, and the degree of underestimation can be considerable unless the number of scale items is large (see, for example, Thomas and Cyr, 2002). The EAP has the advantage that it is easier to compute compared to MAP and MLE (Kim and Nicewander, 1993) and will be used as the IRT-based latent variable predictor compared to CTT-based predictor in this thesis.

4.2 Relative Performance of the IRT- and CTT-based Scores

Theoretical results on the performance of the various prediction methods are all asymptotic in nature. While these provide useful clues to predictor behaviour, an investigation of empirical and Monte Carlo studies comparing various latent variable scores based on finite item size is also essential.

A number of studies have evaluated the extent to which IRT-based scores differ from the CTT-based scores. A study by Lawson (1991) compared predicted latent variable scores (MLE) based on the dichotomous 1-parameter Rasch model with total scores for three different data sets, consisting of one set of 14 items and two sets of 35 items. He found that they are highly correlated (with correlations greater than .986 in all three data sets).

Kim and Nicewander (1993) carried out a detailed investigation of the bias, standard errors, and reliabilities of several IRT-based latent variable predictors (MLE, WLE, MAP, and EAP) and standardized total scores, using Monte Carlo techniques based on the 3-parameter logistic model with 50 binary items. They concluded that except for the MLE with lower reliability, all IRT-based predictors are as reliable as the standardized total scores. Nevertheless, all five exhibited relatively large conditional bias for $|\eta| > 1$, with η measured on an $N(0,1)$ scale, the worst being the standardized total score and the MLE. Conditional bias, reliability, and standard errors were similar for the

WLE, MAP, and EAP, with the WLE exhibiting slightly lower bias but higher standard error than the other two.

Ndalichako & Rogers (1997) compared an IRT-based MLE predictor for the 1-parameter binary logistic model, and two Bayesian predicted scores for the 2- and 3-parameter logistic IRT models, with CTT total scores. They used a dataset of 1,230 respondents containing responses to a 70 item reading comprehension test, and found that the correlations among the latent variable scores were very high (all correlations greater than .977). For the sake of simplicity and ease of application, Ndalichako & Rogers (1997) advocated the continued use of CTT for scoring and item analyses.

Fan (1998) examined latent variable predictions derived from the 1-, 2-, 3-parameter IRT and CTT models with binary items. In the study, several samples of 1000 respondents were randomly drawn from a large pool of more than 193,000 students who participated in a reading exam (48 items) and a math exam (60 items). His findings indicate that IRT-based latent variable scores (MLE's) and total scores are highly comparable, with average correlations greater than .96 for all conditions examined.

Most recently, MacDonald & Paunonen (2002) used Monte Carlo techniques to investigate the behavior of latent variable scores obtained from the 2-parameter binary logistic IRT model and CTT under the conditions of 20, 40, and 60 items. Their findings suggest that IRT- and CTT-based latent variable scores are highly comparable, with an overall average correlation of .985. Both IRT- and CTT-based latent variable scores accurately predict the true latent variable values of the simulated subjects; under all

simulated conditions, both IRT and CTT derived latent variable scores are highly correlated with the true latent values, with correlations of .965 and .952, respectively.

Out of the empirical studies on the performance of CTT- and IRT-based scores, only two used polytomous IRT models. An earlier study by Koch (1983) used Samejima's GRM to analyze a 5-point Likert scale data set with 40 items and found that IRT-based empirical ML latent variable scores and total scores are highly correlated. However, he pointed out that the reason for such high correlation may be because the discrimination parameter estimates did not vary much across the items.

A recent study by Embretson & Reise (2000) examined total scores and latent variable scores under various polytomous models, including the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), the rating-scale model (Andrich, 1978), and other polytomous IRT models. This study used the responses of 350 students to the 12 items on the Neuroticism scale with 5 options of the Neuroticism Extraversion Openness Five-Factor Inventory (NEO-FFI). The results show that all IRT-based scores are highly correlated (over .97), and in turn, each is highly correlated with total scores.

The findings summarized above suggest that IRT- and CTT-based scores are highly comparable for both dichotomously and polytomously scored responses regardless of the IRT model used (please see Table 4.1 for a summary). However, the results of the most of the studies were based on a large number of items, above 20, which is not commonly seen in business research (e.g., scales in *Handbook of Marketing Scales*, by Bearden & Netemeyer, 1999). Moreover, in the polytomous situation, only two empirical studies

were found and both used 5-option scales. Therefore, practical conclusions for business researchers must be based on simulation studies for moderate numbers of items with a selection of numbers of categories. A more thorough examination of the effect of skewness in the discrete variables should also be carried out. Furthermore, the correlations reported in these studies are all Pearson product-moment type.

4.3 Summary

This chapter introduced several methods for predicting latent variable values based on IRT. However, all the latent variable predictors are susceptible to sources of error and bias which are decreasing functions of the number of items in the scale.

Despite the theoretical advantages of IRT over CTT, it is somewhat surprising to find that for both dichotomously and polytomously scored responses, the latent variable scores based on the two measurement frameworks are highly comparable on the average. The overall finding is that IRT- and CTT-based scores are strongly correlated across 1-, 2-, and 3-parameter IRT models and that they are also strongly correlated with the underlying latent variable. At the first sight, this suggests that IRT and CTT methods are equally effective and that either can be used to generate scores for use in subsequent analyses. However, as will be documented in the next section, such a conclusion would be premature. Though IRT and CTT scores are similar in many respects, there are some differences, and the use of such scores in analyses is problematic, despite the apparent high correlations with the latent variable. We will also see later in the thesis that that IRT approach has a natural connection to a structural equation methodology that resolves many of the analytic difficulties. The studies reported above also feature some gaps in their coverage of the various design characteristics of scales or tests. For example, how do IRT and CTT scores compare for the typical scale lengths and item categories used in business research? How close are the IRT and CTT scores to the true latent variable values, when the binary and polytomous manifest variables display appreciable degrees

of skewness? These questions require more rigorous empirical scrutiny, and detailed simulation results on these questions will be described later in this thesis.

Table 4.1: Summary of empirical studies on the performance of IRT- and CTT-based scores

Study	Models	Latent variable scores	Items	Results		Data
I. Dichotomously scored responses						
Lawson (1991)	IRT (1-PL)	MLE	14 35	Pearson correlation	All have strong correlations (greater than .986).	Real dataset
	CTT	NR				
Kim & Nicewander (1993)	IRT: 3PL	MLE, WLE, MAP, & EAP	50	Reliability, Conditional Bias, and standard error	- All have similar reliabilities, except for MLE with lower reliability - MLE and Standardized total scores have larger conditional bias; WLE shows slightly lower conditional bias but higher standard error than MAP and EAP	Monte Carlo
	CTT	Standardized NR				
Ndalichako & Rogers (1997)	IRT: (1-3 PL)	MLE & Bayesian predictors	70	Pearson correlation	Nearly perfect correlations (all above .977)	Real dataset
	CTT	NR				
Fan (1998)	IRT: (1-3 PL)	MLE	48 60	Pearson correlation	Nearly perfect correlations (overall average grater than .96)	Real dataset
	CTT	NR				
MacDonald & Paunonen (2002)	IRT: 2PL	MLE	20 40 60	Pearson correlation	- High correlations (overall average of .985). - High accuracy, with correlations above .95 with true values.	Monte Carlo
	CTT	NR				
II. Polytomously scored responses						
Koch (1983)	IRT: GRM (5 options)	MLE	40	Pearson correlation	Both are highly correlated (.98)	Real dataset
	CTT	Total scores				
Embretson & Reise (2000)	IRT: M-GRM, GRM, PCM, RSM, etc. (5 options)	GRM: MAP M-GRM: EAP Others: MLE	12	Pearson correlation	All are highly correlated (above .97).	Real dataset
	CTT	Total scores				

Note: GRM refers to graded response models, M-GRM to modified GRM, PCM to partial credit models, and RSM to rating scale models. NR refers to the number right score for dichotomously scored responses.

Chapter 5

The Use of Latent Variable Scores in Statistical Analyses

The aforementioned research comparing correlations between IRT and CTT scores is oriented primarily towards the first goal of scaling, namely measurement of latent variables for an individual respondent as in psychometrics. Unlike the case of psychometric assessment, however, the measurement (i.e., scoring) of latent variables for an individual is not the primary goal of research. Rather, it is the relationship among latent variables that is of concern to most researchers.

Most recently, Bollen & Lennox (1991), Fox & Jones (1998), Shevlin, et al., (1997) have pointed out that many research studies involve a two-step approach: predict the latent variable scores first, then use these scores in subsequent analyses, such as regression. Baumgartner & Homburg (1996) reviewed the applications of SEM in four marketing journals (the Journal of Marketing, Journal of Marketing Research, International Journal of Research in Marketing, and the Journal of Consumer Research) between 1977 and 1994. It can be calculated from their articles that at least 22 percent of the 103 models examining relationships among latent variables in a multi-equation system used composite scores for all the latent variables in the model, and that 41 percent used composite scores for at least one latent variable in the model. Another recent study

by Holbert (2002) reviewed the use of SEM from 1995-2000 in 37 communication-related journals. They reported that about 58 percent of the 118 multi-equation models examining latent variable relationships employed single-indicators or composite scores. Note also that these two reviews did not include studies which used the two-step approach in single equation models, like simple or multiple regressions. Thus, they in fact understate the prevalence of the two-step approach. While the two-step approach is very common, it has serious drawbacks, regardless of the type of scoring technique used.

In the remainder of this chapter, we first explain why the direct use of latent variable scores in subsequent statistical analyses may be problematic. Empirical studies on the use of latent variable scores in statistical analysis will be reviewed, especially for scores derived from CTT and factor analysis (FA) of continuous scale items, as well as from IRT analysis of discrete scale items. Thus the goal of this chapter is to investigate the *finite item bias* (i.e., the bias in parameter estimates resulting from inadequate item size) caused by using scores in standard statistical analyses. Methods to correct for, and to remove, this bias will also be briefly described.

While predicted latent variable scores can be used in many types of statistical analyses, regression will receive the most attention in this chapter, as regression is widely used in many business applications. Many of the methods proposed for dealing with bias are applicable no matter what analyses the scores have been used for.

5.1 Problems with Directly Using Latent Variable Scores in Statistical Analyses

The direct use of latent variable scores in regression and other statistical analyses can result in inconsistent or biased estimates of model parameters, and if the latent variable scores are based on too few items, this bias can be severe. The difficulty arises because the distribution of the predicted scores does not converge to the distribution of the latent variable as the sample size grows, *unless the number of scale items used to derive the scores is very large*. Analyses based on the predicted scores are then subject to bias, regardless of sample size (Mislevy et al., 1992). Convergence occurs only when the number of scale items goes to infinity. *Finite item bias* then is the bias that results from not having an infinite number of items.

The problem with the distribution of latent variable scores is related to the discussion of Section 4.1, where we have seen that IRT-based predicted latent variable scores, $\hat{\eta}$, are susceptible to sources of error and bias, which are decreasing functions of the number of items in the scale. The more items, the more precise will be the IRT latent variable scores. Clearly then, for a scale with a fixed number of items, the distribution of IRT-based scores, $\hat{\eta}$, will be a biased estimate of the true latent variable distribution. This bias does not go to zero as sample size goes to infinity, it only goes to zero as the number of items in the scale goes to infinity. This conclusion holds for other methods of latent variable prediction, in particular CTT. It also applied to scores obtained following factor analyses of continuous indicators

Factor Scores with Continuous Indicators. Bollen & Lennox (1991) illustrated the above problems based on composite scores (of which raw total scores are the simplest version and factor scores are a sophisticated version) in factorial analyses, with continuous indicators. Their illustration was based on a simple factor model, with scale items or indicators linearly dependent on a latent variable, i.e.,

$$y_j = \lambda_j \eta + \varepsilon_j \quad \text{for } j = 1, 2, \dots, n, \quad (5.1)$$

with η the latent variable underlying each indicator y_j , and ε_j the measurement error for the j^{th} indicator. Here λ_j is the loading representing the systematic effect of η on y_j . Under the standard factor analysis assumptions, y_j and η are scaled as deviations around their means, ε_j is uncorrelated with η , ε for one indicator is uncorrelated with ε for another indicator, and the mean of ε_j is zero.

Consider the equally weighted linear composite score, c_y , a type of scoring often used by researchers to construct factor scores from items that load heavily on a given factor:

$$\begin{aligned} c_y &= y_1 + y_2 + \dots + y_n \\ &= (\lambda_1 + \lambda_2 + \dots + \lambda_n) \eta + (\varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_n) \\ &= \lambda \eta + \varepsilon \end{aligned} \quad (5.2)$$

where $\lambda = \lambda_1 + \dots + \lambda_n$ and $\varepsilon = \varepsilon_1 + \dots + \varepsilon_n$. Notice that c_y contains measurement error.

The squared correlation between c_y and η is

$$\rho_{c_y, \eta}^2 = \frac{[\text{cov}(c_y, \eta)]^2}{[\text{var}(c_y) \text{var}(\eta)]} = \frac{\lambda^2 \text{var}(\eta)}{[\lambda^2 \text{var}(\eta) + \text{var}(\varepsilon)]}. \quad (5.3)$$

Since the numerator $\lambda^2 \text{var}(\eta)$ is the true variance of the composite score c_y and the denominator is the observed variance of c_y , equation (5.3) is actually the reliability of c_y by definition (2.7). It can be seen that c_y and η are perfectly correlated, i.e., the reliability of c_y will be one, only in the unlikely situation where $\text{var}(\varepsilon) = 0$. In other words, c_y has less than perfect correlation with the latent variable η . Even with optimal weighting, this correlation cannot be raised to one due to the measurement error ε , so that the linear composite score will not be equivalent to the latent variable. Again, this illustrates that the distribution of c_y is a biased estimate of the distribution of η . The latent variable η is not observed, nor will any combination of the observed items be a perfect indicator for it.

Bollen and Lennox (1991) concluded that using composite scores in regression will lead to inconsistent estimates for the regression coefficient. Here, a *consistent estimator* is referred to as an estimator that “converges in probability to the value of the parameter being estimated as the sample size increases” (McDonald, 1996, p.248). To put it simply, if a parameter estimate approaches the true parameter value with high probability as sample size goes to infinity, then we call this estimator “*consistent*”. For a simple linear regression where a composite score variable is used as the predictor, and it is measured with error, then the estimated regression coefficient will tend to be too low. If a multiple regression or multi-equation system is used, the problem due to measurement error

becomes more difficult – coefficient estimates can be higher, lower, or even the same as the true coefficient (Bollen, 1989; Mackenzie, 2001).

However, it can be shown that, for two explanatory variables measured with error in a multiple regression, the estimated regression coefficient will tend to be attenuated, unless there is a large difference in the magnitude of the measurement errors of the explanatory variables (see Appendix C). In this thesis, only cases of attenuation have been explored, as these comprise the most typical and realistic cases, particularly for the case of two exogenous latent variables. For an extensive theoretical discussion and review of measurement error and its implications, please refer to Fuller (1987).

Another problem of using composite total scores to estimate the correlation between two latent variables can be seen as follows. Suppose that c_y and c_x are total scores on two scales, η and ξ are the latent variables underlying c_y and c_x , respectively, ρ_{c_y} and ρ_{c_x} are the corresponding reliability coefficients, ρ_{c_y, c_x} is the correlation between c_y and c_x , and $\rho_{\eta, \xi}$ is the correlation between η and ξ . It can be shown that

$$\rho_{c_y, c_x} = \rho_{\eta, \xi} \sqrt{\rho_{c_y} \rho_{c_x}}. \quad (5.4)$$

The above equation shows that the observed correlation ρ_{c_y, c_x} is less than the actual correlation $\rho_{\eta, \xi}$ (as the reliabilities can not be greater than one). This reduction in correlation is known as *attenuation of the correlations* due to errors of measurement (McDonald, 1999).

To illustrate the impact of ignoring measurement error, consider the correlation between salesperson job satisfaction and job performance. Let the observed correlation between the two scale scores be 0.3 and let the reliability of each scale is 0.5. Apply formula (5.4) and then it can be seen that the resultant correlation of the latent salesperson job satisfaction and job performance is 0.6, which is twice as big as the observed correlation. This shrinkage in the observed correlation is due to measurement error.

From the above discussion, we have seen the consequences of ignoring measurement error; thus, the need to take into account the measurement error in modeling is very important.

5.2 Studies on the Use of Latent Variable Scores in Regression

This section focuses on studies that examine the finite item bias caused by using latent variable scores directly in standard statistical analyses, and in particular regression analysis.

5.2.1 Scores from Continuous Scale Items: CTT and FA

One of the first studies on the use of predicted FA scores in regression and other analyses was by Tucker (1971). He showed that although some factor score prediction methods would result in inconsistent parameter estimates for certain analyses, consistent estimates could also be obtained in specific cases. In particular, for the regression of a dependent variable (measured without error) on one or more factor scores, three of the four score generation methods he studied led to inconsistent estimators of the regression parameters. Of the four methods, he showed that consistent estimates could be obtained only if the FA scores were generated using the “regression” method, which is equivalent to a posterior mean (i.e., an EAP factor analysis score) under the assumption that the items have a multivariate normal distribution (Johnson and Wichern, 1998).

Shevlin et al. (1997) used simulation to examine a similar setup, but with explanatory factor scores computed directly as the sum of the indicator variables loading on each factor (see c_y in equation 5.2), a method that is still commonly used in the behavioural sciences, witness the references cited by Shevlin et al, 1997, and by Skrondal

& Laake, 2001. Their main finding was that for the case of two explanatory factors, the estimated regression coefficients were biased, and that the degree of bias increased as the true factor loadings decreased. When the correlation between the explanatory factors was low, factor loadings of 0.9, 0.5, and 0.3 (corresponding to reliabilities of 0.81, 0.25, and 0.09, respectively) resulted in average (negative) regression coefficient biases of 17%, 21%, and 42%, respectively.

Both Shevlin et al. (1997), and more recently Skrondal & Laake (2001), have noted some confusion in the literature, with some authors asserting that the direct use of factor scores in regression analysis inevitably results in biased and / or inconsistent estimates (see, for example, Bartholomew, 1981), in contrast to Tucker's (1971) findings. This confusion has, however, been cleared up by Skrondal & Laake's (2001) recent article, as described below.

These authors studied the system of latent regression (LR) models given by

$$\eta = \Gamma\xi + \zeta \quad (5.5)$$

where η is a vector of latent response variables, ξ is a vector of latent explanatory variables, Γ is a matrix of regression coefficients and ζ represents a vector of disturbances independent of ξ , where ζ has the covariance matrix ψ . The complete specification of system (5.5), which is a special case of the general linear structural relations model (LISREL), requires the specification of measurement models for η and ξ , namely,

$$\mathbf{x} = \Lambda_x \boldsymbol{\xi} + \boldsymbol{\delta} \text{ and } \mathbf{y} = \Lambda_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad (5.6)$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors of items (or indicator variables) of dimensions $p \times 1$ and $q \times 1$, respectively, and the “loadings” Λ_x and Λ_y are matrices of fixed parameters. The vectors $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ contain random disturbances that are independent of the latent variables and of each other. Consistent estimates of the regression parameters Γ can be obtained by maximum likelihood (along with methods such as weighted and unweighted least squares) using a number of computer programs, e.g., LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1995), AMOS (Arbuckle & Wothke, 1999), Mplus (Muthén & Muthén, 2001), LINCOS (Schoenberg & Arminger, 1990). Robust estimation methods are also available in a number of these packages, which allow for estimation when distributional assumption of normality is violated.

The primary focus of the Skrondal & Laake (2001) study was the use of *blockwise factor score regression*, which consists of the following steps:

1. Separate confirmatory FA of the parameters of the two measurement models in equation (5.6), which are required to be identified by the covariance matrices of $\mathbf{x}$ and $\mathbf{y}$.
2. Computation of separate factor estimates (i.e., $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\xi}}$) from each measurement model, using either the regression method, or the Bartlett method (see, for example, Johnson and Wichern, 1998).
3. OLS regressions of $\hat{\boldsymbol{\eta}}$ on $\hat{\boldsymbol{\xi}}$ to obtain an estimate, $\hat{\Gamma}$, of the regression coefficient matrix, Γ .

They also examined a variant in which separate factors were associated with distinct subsets of indicator variables, which they referred to as *factorwise* FSR (factor score regression). The finding of greatest relevance to this review is that, in general, blockwise FSR yields inconsistent estimators of Γ . The degree of inconsistency depends on the number of response scale items and their coefficient of determination. In their simulation study, they showed that the standard errors of the regression parameters are more severely downward biased than the parameter estimates themselves, so that hypothesis testing will lead to rejection rates exceeding the nominal rates. They concluded that “conventional factor score regression performs very badly and should definitely be abandoned”.

Skrondal & Laake (2001) also proved an intriguing result, namely that if Bartlett scoring is used for η , and regression scoring is used for ξ , this *revised* blockwise FSR approach yields consistent estimates. A corollary of this result confirms Tucker’s (1971) earlier conclusion for regression models featuring a measured response variable. For *factorwise* FSR, where separate one-factor models are used for sets of distinct item subgroups, with corresponding separate estimation of factor scores, Skrondal & Laake (2001) further showed that the Bartlett-and-Regression scheme (as mentioned above) yields consistent estimates if the explanatory latent variables, ξ , are uncorrelated, which is a rather special case. However, for the case of a single explanatory latent variable, consistency is achieved by the Bartlett-and-Regression scheme. Research on the use of linear factor scores in regression continues.

Very recently, Croon (2002) explored theoretically the issue of bias in regression parameter estimates, and showed how the variances and covariances of predicted factor scores and the covariances between predicted factor scores and measured variables can be corrected to obtain consistent estimates of regression coefficients. However, Croon's approach is difficult to apply as a general technique.

Jöreskog (2000) has recently introduced a method of generating factors scores that are constrained to reproduce the full covariance matrix of the latent response and explanatory variables. Under his scheme, which works only for continuous indicators, factor score regression generates consistent regression estimates. For continuous indicators, this warrants further empirical testing. How does it perform for moderate sample sizes? Since the focus of this thesis is on discrete indicators, this question will not be pursued in this study.

5.2.2 Scores from Discrete Scale Items: IRT

While the focus has now switched to scores generated from discrete scale items, the goal of this section is the same as that of the previous, namely to investigate the bias in parameter estimates caused by using scores directly in standard statistical analyses. As in Section 5.2.1, regression analysis will receive the greatest attention.

Examples of the direct use of IRT derived latent variable scores in regression are less common in the literature than examples that use factor scores, which is not surprising given that IRT is not as well known as FA outside the fields of psychometrics and educational testing. Though there seems to be a general agreement in the technical literature that using IRT scores directly is inadvisable (see, for example, Mislevy, et al., 1992; Houtink & Boomsma, 1996) and leads to bias in parameter and standard error estimates, unless the number of items is large, specific empirical evaluations of such biases are scarce.

One study that does provide some evaluation is that of Adams, Wilson and Wu (1997), who studied an extension of the Rasch model that encompasses a number of model types, including the partial credit model (Masters, 1982) which is designed to handle ordered categorical indicator variables. Their estimation scheme incorporated *collateral* information, i.e., they represented the latent variable η as a linear function of collateral information, Y , such as gender, socioeconomic status, etc. Thus their model for η consisted of a simplified latent variable regression model of the form,

$$\eta = Y\beta + \zeta, \quad (5.7)$$

where the vector of disturbances, ζ , was assumed to be normal with mean zero and variance σ^2 .

The Adams et al. (1997) approach is closely related to that used earlier by Mislevy & Sheehan (1989), and differs from that of Bock & Aitkin (1981) in that more prior information is assumed for η . Among other things, Adams et al. (1997) provided a comparison between three estimators of β , namely:

1. The estimator obtained using a simultaneous maximum likelihood (*ML / EM*) estimation of the model item parameters and the regression coefficients.
2. The item parameters were first estimated from the item responses alone, without using the collateral information, then used to generate EAP estimates of the latent variable, η . The regression parameters, β , were then determined by OLS regression of the EAP scores on the collateral information Y (a two-step method).
3. The regression parameters were determined from a set of plausible values of η , a completely different methodology from that discussed hitherto. Plausible values methodology will be described in the next section.

Their simulation results demonstrated that Method 2 (the two-step method) underestimated both R^2 and the magnitude of the regression coefficients, while the parameter estimates from Method 1 and Method 3 were close to the true values.

5.3 Possible Remedies for the Finite Item Bias Generated by the Use of Scores

5.3.1 The Case of Continuous Indicator Variables

Though this thesis focuses on discrete indicator variables, both binary and polytomous, it is instructive to start this discussion by considering the techniques of bias reduction that have been devised for dealing with finite item bias arising from measurement error in the case of continuous variables.

For scores based on continuously measured indicator variables, Bollen (1989), and Bagozzi & Heatherton (1994), introduced the following procedure which can account for measurement error in a limited way. This approach involves the following four steps for each construct: (a) generate the composite score from the multiple-item scale (e.g., total scores); (b) compute the reliability of the scale (e.g., Cronbach's alpha); (c) specify the composite score as a single variable loading on a latent variable; (d) fix the error variance of the composite score to $[(1 - \text{reliability}) \times \text{variance of the composite score}]$. This procedure can be implemented via any of the modern structural equation modelling (SEM) packages listed in Section 5.3.4. Stephenson & Holbert's (2003) simulation results showed that for continuous variables this approach yields more accurate path coefficients than the method that treats single-item scores or composite multiple-item scores as observed variables in path analysis (i.e., assuming no measurement error in the variable). However, it does not perform as well as the SEM approach to be introduced in Section 5.3.4.

For the case of continuous indicators, as noted previously, Skrondal & Laake's (2001) revised blockwise FSR approach yields consistent estimates. Croon (2002) also argued that there are some situations in which a two-step approach to latent variable regression and other analyses might be advantageous. For example, the search for appropriate measurement models, and the search for the correct functional form of a latent variable regression model, may be easier to undertake if the measurement models are estimated separately from the latent variable model. Similar arguments can be advanced for the case of discrete indicator variables, when IRT scores must be used as proxies for latent variables. If a two-step approach to latent variable regression is to be used in the discrete case, methods for incorporating bias corrected IRT scores become important. However, this is more difficult to do in the discrete case than in the continuous case, and only one effort in this direction has been founded by the author in the literature. This is described in the next sub-section, followed by some other possible remedies for the case of discrete indicator variables.

5.3.2 Bias-Corrected IRT Scores: Discrete Indicators

The results of Kim & Nicewander (1993) reported earlier showed that the least biased IRT predictors were the posterior mean (EAP), the posterior mode (MAP), and the weighted likelihood predictor (WLE). Hoijtink & Boomsma (1996) investigated the bias in MAP and WLE predictors both empirically and theoretically, and developed asymptotic expressions, correct to order $O(1/n)$, for the conditional mean and variance of the latent variable, η , the covariance between η and a measured variable, y . For the WLE predictor, these expressions can readily be evaluated and used in many kinds of analyses, including ANOVA and multiple regression. Based on an ANOVA simulation, they concluded that their asymptotic bias correction method was necessary, but that it still required at least 15 binary items in the IRT model to keep true biases in the ANOVA means and error estimates to an acceptable level. However, Hoijtink & Boomsma's (1996) method does not comprise an easily generalized approach to the problem of latent variable analysis when the indicator variables are discrete. Each specific analysis must be treated as a special case, and the mathematical treatments required will not appeal to most applied researchers.

5.3.3 Plausible Values: Discrete Indicators

The *plausible values* technique differs from the techniques defined above in that M multiple sets of scores (five or more) are provided for each latent variable. Plausible values in fact represent multiple imputations (Rubin, 1987) drawn from the predictive distribution of the latent variable, the latent variable in this case being treated as missing data. The predictive distribution in this case is similar to the posterior distribution described in section 4.1.2, except that it also incorporates the collateral information Y available on each respondent (see the discussion of Adams, Wilson, & Wu, 1997). The technique consists of running the regression separately for each of the $M = 5$ or more sets of predicted latent variables, and then averaging the regression coefficients over the M multiple imputations. The technique was originally designed for surveys such as the biannual U.S. National Assessment of Education Progress (NAEP) survey, in which individual students may be asked to answer as few as eight questions from a particular test, so that none of the likelihood based or posterior distribution based predictions will provide sufficient accuracy. The approach is quite general, however, and could be adopted in marketing and other business applications, though it does not appear to have been used to date, no doubt due to the labour involved in computing the predictive distribution. The predictive distribution incorporates the variability arising from using a finite number of scale items. Therefore, plausible values estimators, which consist of estimated expected values taken with respect to the predictive distribution, will be free of finite item bias.

Individual plausible values cannot be interpreted as predicted scores for particular individuals because each of the M plausible values consists only of a random draw from the predictive distribution of the latent variable, hence the name. The M values for each regression coefficient can also be used to provide an estimate of the “between imputation” component of the variance of each estimated regression coefficient. The method is widely applicable and can be used to obtain estimates of sub-class means, regression coefficients, etc. Even though the number of items administered to each subject might be small, the plausible values / multiple imputation approach will yield consistent estimates provided certain conditions are satisfied. A key condition is that the predictive distribution from which the plausible values are drawn must be conditioned on the variables featured in the aggregate analysis. Let us denote the required predictive distribution as $P(\boldsymbol{\eta} | \mathbf{x}, \mathbf{Y})$, where $\mathbf{x}$ represents item response patterns as before, and $\mathbf{Y}$ represents other collateral information available on all respondents. In theoretical terms, the plausible values approach obtains an estimate of an aggregate statistic, $t(\boldsymbol{\eta}, \mathbf{Y})$, via its conditional expectation,

$$E[t(\boldsymbol{\eta}, \mathbf{Y}) | \mathbf{x}, \mathbf{Y}] = \int t(\boldsymbol{\eta}, \mathbf{Y}) p(\boldsymbol{\eta} | \mathbf{x}, \mathbf{Y}) d\boldsymbol{\eta} \quad (5.8)$$

which is evaluated by averaging over the multiple imputations. It is clear from equation (5.8) that all variables included in the $\mathbf{Y}$ of $t(\boldsymbol{\eta} | \mathbf{Y})$ must also be included in the $\mathbf{Y}$ on which the posterior $P(\boldsymbol{\eta} | \mathbf{x}, \mathbf{Y})$ is conditioned. This is difficult to ensure in practice, particularly when the data set on which the plausible values are recorded is made widely available for secondary research.

In the NAEP implementation, approximately 300 collateral variables are included in the conditioning, and efforts of a similar magnitude are undertaken for the Program for International Student Assessment (PISA) survey, in which Statistics Canada participates. Nevertheless, it is never possible to guarantee that all possibilities are covered. With 300 variables, for example, there is no way that all interaction effects can be anticipated. In such cases, one must simply hope that the amount of new information contributed by unanticipated aggregating variables is small and that, as a result, the corresponding bias is small. Clearly, this method is not feasible for the individual researcher. It is feasible, however, for large statistical agencies that provide data of interest to business researchers and it also may be feasible for the large research companies involved in large scale business research projects.

5.3.4 Structural Equation Modeling (SEM): Simultaneous Estimation of Item and Regression Parameters with Discrete Indicators

The work of Adams et al. (1997) that uses the ordinal partial credit IRT model, and the earlier IRT work of Mislevy (1987) for dichotomous indicator variables, are methods whereby the IRT model item parameters and the regression parameters are simultaneously (and consistently) estimated, bypassing the need to predict latent variable scores, and thus avoiding finite item bias. Similar simultaneous modeling of the regression parameters was also undertaken by Zwinderman (1991) in the context of a Rasch model. These methods were developed to improve precision of estimation by incorporating collateral information via equation (5.7), which features a latent response variable, but predictor variables measured without error. Thus they do not provide a general approach to analyzing latent variable regression models.

Nevertheless, general methods do exist for obtaining consistent parameter estimates of systems of linear latent variable equations with both continuous and discrete indicator variables (see section 5.2.1 for the special case of the general linear structural relations model). Methods for handling latent variable systems with discrete indicators can be used to simultaneously estimate IRT item parameters and the parameters of the latent regression models, as will be described in the next chapter. The terminology associated with systems of linear latent variable equation is somewhat confusing. From a theoretical viewpoint, a research hypothesis can be specified in terms of a system of linear equations, featuring latent dependent and/or independent variables. Such systems are referred to as

latent variable structural equations. The study of such systems, which includes the estimation of the regression parameters or structural parameters, is referred to as *structural equation modeling* (SEM). Strictly speaking, the SEM terminology should include all methods of estimating the structural parameters. However, the convention now appears to be to associate SEM only with estimation methods, such as “covariance-structural” analyses that yield consistent estimates for large samples (Chin, 1998). This convention will be followed in this thesis. There are several statistical packages that implement such estimation methodologies, the best known being LISREL, EQS, AMOS, Mplus, and LINCOS. All of these packages can also perform discrete SEM, i.e., SEM analysis based on discrete indicators, though the most flexible in this regard appears to be Mplus. Note that estimation methods such as discrete covariance structure analysis are akin to the MML methods described previously for IRT in that they operate on parameters of marginal distributions obtained by averaging over (i.e., integrating out) the latent variables. This essentially is what guarantees consistent estimates.

As noted in the introduction, the use of discrete SEM techniques in this thesis is motivated by purely technical concerns of estimation, primarily bias reduction. Though SEM (primarily for continuous indicators) has become very popular in business research (see reviews by Baumgartner & Homburg, 1996; Shook, Ketchen, Hult, & Kacmar, 2004; Steenkamp & Baumgartner, 2000), a discussion of SEM in the context of general modeling philosophy is beyond the scope of this thesis. It is noted, however, that some researchers have criticized the value and the use of SEM in empirical research, and a discussion on the limitations of SEM can be found in Rigdon (1998, p.p.277-279), for example. For a thorough discussion on the pros and cons of SEM from both theoretical

and practical viewpoints, please refer to the Journal of Educational Statistics' special issue on SEM in 1987. For interested readers, the use of SEM in specific areas of business research has been addressed by Brannick (1995) in organizational behaviour, Chin (1995, 1998) in management information system, Steenkamp & Baumgartner (2000) in marketing, and Shook, et al. (2004) in strategic management.

5.3.5 Partial Least Squares: Continuous Indicators

There is an alternative estimation technique known as *partial least squares* (PLS) which is also used by researchers to analyze systems of latent variable structural equations, when the indicator variables are continuous or have sufficient number of categories that the latent variables can be approximated by continuous variables. PLS has been used in many business disciplines, some examples being Arnett, Laverie, & Meiers (2003), Auh, Salisbury, & Johnson (2003), and Smith & Bristor (1994) in marketing, Duxbury & Higgins (1991) and Kahai, Sosik, & Avolio (2004) in organizational behaviour, Ravichandran & Rai (2000) and Wixon & Watson (2001) in management information systems, and Birkinshaw, Morrison, & Hulland (1998) and Johansson & Yip (1994) in strategic management.

PLS is based on a very different methodology than SEM, and is closer in concept to the two-step methods discussed previously. PLS has become popular for a number of reasons, in particular because of its apparent freedom from distributional assumptions. PLS is also advocated because it can yield estimates even when the samples are so small that covariance-based methods do not converge (e.g., Chin, 1998). The main difference between PLS and covariance-based SEM approaches is that latent variable scores are computed in PLS as intermediate steps in an iterative least squares process. However, PLS parameter estimates do suffer from finite item bias, i.e., they produce inconsistent estimates of the structural regression parameters (Schneeweiss, 1993). PLS estimates approach the true values only as the number of items and the sample size increase. Though, this limiting property, termed *consistency at large* (Wold, 1982, p.25), has been

reported in the applied literature (see, for example, Barclay, Higgins, & Thompson, 1995; Chin, 1998), it appears not to be well known by practitioners. For example, in his commentary on the use of PLS, Hulland (1999) urges researchers to understand its assumptions and limitations, but omits mention of the critical property of “consistency at large”. Thus, it is relevant to explore the extent of the finite item bias in PLS before proposing it as a solution to accounting for measurement error.

Some of the benefits of PLS were alluded to above. It should also be noted that PLS has its detractors, as evidenced by the following quote from McDonald (1996):

The PLS methods are difficult to describe and extremely difficult to evaluate, partly because PLS constitutes a set of ad hoc algorithms that have generally not been formally analyzed, or shown to possess any clear global optimizing properties (except in the well understood case of just two composites), and partly because these devices are represented as a form of path analysis with latent variables, and it can be difficult to determine what properties of latent variable models they possess, if any. (p.240)

Clearly, the relative merits of SEM, PLS and other modeling approaches are controversial, and again, we note that this thesis is not oriented towards a philosophical discussion of their relative merits. This thesis is focused only on the properties of various methods of estimating regression coefficients in a latent structural equation, when the indicators are discrete.

5.4 Summary

Most recently, Bollen & Lennox (1991), Fox & Jones (1998), Shevlin, et al., (1997) have pointed out that many research studies involve the two-step approach: predict latent variable scores first, then use these scores in subsequent analyses. From Baumgartner & Homburg's (1996) review of SEM applications in four major marketing journals between 1977 and 1994, it can be seen that at least 63 percent of the studies that examined relationships among latent variables in a multi-equation system used composite scores for at least one latent variable in the model. Another more recent study by Holbert (2002) reviewed the use of SEM from 1995-2000 in 37 communication-related journals. They reported that about 58 percent of the 118 multi-equation models examining latent variable relationships employed single-indicators or composite scores. Note also that these two reviews did not include studies which used the two-step approach in single equation models, like simple or multiple regressions. Thus, they in fact understate the prevalence of the two-step approach. However, regardless of the type of scoring technique used, CTT-, IRT-, or FA-based, analyses based on latent variables scores are generally subject to finite item bias, and thus any set of scores should be used with caution.

In a simple regression analysis, the consequence of ignoring measurement error (i.e., directly using an observed single-indicator measure or composite score from a multiple scale and treating it as true latent variable value in analysis) is the attenuation of parameter estimates; but in multiple regression or multi-equation systems, ignoring measurement error can lead both to attenuated and inflated parameter estimates. Since it

is difficult to predict the consequence of measurement error, the need to take into account the measurement error in modeling becomes extremely important.

There are some methods available for correcting for finite item bias, though these do not constitute generally applicable solutions, particularly for discrete indicator variables. There are also methods for removing finite item bias in the case of discrete indicators by avoiding the prediction of scores, such as the plausible values approach and the use of discrete SEM. The plausible values approach for discrete indicators is not practical for any but the largest research organizations, while both continuous and discrete SEM based methods, on the other hand, are now readily available through several commercially available software packages. Clearly, the relative merits of SEM, PLS and other modeling approaches are controversial, and again, we note that this thesis is not oriented towards a philosophical discussion of their relative merits. Instead, this thesis is focused on the properties of various methods of estimating regression coefficients in a latent structural equation, when the indicators are discrete.

Since IRT methods comprise a model-based approach to scale construction, and offer several advantages over the CTT approach, it is natural to enquire if the use of IRT scores will also result in lower finite item bias than CTT scores when they are used in subsequent regression analysis. If these advantages do not carry over when IRT scores are used in a two-step analysis, it will then be relevant to enquire if and how the IRT approach can be extended to allow for bias free regression analysis. Despite the long history of IRT, however, this issue has received relatively little empirical investigation, and, as a result, very little is known. This issue must be addressed before suggestions for the use of IRT in business research (e.g., McCollam, 1998; Singh, 2004) can be

supported. Therefore, a simulation investigation of finite item biases in regression parameters caused by using IRT- and CTT-based scores will be provided later in this thesis.

Approaches to estimating regression coefficients in latent regression models, and literature reports of empirical studies on the use of predicted latent variable scores in regression, are summarized in Table 5.1 and Table 5.2.

Table 5.1: Approaches to estimating regression coefficients in latent regression model

Data Type	Predict latent variable scores, then run OLS regression	Sidestep prediction of latent variable scores
Continuous Indicators	<ul style="list-style-type: none"> • Total scores • Factor scores: Regression, Bartlett methods, and other methods (see Tucker, 1971) • Jöreskog / LISREL 8.6 scores • PLS 	<ul style="list-style-type: none"> • Plausible values methodology • SEM: Simultaneous estimation of the measurement model parameters and latent regression parameters
Discrete Indicators	<ul style="list-style-type: none"> • Total scores • IRT scores (MLE, WLE, MAP, and EAP) 	<ul style="list-style-type: none"> • Simultaneous IRT/SEM: Simultaneous estimation of IRT item parameters and latent regression parameters [to be described in Chapter 6] • Constrained IRT/SEM: As above, but with IRT item parameters constrained at known or pre-estimated values [to be described in Chapter 6]

Table 5.2: Summary of the studies on the use of predicted latent variable scores in regression

Study	Findings
I. Continuous Scale Items	
Tucker (1971) examined regression of a measured dependent variable on one or more factor scores	Only 'regression' method produced consistent regression parameter estimate
Shevlin, Miles, & Bunting (1997) examined a setup similar to Tucker's (1971) but with 'scores' computed as sums of indicator variables associated with each factor	Biased estimate, more so as true factor loadings decreased
Skrondal & Laake (2001) studied <i>blockwise factor score regression</i> , which computes separate factor scores from the two confirmatory FA measurement models, and then runs OLS regression	Inconsistent estimate, exception when Bartlett/Regression method is used, concludes: "conventional factor score regression performs badly and should definitely be abandoned"
Croon (2002) theoretically showed how the variances and covariances of predicted factor scores, and the covariances between predicted factor scores and measured variables can be corrected to obtain consistent estimates of regression parameters	<ul style="list-style-type: none"> - Consistent estimates of regression parameters - Difficult to implement as a general technique
Jöreskog (2000) used Anderson & Rubin method (1956) to generate scores that are constrained to reproduce the full covariance matrix of the latent response and explanatory variables	Consistent estimates of regression parameters
II. Discrete Scale Items	
Adams, Wilson, & Wu (1997) examined $\eta = Y\beta + \zeta$ and used discrete partial credit model incorporating collateral information Y to estimate β and then regressed EAP scores of η on Y	<ul style="list-style-type: none"> - Attenuated R-square and regression coefficients - Not a general solution
Hojtink & Boomsma (1996) used bias corrected WLE estimators for regression and ANOVA	<ul style="list-style-type: none"> - Still required items at least 15 for acceptable parameter bias - Difficult to apply as a general technique

Chapter 6

Structural Equation Modeling (SEM)

From the mid 1970's on, extensive work has been carried out by numerous authors (e.g., Christopherson, 1975; Muthén, 1983, 1984) to extend the techniques of SEM to handle discrete indicator variables. The methodology currently implemented in programs such as Mplus (Muthén & Muthén, 2001) allows for the analysis of mixtures of discrete and continuous indicator variables, and of combinations of continuous and categorical latent variables, and includes robust procedures (against deviations from normality) for estimating standard errors and fit statistics that do not rely on multivariate normal indicators. Recall that we use the term SEM not only to refer to the system of latent variable structural equations, but also to the estimation methodology that provides consistent parameter estimates.

Discrete SEM is very general and subsumes IRT modeling and the simultaneous estimation of both IRT item parameters and the parameters of latent regression models. The IRT connection to general SEM appears not to be as well known as the corresponding connection between factor measurement models and SEM in the continuous case, despite the fact that the IRT connection was very clearly spelled out by Takane and de Leeuw (1987). This connection was used by Muthén (1988) to relate IRT

models to external variables such as grouping indicators as well as to continuous background variables, and was subsequently extended by Muthén, Kao and Burstein (1991).

This chapter begins by reviewing SEM-based latent regression models with continuous indicators, and then introduces discrete SEM models with both dichotomous and polytomous indicators. The connection between IRT and SEM is then described.

6.1 SEM-Based Latent Regression Models with Continuous Indicators

The structural SEM model studied in this thesis is given by

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \quad (6.1)$$

where $\boldsymbol{\eta}$ is a vector of latent response variables, $\boldsymbol{\xi}$ is a vector of latent explanatory variables, $\boldsymbol{\Gamma}$ is a matrix of regression coefficients and $\boldsymbol{\zeta}$ represents a vector of disturbances independent of $\boldsymbol{\xi}$, with covariance matrix $\boldsymbol{\Psi}$. The model of (6.1) is a special case of the general linear structural relations model (LISREL) of Jöreskog & Sörbom (2001) with $\mathbf{B}=\mathbf{0}$. Hence, a latent variable is either an explanatory or a response variable, but never an intervening variable. The complete specification of system (6.1) requires the specification of factor measurement models for $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, namely,

$$\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \quad \text{and} \quad \mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad (6.2)$$

where $\mathbf{x}$ and $\mathbf{y}$ are themselves observable indicator variables consisting of vectors of items (or indicator variables) of dimensions $\mathbf{p}$ and $\mathbf{q}$, respectively, and the “loadings” $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Lambda}_y$ comprise matrices of fixed parameters. The vectors $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ contain random disturbances that are independent of the latent variables and of each other.

Consistent estimates of the regression parameters, $\boldsymbol{\Gamma}$, can be obtained by maximum likelihood using a number of computer programs, e.g., LISREL, EQS, AMOS, Mplus, LINCOS. Robust estimation methods are also available in a number of these packages.

6.2 SEM-Based Latent Regression Models with Dichotomous Indicators

The structural discrete SEM model with dichotomous indicators is again described by equation (6.1). The detailed measurement models corresponding to equation (6.2) now have to be modified, however, i.e.,

$$\mathbf{x}^* = \Lambda_x \boldsymbol{\xi} + \boldsymbol{\delta} \quad \text{and} \quad \mathbf{y}^* = \Lambda_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad (6.3)$$

where $\mathbf{x}^*$ and $\mathbf{y}^*$ are now unobservable variables consisting of vectors of items (or indicator variables) of dimension $p \times 1$ and $q \times 1$ respectively, and the “loadings” Λ_x and Λ_y comprise matrices of fixed parameters. The vectors $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ contain random disturbances that are independent of the latent variables and of each other. The difference between the measurement models (6.2) and (6.3) is that $\mathbf{x}$ and $\mathbf{y}$ in the former represent observable item indicators whereas the $\mathbf{x}^*$ and $\mathbf{y}^*$ in the latter are unobservable. For the single population model discussed here, the means of $\mathbf{x}^*$ and $\mathbf{y}^*$ can be set to zero. Since $\mathbf{x}^*$ and $\mathbf{y}^*$ cannot be observed, their variance is arbitrary and usually set to one for convenience. Observable discrete indicator variables, denoted $\mathbf{x}$ and $\mathbf{y}$, are then modeled as

$$\begin{cases} x_j = 1 \text{ if } x_j^* \geq \tau_j, & x_j = 0 \text{ otherwise, } & j = 1, \dots, p \\ y_j = 1 \text{ if } y_j^* \geq \tau_j, & y_j = 0 \text{ otherwise, } & j = p+1, \dots, p+q \end{cases}, \quad (6.4)$$

where the τ 's represent thresholds that must be estimated from the data, and p and q are the number of indicator variables, or items, in the measurement models for η and ξ , respectively. The continuous x^* and y^* can be thought of as latent response variables underlying each x and y . In other words, the observed responses, x and y , are assumed to reflect the latent continuous variables x^* and y^* . It is usually assumed that x^* and y^* are normal.

Maximum likelihood estimation of the parameters of the above model is difficult as it involves integration of $(p + q)$ correlated multivariate normal variables over $(p + q)$ dimensional space defined by the thresholds. For this reason, a *limited information generalized least squares* approach to estimation is used, based on one-way and two-way marginal probabilities of the discrete indicator variables. Consistency of the parameter estimates has been established (Browne & Arminger, 1995; Christopherson, 1975; Muthén, 1984).

6.2.1 The connection between IRT and SEM for Dichotomous Indicators

Takane and de Leeuw (1987) showed that a discrete measurement model of the form given in equation (6.4) is formally equivalent to the normal ogive IRT model with two item parameters given in equation (3.2), which can be re-expressed as follows:

$$P_j(\eta) = \int_{-\infty}^{a_j\eta + d_j} \phi(z) dz = \Phi(a_j\eta + d_j), \quad (6.5)$$

where ϕ is the density function of the standard normal distribution, Φ is the normal ogive function (i.e., the cumulative distribution function of the standard normal distribution), $d_j = -a_j \times b_j$. Recall that a_j is the discrimination parameter and b_j is the difficulty parameter for item j . In particular, for a discrete measurement model of the form $\mathbf{y}^* = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, with $\text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$, they showed that the item parameters a_j and d_j of equation (6.4) can be expressed in terms of the parameters of the discrete measurement model as

$$a_j = \lambda_j / \theta_j^{1/2} \quad \text{and} \quad d_j = -\tau_j / \theta_j^{1/2}, \quad (6.6)$$

where λ_j is the j^{th} row of $\mathbf{\Lambda}$, and θ_j is the j^{th} diagonal element of $\boldsymbol{\Theta}$. Thus estimation of the parameters of the discrete SEM defined by equations (6.1), (6.3) and (6.4) can be regarded as a simultaneous and consistent estimation of IRT item parameters and the parameters of a latent regression model. In other words, the IRT model is automatically embedded in the structural equation model. Furthermore, the two-parameter normal

ogive IRT model can be considered as a special unidimensional case of the general item factor analysis model (McDonald, 1967; Takane & de Leeuw, 1987). It is noteworthy that the above relationship holds approximately with the 2-parameter logistic IRT model to the extent that the logistic distribution provides a good approximation to the normal distribution (Takane & de Leeuw, 1987).

Whenever the IRT item parameters a_j and b_j are known, the corresponding measurement model parameters, λ_j and τ_j , can be calculated and held fixed during the SEM estimation. This will be referred to as the *constrained* IRT-SEM approach to distinguish it from the case where both IRT and structural parameters are simultaneously estimated. The constrained IRT-SEM approach will be useful when a set of pre-calibrated IRT items are used in the study, so that only the structural regression parameters need to be estimated from the research data. The constrained IRT-SEM approach will then again provide consistent estimates of the structural equation parameters, and it is expected that convergence of the regression estimates will be faster under constrained than simultaneous IRT-SEM because a smaller number of fixed parameters will be being estimated. Thus, for a given finite sample size N , it is expected that constrained IRT-SEM will yield a smaller finite sample bias than simultaneous IRT-SEM.

6.3 SEM-Based Latent Regression Models with Polytomous Indicators

A general SEM model with ordered categorical indicators is an extension of the SEM model with dichotomous indicators described in Section 6.2. Suppose a respondent's response to an item can be categorized into one of $(m + 1)$ ordered categories. For an ordered polytomous response x_j with $(m + 1)$ categories on the j^{th} item, and $x_j = 0, 1, \dots, m$, we assume the monotonic relationship as follows:

$$x_j = \begin{cases} m, & \text{if } \tau_{j,m} < x_j^* \\ m-1, & \text{if } \tau_{j,m-1} < x_j^* \leq \tau_{j,m} \\ \dots & \\ 1, & \text{if } \tau_{j,1} < x_j^* \leq \tau_{j,2} \\ 0, & \text{if } x_j^* \leq \tau_{j,1} \end{cases} . \quad (6.7)$$

For an ordered polytomous response y_j with $(m + 1)$ categories on the j^{th} item, we similarly assume:

$$y_j = \begin{cases} m, & \text{if } \tau_{j,m} < y_j^* \\ m-1, & \text{if } \tau_{j,m-1} < y_j^* \leq \tau_{j,m} \\ \dots & \\ 1, & \text{if } \tau_{j,1} < y_j^* \leq \tau_{j,2} \\ 0, & \text{if } y_j^* \leq \tau_{j,1} \end{cases} . \quad (6.8)$$

6.3.1 The connection between IRT and SEM for Polytomous Indicators

In Section 6.2.1, we discussed the relationship between IRT and SEM models for dichotomous data. An analogous relationship holds for general ordered categorical data. Takane & de Leeuw (1987) show that Samejima's (1969) normal ogive graded response IRT model is equivalent to a factor analytic model of ordered categorical data proposed by Muthén (1984). That is, an ordered polytomous response measurement model of the form given in Equation (6.7) or (6.8) is formally equivalent to the graded response model given in Equation (3.6), which can be rewritten as,

$$P_{x_j}^*(\eta) = P_{x_j}(\eta) - P_{(x+1)_j}(\eta) = \Phi[a_j\eta + d_{x_j}] - \Phi[a_j\eta + d_{(x+1)_j}], \quad (6.9)$$

where $P_{x_j}^*(\eta)$ is the probability of endorsing option x on item j for a randomly sampled respondent with ability η , and $P_{x_j}(\eta)$ is the probability of endorsing option x or higher, while $P_{(x+1)_j}(\eta)$ is the probability of endorsing the next highest option $(x + 1)$ or higher, and $d_{x_j} = -a_j \times b_{x_j}$. They showed that the item parameters a_j and d_{x_j} of equation (6.9) can be expressed in terms of the parameters of the discrete measurement model as

$$a_j = \lambda_j / \theta_j^{1/2} \quad \text{and} \quad d_{x_j} = -\tau_{x_j} / \theta_j^{1/2}. \quad (6.10)$$

As in the dichotomous case, an approximate relationship holds between polytomous factor analytic model and the 2-parameter logistic IRT model for ordered categorical data

to the extent that the logistic distribution provides a good approximation to a normal distribution (Takane & de Leeuw, 1987).

In this polytomous IRT-SEM model, if the item parameters are held constant (i.e., fixed at random values) then this is also called the *constrained* IRT-SEM approach to distinguish it from the case where both IRT and structural parameters are simultaneously estimated.

6.4 Summary

This chapter reviews two types of structural equation modeling (SEM) approaches: standard full-information SEM for analyzing continuous indicators and IRT-SEM for analyzing discrete (binary and ordinal) indicators. Unlike two-step approach, SEM methods produce consistent estimates and do not suffer from finite item bias. A variant of IRT-SEM, namely the constrained IRT-SEM approach, is also introduced, which again provides consistent estimates of the structural equation parameters. For a given finite sample size N , it is expected that constrained IRT-SEM will yield a smaller finite sample bias than simultaneous IRT-SEM.

Chapter 7

Research Questions

This chapter summarizes the motivations and justifications for exploring the research questions that will be examined in the next chapter.

Researchers in business frequently measure latent variables by using a multiple-item scale constructed through CTT procedures. Despite its common acceptance in behavioural research, however, CTT-based procedures of scale construction have many well-documented shortcomings. In an attempt to remedy these shortcomings, IRT has been developed and has become the scale construction method of choice in psychometrics. IRT methods are applicable to binary and polytomous items and have also been recommended for use in business by McCollam (1998) and Singh (2004). Given this trend, it is natural to enquire whether or not from a quantitative perspective IRT methods yield better measures than does CTT when the goal of measurement is to obtain latent variable scores for individuals.

After examining the empirical and theoretical studies on the relative performance of IRT- and CTT-based scores, it is somewhat surprising to find that for both dichotomously and polytomously scored responses, the latent variable scores based on the two measurement frameworks are highly comparable. Nevertheless, the literature also

features some gaps in the coverage of the various design characteristics of scales or tests. For example, how do IRT and CTT scores compare for the typical scale lengths and item categories used in business research? How close are the IRT and CTT scores to the true latent variable value, when the binary and polytomous manifest variables display appreciable degrees of skewness? These questions require more detailed empirical scrutiny. Therefore, the first research question for the study is:

Research Question 1.

How comparable and how accurate are the CTT- and IRT-based scores?

This question will be examined under various scale design characteristics.

The study of the relationships among latent variables is also very common in business research, where many researchers use a *two-step approach*, i.e., obtain latent variable predicted scores and then use these scores directly in subsequent analyses such as regression. While this approach is very common (Baumgartner & Homburg, 1996; Fox & Jones, 1998; Holbert, 2002; Osgood, et al., 2002), it has serious drawbacks, regardless of the individual optimality of the measurement technique used to derive the scores. The major drawback is that analyses based on the predicted scores are subject to finite item bias, regardless of sample size. Nevertheless, given some of the potential advantages of IRT over CTT, one might expect there to be an appreciable reduction in finite item bias from using IRT-based versus CTT-based scores. Thus, this leads to an interesting question: are there appreciable differences in the finite item bias in parameter estimates caused by using IRT-based versus CTT-based scores in two-step analyses? In other words, is using IRT-based scores more profitable than using CTT-based scores in standard statistical analyses? Little attention has been focused on this issue in the

psychometric or other literature. The empirical silence on this issue is an anomaly which must be removed before suggestions for the use of IRT in business research can be supported.

Another approach to studying relationships among latent variables for discrete indicators is the IRT-SEM approach, which is different from the aforementioned two-step approach in that it simultaneously estimates the IRT item parameters in the latent variable measurement models *and* estimates the relationships between the latent variables, while bypassing the prediction of latent variable scores altogether. In other words, the IRT model is automatically embedded in the SEM model. We call this approach *simultaneous IRT-SEM* estimation, which yields consistent estimates of all parameters, that is, it avoids the problem of finite item bias even when the individual items are measured discretely (Lu, et al., 2003; Lu, et al, 2004).

Another possible approach can be used whenever the parameters of an IRT scale are known a priori. This situation occurs if an IRT scale has been calibrated using a large normative sample and made available for use in subsequent research. In such a situation, the IRT parameters can be translated into discrete SEM parameters and treated as fixed in the IRT-SEM estimation. It is hoped that this approach, referred to as *constrained IRT-SEM*, might perform better than the simultaneous IRT-SEM for small samples, since fewer parameters are involved in the estimation. Therefore, another goal of the research will be to see how well the theoretically consistent IRT-SEM estimates perform in practice, i.e., to examine the extent of the finite sample bias and to compare its magnitude to that of the finite item bias of the IRT (and CTT) based two-step methods. This leads to the following research question:

Research Question 2.

Are there appreciable differences in the bias of regression parameter estimates for the following four methods?

1. CTT two-step approach: using CTT-based scores in regression analysis
2. IRT two-step approach: using IRT-based scores in regression analysis
3. Simultaneous IRT-SEM approach
4. Constrained IRT-SEM approach

Finally, another technique known as *partial least squares* (PLS) has been used by business researchers to estimate the parameters of a structural model. PLS parameter estimates, however, are not consistent as the PLS technique generates latent variable scores as an intermediate step and so suffers from finite item bias unless the number of items is very large (Schneeweiss, 1993). Thus, another goal of the research will be to carry out a preliminary comparison of PLS to the two-step and SEM-based approaches, in order to assess the extent of finite item bias in PLS. The final research question then becomes:

Research Question 3.

How does the finite item bias of PLS compare with that of the two-step and IRT-SEM approaches?

It is important to emphasize why we need to be concerned with *bias*. Specifically speaking, *bias* is a systematic difference between the estimated value of a parameter and the true value of the parameter itself, i.e., bias is systematic error. As such, a particular estimator may be biased (e.g., the sample median tends to *overestimate* the true population median), or on a larger scale, a certain methodology may also be considered biased (e.g., using CTT-based scores in simple latent regression tends to *underestimate* the true latent parameters). Of course, we can not expect every instance of an estimation to match the true parameter value precisely, but it is very problematic to employ an estimator or methodology that consistently under- or over-estimates. When there is no bias, the true parameter value will begin to emerge over repeated studies; when there is bias, however, the true parameter value will never be discovered no matter how often the study is replicated. Since a primary goal in data analysis is to estimate or describe population characteristics, not simply to decide whether or not some coefficient is different from zero, it is very important that our analysis produce estimates that are free of bias.

Chapter 8

Study Design and Implementation

A series of Monte Carlo (computer simulation) experiments were carried out to answer the aforementioned research questions, and to quantify the effects on estimation bias of a number of model conditions (or “factors”). In other words, we “simulate” or create different conditions (e.g., data) to see how a model or methodology will “perform” under those situations.

Normally, the asymptotic properties of an estimator are known to us to some degree at least, but not its finite sampling properties. Monte Carlo simulation is a technique which allows us to assess research questions that cannot be satisfactorily answered through asymptotic theory alone. Asymptotic properties of an estimator rely on strict conditions such as large sample size or ideal distributional assumptions for variables. If a researcher is working with large sample size (say, 1,000) and normally distributed variables, asymptotic theory provides useful information about the sampling distribution of the estimator; on the other hand, if a researcher is working with small sample size (say, 100) and highly skewed variables, Monte Carlo technique may be the only way to assess the finite sampling performance of an estimator (Paxton, Curran, Bollen, Kirby, & Chen, 2001). Thus, a major advantage of a Monte Carlo experiment is the ability to observe

the behaviour of a given statistic, e.g., the latent regression parameter estimates in our study, across random samples by manipulating experimental factors that would normally be inaccessible or extremely expensive to manipulate via any other means of analysis (e.g., MacDonald & Paunonen, 2002; Stephenson & Holbert, 2003). This technique is seen as "far superior to the use of a single or small number of data sets" (Stephenson & Holbert, 2003, p.312).

Two structural regression models were selected for the simulation study. The first model is a simple structural regression model consisting of one latent response variable, and one latent explanatory variable, i.e., $\eta = \gamma \xi + \zeta$. The second model then is a multiple structural regression model with one latent response variable, and two latent explanatory variables, i.e., $\eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \zeta$. The simulations were conducted for discrete observations (both dichotomous and ordinal polytomous). The specific values of the population model parameters were selected to be representative from an applied standpoint.

This chapter outlines detailed simulation designs and implementation procedures. Table 8.1 provides a summary of the simulations for the research questions. The path diagrams of simulated simple and multiple latent regression models are depicted in Figures 8.1 and 8.2.

Table 8.1: Summary of the simulations for the research questions

<div>Model</div> <div>Data Type</div>	<div>Simple latent regression: $\eta = \gamma \zeta + \zeta$</div> <div>&</div> <div>Multiple latent regression: $\eta = \gamma_1 \zeta_1 + \gamma_2 \zeta_2 + \zeta$</div>
<div>Discrete (binary & ordinal)</div>	<p><i>Research Question 1.</i> How comparable and how accurate are the CTT-based and IRT-based scores?</p> <p><i>Research Question 2.</i> Would there be appreciable differences in the bias of the regression parameter estimates for the following four methods?</p> <ol style="list-style-type: none"> 1. Two-step approach: using CTT-based scores in regression analysis 2. Two-step approach: using IRT-based scores in regression analysis 3. Simultaneous IRT-SEM approach 4. Constrained IRT-SEM approach <p><i>Research Question 3.</i> How does the finite item bias of PLS compare with that of the two-step and IRT-SEM approaches?</p>

Figure 8.1: Path diagrams of the simulated simple latent regression model

A. Simulated simple latent regression model: $\eta = \gamma\xi + \zeta$

B. Two-step approach

C. IRT-SEM approach (discrete SEM)

Note: x_p^* and y_q^* are the normally distributed continuous latent variables underlying the observable discrete indicators x_p and y_q .

Figure 8.1: Path diagrams of the simulated simple latent regression model (contd.)

Note: 1. x_p and y_q are the discrete observable indicators, to be analyzed as continuous variables here. 2. Dotted lines indicate successive iterations.

Figure 8.2: Path diagram of the simulated multiple latent regression model

Simulated multiple latent regression model: $\eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \zeta$

Note: All procedures used with this multiple latent regression model are analogous extensions of the procedures used with the simple regression model.

8.1 Design of the Simulation Study for a Simple Latent Regression Model

This section describes the detailed simulation procedures for exploring each of the research questions under the simple latent regression model.

8.1.1 Simulation Procedures for Assessing Comparability and Accuracy of IRT- and CTT-based Scores

Simulating Discrete Item Responses. The process of generating discrete item response data under IRT and CTT measurement models for the simulation involved four stages, which were programmed using procedure IML in the software package SAS 8.2 (summarized in Table 8.2):

Stage 1. Generating two vectors of true latent variable values of η and ξ for N simulated respondents, corresponding to the preset R^2 for the structural regression model. The variances of the latent variables η and ξ were both set to one. Two vectors of realizations of the latent explanatory variable ξ and the disturbance variable ζ were first generated from a standard normal distribution $N(0,1)$ and a normal distribution $N(0, 1 - R^2)$, respectively, corresponding to a coefficient of determination, R^2 , for the structural regression model, which implies a regression coefficient γ of $\sqrt{R^2}$. Note that the terminology R^2 is used here for reasons of familiarity, although, strictly speaking, it refers to a population or model parameter in this study. A vector of values of η was then computed via $\eta = \gamma \xi + \zeta$.

Stage 2. *Generating the multivariate normal item responses, y_j^* and x_j^* , to correspond to the structural and measurement models.* The measurement models for η and ξ were $y_j^* = \lambda_j \eta + \varepsilon_j$ and $x_j^* = \lambda_j \xi + \delta_j$, respectively, where $j = 1, \dots, n$, with n being the number of items in each scale. All the items were designed to load only on a particular factor; that is, each item only measures one latent variable. Loading parameters were made identical for both measurement models (except where indicated otherwise in the tables), and variances of the individual y_j^* and x_j^* were set equal to one. The measurement errors, ε_j and δ_j , were generated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\theta}_\varepsilon)$ and $N(\mathbf{0}, \boldsymbol{\theta}_\delta)$, respectively, where the diagonal matrices $\boldsymbol{\theta}_\varepsilon$ and $\boldsymbol{\theta}_\delta$ were made the same, yielding the loadings $\lambda_j = \sqrt{1 - \theta_j}$, with corresponding coefficients of determination (CD) for the measurement models equal to λ_j^2 . The column vectors of y_j^* and x_j^* were then determined via the measurement model equations.

Stage 3. *Creating category thresholds within each item, used for truncating y_j^* and x_j^* into discrete response data, y_j and x_j , in the final stage.* Three alternative marginal frequency distributions for the response data, y_j and x_j , were examined: a symmetric distribution, a uniform distribution, and a skewed distribution, as shown in Figure 8.3. For skewed distributions, the level of skew ranged from a low level of approximately one (0.99-1.05) to a high level of approximately 2 (2.0-2.1), with kurtosis level varying from -0.17 to 0.39, and 2.0 to 3.8, respectively. The marginal frequency distributions of y_j and x_j can be obtained from the normal distributions of y_j^* and x_j^* by appropriate choice

of thresholds (see Figure 8.4 for illustration of threshold creation). For example, for a symmetric three frequency distribution (22%, 56%, 22%), the thresholds would be -0.775 and 0.775. The skewed marginal frequency distributions of y_j and x_j can also be obtained from the normal distributions of y_j^* and x_j^* using this strategy. For example, for a three category skewed marginal frequency distribution (80%, 15%, 5%), the thresholds would be 0.842 and 1.645. A set of thresholds was designed to be identical for y_j^* and x_j^* .

Stage 4. *Generating the discrete (binary and ordinal) item response data, y_j and x_j .* This was carried out by categorizing the multivariate item response data, y_j^* and x_j^* , into different number of categories, $m+1$, using the thresholds created from the previous stage and a natural extension of the system exhibited in equation (6.4) for binary data, or equations (6.7) and (6.8) for ordinal data. The categorization of the underlying item response variables into discrete item indicators is illustrated in Figure 8.4.

Figure 8.3: Marginal frequency distributions of discrete indicator variables

Note: The height bars represent percentages.

Figure 8.4: Illustrations for threshold creation and categorization of underlying response variables into discrete item indicators

Generating IRT- and CTT-based Scores. EAP predicted latent variable scores were calculated for each of the respondents from the simulated discrete response data, based on the normal ogive IRT model corresponding to the measurement models used in Stage 2 and 3 above (also see Equation 6.6 and 6.10). In other words, here the IRT item parameters were treated as known and fixed. However, the work of Kim and Nicewander (1993) suggests that results obtained using fixed parameters will not differ greatly from results based on estimated parameters, for the sample sizes considered here. The EAP latent variable scores were calculated using 24 point Gauss-Hermite integration of the posterior latent distribution. The standardized total scores were also calculated from the discrete response data for each respondent. The above processes for generating IRT- and CTT-based predicted scores were programmed using procedure IML in SAS 8.2.

Assessing Comparability. Both Pearson and Spearman rank correlations were used to assess the comparability of the IRT- and CTT-based scores obtained from the same simulated data. Specifically, Pearson correlation is employed to measure the linear relationship between the IRT- and CTT-based scores, while Spearman rank correlation is a measure of association between the respondents' ranking on the IRT- and CTT-based scores.

Assessing Accuracy. Note that the true latent variable values for each simulated individual are known in the study; this allows us to assess the accuracy of the predicted scores. The squared Pearson correlation between the predicted latent variable scores (obtained with both IRT and CTT) and the known latent variable values was calculated as a measure of accuracy of predicted scores based on the two measurement frameworks

(see, for example, MacDonald and Paunonen, 2002). This measure indicates how much variance the true latent values and the predicted scores share in common (see, for example, O'Brien, 1981). As noted in Section 2.2, this measure could be considered as an alternative reliability index; however, to avoid confusion it will be referred to only as “*accuracy*” or “*shared variance*” in this thesis. Though this reliability index is very appealing, it is worth noting that it is not commonly used and is different from the classical *reliability* index based on the classical test theory for discrete indicators. Under the classical test theory, the reliability is calculated as the squared correlation of the predicted scores and the true scores (the expected predicted scores conditioned on the true latent variable) (see Section 2.2), and is usually estimated using Cronbach’s alpha, a lower bound. It will be seen later in Chapter 9 and Appendix B that Cronbach’s alpha is not a lower bound for the shared variance.

8.1.2 Simulation Procedures for the Two-Step Approach

The simulation of the two-step approach for using both IRT and CTT-based predicted scores in regression was carried out in the following steps using the software package SAS 8.2.:

Step 1. The true latent variable values and the discrete response indicators for both response and explanatory latent variables were created (see Stages 1-4 in Section 8.1.1). Note that the true latent variable values were known for each of the N respondents by design.

Step 2. EAP and CTT predicted scores were then calculated for each of the respondents from the simulated discrete response data (see Section 8.1.1).

Step 3. OLS regression using the above predicted scores for response and explanatory latent variables was then employed to estimate the latent regression parameter γ .

8.1.3 Simulation Procedures for the IRT-SEM Approach

This part of the simulation was conducted in the following steps using the Monte Carlo simulation feature included in the software package Mplus 2.14.

Step 1. The means and correlations of y_j^* and x_j^* were generated as input data in Mplus for a structural regression model with a regression coefficient $\gamma = \sqrt{R^2}$, and corresponding to the measurement models, $y_j^* = \lambda_j \eta + \varepsilon_j$ and $x_j^* = \lambda_j \xi + \delta_j$, with the indicator loadings $\lambda_j = \sqrt{CD}$.

Step 2. Category thresholds within each item, used for truncating y_j^* and x_j^* into discrete response data, y_j and x_j , were created as in Stage 3 in Section 8.1.1 using SAS.

Step 3. The discrete (binary and ordinal) item response data, y_j and x_j was generated by categorizing the multivariate item response data, y_j^* and x_j^* , using the thresholds created from the previous step.

Step 4. A theoretically consistent regression estimate, $\hat{\gamma}$, was then obtained. Note that Mplus uses a limited information generalized least squares approach, based on one-way and two-way marginal probabilities of the discrete item response data. Both simultaneous and constrained IRT-SEM approaches were carried out in Mplus with item parameters estimated for the former approach and fixed for the latter approach.

8.1.4 Simulation Procedures for Partial Least Squares Approach with Discrete Indicators

This part of the simulation was carried out as described below in the following steps:

Step 1. Discrete item response data was simulated as in Step 1 to Step 4 in section 8.1.1 by using the software package SAS 8.2.

Step 2. The regression estimate, $\hat{\gamma}$, was then estimated by the partial least squares approach using the software package PLS Graph 3 developed by Dr. Wynne Chin.

Table 8.2: Summary of the simulation procedures for generating discrete item response data and the simulation conditions in simple latent regression model

Simulation Procedures	Simulation Conditions	
	Description	Full set of values used*
Stage 1. Generating true latent variable values for N simulated respondents, corresponding to the preset R^2 of the structural regression model.	the coefficients of determination of the latent regression structure model: R^2	<ul style="list-style-type: none"> 0.1 and 0.5
	the sample size: N	<ul style="list-style-type: none"> 300, 500, 800, and 2000
Stage 2. Generating the multivariate normal item responses, y_j^* and x_j^* , corresponding to the structural and measurement models.	the coefficients of determination of the measurement models: CD	<ul style="list-style-type: none"> all 0.3 all 0.6 varying uniformly between 0.4 and 0.6 (loadings: 0.632-0.775) $CD=0.6$ for 1/5 of items $CD=0.1$ for 4/5 of items $CD=0.6$ for 2/5 of items $CD=0.1$ for 3/5 of items
	the number of simulated items in the scale: n	<ul style="list-style-type: none"> 5, 10, 20, and 30
Stage 3. Creating category thresholds within each item, used for truncating y_j^* and x_j^* into discrete response data, y_j and x_j , in the final stage. Stage 4. Generating the discrete (binary and ordinal) item response data, y_j and x_j .	the number of item response options (categories): $m+1$	<ul style="list-style-type: none"> 2, 3, 5, and 7
	the marginal frequency distribution of the discrete observable item indicators, as controlled by the item thresholds	<ul style="list-style-type: none"> all symmetric all uniform all skewed mixed: <ul style="list-style-type: none"> 2/5 of items are symmetric 1/5 of items are uniform 1/5 of items are skewed with level 1 1/5 of items are skewed with level 2

* Not all simulation condition values are used for all approaches. Specific condition values used for each approach are reported wherever appropriate.

8.2 Design of Simulation Study for a Multiple Latent Regression Model

All procedures used with the multiple latent regression model are analogous extensions of the procedures used with the simple regression model.

Simulating Discrete Item Responses. As for the simple latent regression model, the process of generating discrete item response data for the simulation involved four stages, all implemented using SAS 8.2. (summarized in Table 8.3).

Stage 1. Three vectors of true latent variable values of η , ξ_1 , and ξ_2 were generated for N simulated respondents, subject to the preset R^2 for the structural regression model. The variances of the latent variables η , ξ_1 , and ξ_2 were all set to one; the two regression coefficient parameters, γ_1 and γ_2 , were set to the same values. Vectors of the latent explanatory variables ξ_1 and ξ_2 were generated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\rho})$, where $\boldsymbol{\rho}$ is a 2 by 2 diagonal correlation matrix with off-diagonal element ρ . A vector of the disturbance variable ζ was created from a univariate normal distribution $N(0, \psi)$, yielding the coefficient of determination of, i.e., R^2 for the structural regression model. In terms of the parameters ψ and ρ , the regression coefficients, γ_1 and γ_2 are then given by $\sqrt{(1-\psi)[2(1+\rho)]^{-1}}$. The vector of realization of the dependent latent variable η was then computed via $\eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \zeta$.

Stage 2. Generating the multivariate normal item responses, y_j^* , x_{1j}^* and x_{2j}^* , to correspond to the structural and measurement models. The settings in this stage were

similar to those used for simple latent regression model in section 8.1.1. The three measurement models of η , ξ_1 , and ξ_2 were assigned identical loading parameter values and the same coefficient of determination.

Stage 3. *Creating category thresholds within each item, used for truncating y_j^* , x_{1j}^* and x_{2j}^* into discrete response data, y_j , x_{1j} , and x_{2j} , in the next stage.* The settings in this stage were also similar to those in the latent multiple regression model. The set of thresholds was designed to be identical for y_j^* , x_{1j}^* and x_{2j}^* .

Stage 4. *Generating the discrete (binary and ordinal) item response data, y_j , x_{1j} , and x_{2j} .* This stage was carried out by categorizing the multivariate item response data, y_j^* , x_{1j}^* and x_{2j}^* , into different number of categories, $m+1$, using the thresholds created from the previous stage, following system (6.4), (6.7) and (6.8).

The simulation procedures used for the multiple latent variable regression follow analogously from those described earlier for the simple regression model. The two-step approach is described in Section 8.1.2, IRT-SEM approach in Section 8.1.3, and PLS approach is described in Section 8.1.4.

Table 8.3: Summary of the simulation procedures for generating discrete item response data and the simulation conditions in multiple latent regression model

Simulation Procedures	Simulation Conditions	
	Description	Full set of values used*
Stage 1. Generating true latent variable values for N simulated respondents, corresponding to the preset R^2 of the structural regression model.	the coefficients of determination of the latent regression structure model: R^2	<ul style="list-style-type: none"> 0.3 and 0.5
	the sample size: N	<ul style="list-style-type: none"> 300, 500, 800, and 2000
	the correlation (ρ) between two latent explanatory variables in the multiple latent regression model	<ul style="list-style-type: none"> -0.5, -0.3, -0.1, 0, 0.1, 0.3, and 0.5
Stage 2. Generating the multivariate normal item responses, y_j^* , x_{1j}^* and x_{2j}^* , corresponding to the structural and measurement models.	the coefficients of determination of the measurement models: CD	<ul style="list-style-type: none"> all 0.3 all 0.6 varying uniformly between 0.4 and 0.6 (loadings: 0.632-0.775) $CD=0.6$ for 1/5 of items $CD=0.1$ for 4/5 of items $CD=0.6$ for 2/5 of items $CD=0.1$ for 3/5 of items
	the number of simulated items in the scale: n	<ul style="list-style-type: none"> 5, 10, 20, and 30
Stage 3. Creating category thresholds within each item, used for truncating y_j^* , x_{1j}^* and x_{2j}^* into discrete response data, y_j , x_{1j} , and x_{2j} , in the final stage. Stage 4. Generating the discrete (binary and ordinal) item response data, y_j , x_{1j} , and x_{2j} .	the number of item response options (categories): $m+1$	<ul style="list-style-type: none"> 2, 3, 5, and 7
	the marginal frequency distribution of the discrete observable item indicators, as controlled by the item thresholds	<ul style="list-style-type: none"> all symmetric all uniform all skewed mixed: <ul style="list-style-type: none"> 2/5 of items are symmetric 1/5 of items are uniform 1/5 of items are skewed with level 1 1/5 of items are skewed with level 2

* Not all simulation condition values are used for all approaches. Specific condition values used for each approach are reported wherever appropriate.

8.3 Simulation Conditions

In the simulation study, a number of conditions (factors) were manipulated: (a) coefficient of determination, R^2 , of the structural regression model, (b) sample size, (c) coefficient of determination (CD) of the measurement model, (d) number of items in the scale, (e) number of item response options (categories), (f) marginal frequency distribution of the discrete observable item indicators, as controlled by the item thresholds (see Figure 8.4), and (g) the correlation between two latent explanatory variables in the multiple latent regression model. All possible combinations of these conditions (factors) led to various simulation settings (treatments). For each simulation setting, a number of data sets were generated, which are called “replications”. Tables 8.2 and 8.3 summarize the simulation conditions under simple and multiple latent regression models, respectively.

The design of the Monte Carlo experiments in this thesis is similar to the following design recommendations made by Skrondal (2000): 1. Incomplete factorial designs were used instead of full factorial designs which may exhibit pronounced redundancy, especially when the number of simulation factors is large. That is, not all simulation condition values reported in Tables 8.2 and 8.3 are used for all approaches. The specific condition values used for each approach will be reported wherever appropriate. 2. The same set of seeds for the pseudo random number generators were utilized for each replication across all settings, in order to improve the precision in estimating the meta-models (presented in the next chapter). 3. Fewer replications (100) were used for each simulation setting, allowing us to investigate more potentially relevant factors to enhance

external validity at the expense of the “conventionally excessive precision”, given the limitations of computer resources and time (Skron dal 2000, p.165). The Monte Carlo standard errors of the regression coefficient estimates will be given in the results tables in the next chapter. 4. Specific values for the population model parameters (i.e., simulation factors) reflect values commonly encountered in applied research.

8.4 Summary

Monte Carlo simulation is a popular technique which allows us to assess the finite sampling performance of an estimator. A series of Monte Carlo experiments were carried out in this thesis for the following purposes: (a) to assess the comparability and accuracy of IRT- and CTT-based scores, (b) to investigate the finite item bias attributable to using predicted scores and to determine if the theoretically predicted bias was of a magnitude that warranted further investigation, (c) to evaluate how well the theoretically consistent simultaneous and constrained IRT-SEM estimates perform in practice, and (d) to explore (in preliminary fashion) the finite item bias exhibited by PLS.

Chapter 9

Research Results

To evaluate the research questions, a series of Monte Carlo experiments were conducted following the procedures described in Chapter 8. Selected simulation results are presented as tables and discussed as responses to the three research questions. For the second research question, as a supplementary aid for summarizing the tables and understanding the complete set of simulation results derived from each approach, explicit statistical modeling of the simulation outputs, referred to as *meta-models*, were also used. These models specify the relationships between the performance criteria (e.g., bias in a parameter estimate) and the experimental factors (e.g., the number of items, the number of item categories, etc). We refer to Skrondal (2000) for details regarding meta-modeling in Monte Carlo experiments. The graphical summary results based on the fitted meta-models will be used to describe the main trends in the findings. Overall, we used two main ways to present the simulation output: tabular displays of the simulation results; second, meta-model approximations to these simulation results which are used to generate descriptive plots of the main trends. In addition, the fitted meta-models can be

used to predict parameter bias for specific simulation settings, and this application will be described in detail.

The tables displayed in this chapter include the values of Cronbach's alpha for each particular simulation setting. This inclusion is made to give readers a practical appreciation for the scale corresponding to each particular simulation setting. In fact, as has been discussed in Chapter 2, Cronbach's alpha is not the measure of reliability most relevant to this study.

9.1 Comparability and Accuracy of IRT- and CTT-based Scores

For the first research question (*how comparable and how accurate are the CTT- and IRT-based scores?*), the results of the computer simulations under different simulation conditions are presented in Tables 9.1 through 9.6. Each correlation in these tables was derived based on the procedures described by Fan (1998) and MacDonald and Paunonen (2002), namely: first (a) obtain IRT- and CTT-based scores for each respondent in the simulated sample; (b) correlate the IRT- and CTT-based scores for each respondent in the same sample; and (c) average the correlations across all simulated samples under the same simulation conditions. Each entry in the table is the average of 100 correlation coefficients obtained from 100 randomly simulated samples of 300 respondents.

Comparability. Table 9.1 gives IRT-CTT correlations for symmetrically distributed discrete items, with measurement model *CD*'s of 0.3 and 0.6. The results show that IRT- and CTT-based scores are extremely highly correlated for both Pearson correlation and Spearman rank order correlation, ranging from 0.977 to 1 and from 0.992 to 1, respectively. The overall averages are 0.996 for Pearson correlation and 0.999 for Spearman correlation, and 0.851 for the Cronbach's alpha which is also displayed. Table 9.2 shows IRT-CTT correlations for symmetrically distributed items when the items in the measurement model have different precision (i.e., *CD* values), and Table 9.3 displays results for asymmetric items. Again, the correlations are generally very strong, though not as strong as those displayed in Table 9.1, with the majority of correlation coefficients exceeding 0.900. However, we did find situations where the Pearson correlations are less

than 0.900: (a) symmetric items with mixed measurement precision (4/5 of items with $CD=0.1$, i.e., loading = 0.316, while remaining items have $CD=0.6$, i.e., loading=0.775), displayed in Table 9.2, and (b) all asymmetric items with skewness level of two, displayed in Table 9.3.

Table 9.1: Comparability of IRT- and CTT-based scores with symmetrically distributed items

Category	CD	Scale length	Cronbach's alpha	Pearson correlation	Spearman correlation
2	0.3	5	0.543	0.999	0.992
		10	0.704	0.998	0.997
		20	0.828	0.997	0.999
		30	0.878	0.996	0.999
	0.6	5	0.774	0.995	0.998
		10	0.873	0.989	0.999
		20	0.933	0.982	1.000
		30	0.954	0.977	1.000
3	0.3	5	0.605	1.000	0.993
		10	0.756	1.000	0.997
		20	0.862	0.999	0.999
		30	0.903	0.999	0.999
	0.6	5	0.821	0.999	0.997
		10	0.902	0.998	0.999
		20	0.949	0.995	1.000
		30	0.965	0.992	1.000
5	0.3	5	0.651	0.999	0.996
		10	0.789	0.999	0.999
		20	0.883	0.999	1.000
		30	0.919	0.999	1.000
	0.6	5	0.858	0.999	0.998
		10	0.924	0.998	0.999
		20	0.961	0.994	1.000
		30	0.973	0.991	1.000
7	0.3	5	0.663	0.999	0.998
		10	0.797	1.000	0.999
		20	0.888	1.000	1.000
		30	0.922	0.999	1.000
	0.6	5	0.868	0.999	0.999
		10	0.930	0.998	1.000
		20	0.964	0.994	1.000
		30	0.976	0.991	1.000
Overall Average			0.851	0.996	0.999

Note: Each entry is based on 100 replications with sample size of 300.

Table 9.2: Comparability of IRT- and CTT-based scores with symmetrically distributed items, where items in the measurement model differed in *CD*

Category	CD	Scale Length	Cronbach's alpha	Pearson Correlation	Spearman Correlation
2	0.6 in 1/5 items 0.1 in 4/5 items	5	0.356	0.893	0.913
		10	0.532	0.926	0.922
		20	0.702	0.953	0.948
		30	0.780	0.965	0.963
	0.6 in 2/5 items 0.1 in 3/5 items	5	0.474	0.918	0.938
		10	0.651	0.947	0.948
		20	0.794	0.966	0.967
		30	0.853	0.973	0.976
3	0.6 in 1/5 items 0.1 in 4/5 items	5	0.410	0.888	0.900
		10	0.595	0.924	0.916
		20	0.753	0.953	0.947
		30	0.820	0.966	0.961
	0.6 in 2/5 items 0.1 in 3/5 items	5	0.533	0.913	0.920
		10	0.705	0.946	0.941
		20	0.832	0.969	0.965
		30	0.881	0.977	0.975
5	0.6 in 1/5 items 0.1 in 4/5 items	5	0.582	0.880	0.876
		10	0.742	0.921	0.917
		20	0.854	0.953	0.949
		30	0.898	0.966	0.963
	0.6 in 2/5 items 0.1 in 3/5 items	5	0.692	0.911	0.908
		10	0.823	0.946	0.943
		20	0.905	0.969	0.968
		30	0.934	0.977	0.977
Overall Average			0.713	0.942	0.942

Note: Each entry is based on 100 replications with sample size of 300.

Table 9.3: Comparability of IRT- and CTT-based scores with asymmetrically distributed items at skew level of 2

Category	CD	Scale Length	Cronbach's alpha	Pearson Correlation	Spearman Correlation
2	0.3	5	0.453	0.991	0.995
		10	0.626	0.982	0.995
		20	0.771	0.969	0.998
		30	0.836	0.961	0.998
	0.6	5	0.723	0.975	1.000
		10	0.839	0.954	0.999
		20	0.913	0.931	0.999
		30	0.940	0.919	0.999
3	0.3	5	0.509	0.980	0.995
		10	0.677	0.969	0.997
		20	0.808	0.957	0.997
		30	0.864	0.950	0.998
	0.6	5	0.775	0.955	1.000
		10	0.873	0.933	0.999
		20	0.933	0.912	0.999
		30	0.954	0.902	0.999
5	0.3	5	0.530	0.962	0.993
		10	0.695	0.956	0.990
		20	0.821	0.948	0.992
		30	0.874	0.944	0.994
	0.6	5	0.794	0.936	0.998
		10	0.886	0.918	0.997
		20	0.940	0.902	0.998
		30	0.959	0.893	0.998
7	0.3	5	0.567	0.953	0.986
		10	0.726	0.951	0.987
		20	0.842	0.947	0.991
		30	0.889	0.945	0.994
	0.6	5	0.820	0.929	0.996
		10	0.901	0.916	0.997
		20	0.948	0.904	0.998
		30	0.965	0.897	0.998
Overall Average			0.802	0.942	0.996

Note: Each entry is based on 100 replications with sample size of 300.

Accuracy. As noted previously, the squared Pearson correlation between the predicted scores (obtained with both IRT and CTT) and the known latent variable values was used as a measure of predicted score accuracy. This measure indicates how much variance the true latent variable values and the predicted scores share in common and is regarded as an accuracy index in this study.

The results in Table 9.4 show that shared variances with the true latent variable values for both IRT- and CTT- scores are very similar, with IRT-CTT pairs of shared variances ranging from a low of (0.540, 0.539) to a high (0.961, 0.973), respectively, across all levels of categories, *CD*, and scale length, for symmetrically distributed items. The accuracy of the IRT- and CTT-based scores could be considered strong or somewhat strong (0.800s to 0.900s) under some conditions, while the accuracy is moderate (0.500s to 0.600s) when only 5 items with *CD* = 0.3 are used. The overall average of the shared variance is 0.843 for IRT-scores and 0.841 for CTT-scores. It is worth noting that the overall average of Cronbach's alpha, 0.851, is greater than the shared variance for CTT-scores (see Appendix B for a mathematical explanation), which underlies the difference between Cronbach's alpha and shared variance as measures of reliability (see Section 2.2).

As in the assessment of comparability, two situations were found where IRT- and CTT-based scores performed differently, with IRT scores showing more accuracy. The first situation is illustrated in Table 9.5, where 3/5 or 4/5 of items have low measurement precision (*CD* = 0.1) and the rest have good measurement precision (*CD* = 0.6). In this case, the overall average difference in accuracy between IRT and CTT scores is 8.09%.

For the second situation, when items are skewed (Table 9.6), the overall averaged difference is 8.50%. In the first situation, the greatest differences in accuracy between the two predicted scores, ranging from 11.3% to 14%, were found when 5 items with 3 or 5 categories are used. For the second situation where all the items are asymmetrically distributed with skewness level 2, the discrepancy between the accuracy of IRT and CTT-based scores increases as the scale length increases. The greatest difference in IRT and CTT accuracy, 16.96%, occurs when 30 items are used with $CD = 0.6$ and 5 categories.

To summarize the results for comparability and accuracy thus far, IRT-based scores are closely comparable with CTT-based scores in the sense that the majority of their average correlation coefficients are above 0.900, which is typical of the past empirical and Monte Carlo results shown in Table 4.1, e.g., Embretson & Reise (2000), Fan (1998), Koch (1983), Lawson (1991), MacDonald & Paunonen (2002), and Ndlichako & Rogers (1997). Both types of the predicted scores contain measurement errors, but their shared variances with the true latent values (i.e., accuracy) nevertheless range from moderate to high (0.500s to 0.900s) when the items are symmetrically distributed with $CD = 0.3$ or 0.6 . Our findings confirm the simulation results based on 20 items by MacDonald & Paunonen (2002) who examined the accuracy of IRT and CTT scores for the binary case. With regards to the accuracy of predicting the true latent values, IRT and CTT may yield noticeable discrepancies in situations where some of the items have low measurement precision ($CD = 0.1$) and the rest have good measurement precision ($CD = 0.6$), and in the case when all items are skewed.

In conclusion, regardless of the measurement framework, IRT and CTT perform similarly in terms of latent variable prediction. Nevertheless, IRT is a slightly more

robust method in that it can tolerate skewness and poor measurement precision of individual item better than can CTT.

Table 9.4: Accuracy of IRT- and CTT-based Scores with symmetrically distributed items

Category	CD	Scale Length	Cronbach's alpha	Shared Variance with True Latent Trait Value		
				IRT (A)	CTT (B)	(A-B)%
2	0.3	5	0.543	0.540	0.539	0.10
		10	0.704	0.701	0.699	0.20
		20	0.828	0.823	0.817	0.60
		30	0.878	0.873	0.866	0.70
	0.6	5	0.774	0.731	0.725	0.60
		10	0.873	0.835	0.818	1.70
		20	0.933	0.900	0.872	2.80
		30	0.954	0.922	0.891	3.10
3	0.3	5	0.605	0.609	0.608	0.10
		10	0.756	0.757	0.757	0.00
		20	0.862	0.862	0.861	0.10
		30	0.903	0.902	0.901	0.10
	0.6	5	0.821	0.817	0.815	0.20
		10	0.902	0.897	0.893	0.40
		20	0.949	0.942	0.939	0.30
		30	0.965	0.954	0.956	-0.20
5	0.3	5	0.651	0.652	0.652	0.00
		10	0.789	0.791	0.791	0.00
		20	0.883	0.883	0.882	0.10
		30	0.919	0.918	0.917	0.10
	0.6	5	0.858	0.858	0.855	0.30
		10	0.924	0.923	0.921	0.20
		20	0.961	0.953	0.957	-0.40
		30	0.973	0.960	0.969	-0.90
7	0.3	5	0.663	0.665	0.664	0.10
		10	0.797	0.800	0.800	0.00
		20	0.888	0.888	0.888	0.00
		30	0.922	0.922	0.922	0.00
	0.6	5	0.868	0.868	0.866	0.20
		10	0.930	0.929	0.928	0.10
		20	0.964	0.955	0.961	-0.60
		30	0.976	0.961	0.973	-1.20
Overall Average			0.851	0.843	0.841	0.28

Note: Each entry is based on 100 replications with sample size of 300.

Table 9.5: Accuracy of IRT- and CTT-based Scores of simple latent regression model with symmetrically distributed items, where items in the measurement model differed in *CD*

Category	CD	Scale Length	Cronbach's alpha	Shared Variance with True Latent Trait Value		
				IRT (A)	CTT (B)	(A-B)%
2	0.6 in 1/5 0.1 in rest	5	0.356	0.469	0.371	9.80
		10	0.532	0.630	0.539	9.10
		20	0.702	0.768	0.696	7.20
		30	0.780	0.829	0.772	5.70
	0.6 in 2/5 0.1 in rest	5	0.474	0.580	0.485	9.50
		10	0.651	0.722	0.646	7.60
		20	0.794	0.832	0.776	5.60
		30	0.853	0.878	0.832	4.60
3	0.6 in 1/5 0.1 in rest	5	0.410	0.555	0.435	12.00
		10	0.595	0.712	0.607	10.50
		20	0.753	0.830	0.755	7.50
		30	0.820	0.880	0.822	5.80
	0.6 in 2/5 0.1 in rest	5	0.533	0.673	0.560	11.30
		10	0.705	0.801	0.718	8.30
		20	0.832	0.889	0.835	5.40
		30	0.881	0.922	0.883	3.90
5	0.6 in 1/5 0.1 in rest	5	0.582	0.613	0.473	14.00
		10	0.742	0.760	0.644	11.60
		20	0.854	0.865	0.785	8.00
		30	0.898	0.905	0.845	6.00
	0.6 in 2/5 0.1 in rest	5	0.692	0.730	0.603	12.70
		10	0.823	0.842	0.753	8.90
		20	0.905	0.915	0.860	5.50
		30	0.934	0.939	0.902	3.70
Overall Average			0.713	0.722	0.692	8.09

Note: Each entry is based on 100 replications with sample size of 300.

Table 9.6: Accuracy of IRT- and CTT-based scores with asymmetrically distributed items at skew level of 2

Category	CD	Scale Length	Cronbach's alpha	Shared Variance With True Latent Trait Value		
				IRT (A)	CTT (B)	(A-B)%
2	0.3	5	0.456	0.403	0.396	0.65
		10	0.626	0.564	0.543	2.03
		20	0.774	0.712	0.668	4.41
		30	0.836	0.781	0.722	5.89
	0.6	5	0.719	0.557	0.529	2.76
		10	0.837	0.676	0.614	6.21
		20	0.912	0.768	0.667	10.06
		30	0.940	0.809	0.688	12.15
3	0.3	5	0.514	0.464	0.444	1.97
		10	0.676	0.628	0.589	3.94
		20	0.811	0.763	0.697	6.64
		30	0.865	0.823	0.744	7.95
	0.6	5	0.774	0.631	0.573	5.80
		10	0.873	0.740	0.645	9.53
		20	0.933	0.820	0.686	13.39
		30	0.954	0.854	0.702	15.26
5	0.3	5	0.535	0.504	0.464	3.95
		10	0.695	0.669	0.608	6.10
		20	0.823	0.796	0.713	8.25
		30	0.875	0.849	0.757	9.19
	0.6	5	0.793	0.679	0.594	8.59
		10	0.885	0.783	0.661	12.23
		20	0.940	0.854	0.699	15.48
		30	0.959	0.884	0.714	16.96
7	0.3	5	0.571	0.560	0.507	5.21
		10	0.726	0.719	0.648	7.06
		20	0.844	0.833	0.747	8.66
		30	0.890	0.879	0.787	9.19
	0.6	5	0.821	0.746	0.644	10.26
		10	0.902	0.840	0.706	13.41
		20	0.949	0.898	0.741	15.69
		30	0.965	0.918	0.754	16.40
Overall Average			0.802	0.731	0.645	8.60

Note: Each entry is based on 100 replications with sample size of 300.

9.2 Simulation Results on the Percentage Bias of the Two-Step and IRT-SEM Approaches

The Monte Carlo results described in this section provide a comparison of the parameter biases arising from the use of the two-step methods with those arising from IRT-SEM approaches. It will also show how the performance of each latent regression method is related to different simulation conditions.

9.2.1 Results on Percentage Bias in the Simple Latent Regression Model

Tables 9.7-9.11 display the effect of various simulation conditions on the bias in the simple regression estimates. Each bias entry in these tables was derived based on the following procedures: (a) obtain the parameter estimate (R^2 or regression coefficient) from the latent variable regression for each simulated sample, and (b) average the percentage biases across all simulated samples under the same simulation setting. Each bias entry is the average of percentage R^2 biases or percentage regression coefficient biases obtained from 100 randomly simulated samples of N respondents. The percentage biases of the R^2 and regression coefficient are defined as follows, respectively:

$$\% \text{bias of } R^2 = \{[(\hat{R}^2 - R^2)/R^2] 100\} \quad (9.1)$$

and

$$\% \text{bias of } \gamma = \{[(\hat{\gamma} - \gamma)/\gamma] 100\}. \quad (9.2)$$

In other words, the table entries present relative bias, though this is referred to throughout the following discussions simply as *bias* to simplify the terminology.

In particular, Table 9.7 shows the effect of scale length on bias in the regression estimates, for a sample of size 300. The first feature of note is that the biases in the IRT and CTT two-step methods are very similar for all conditions, and are unacceptably large in most cases. For a 5 category scale containing 5 items, for example, the percentage bias in R^2 is -26% for both IRT and CTT two-step methods. As predicted, the parameter bias

with IRT and CTT regression decreases as number of items increases and is appreciably larger than in simultaneous IRT-SEM. Table 9.7 also shows that for both two-step approaches, the bias decreases as the number of categories increases, for a sample of size 300, and symmetrically distributed items. For scales containing 5 or 10 items, the bias decreases as the number of categories increases from 2 to 5. There is relatively little improvement in bias from 5 to 7 categories. For scales containing 20 or 30 items, most of the improvement in bias occurs between two and three categories.

The simultaneous IRT-SEM approach is not as sensitive to number of categories as are the two-step methods, though a decrease in bias can be seen when the number of categories increases from 2 to 3, particularly for the two shortest scale conditions (see Table 9.7). This effect is due to *finite sample* bias, and would decrease if the sample size were increased from 300. In all cases the bias in IRT-SEM estimates is less than that in either of the two-step methods, the relative benefits of IRT-SEM being particularly noticeable for 10 or fewer items (e.g., a bias of 44% in R^2 dropping to 11% for the case of 5 items and 2 categories). The results based on the constrained IRT-SEM approach are reported in the highlighted bottom entry in each cell of Tables 9.7-9.11. It can be seen in Table 9.7 that the bias of the constrained IRT-SEM approach is essentially negligible in all cases.

Table 9.7: Effects of scale Length and number of categories on bias in the simple latent regression

True R^2 = 0.5, True Reg. Coef.= 0.707, Sample size = 300, CD = 0.6, Symmetrically distributed items				
Scale length	2 categories	3 categories	5 categories	7 categories
05y-05x	0.774 -43.61 / -25.09 -44.01 / -25.40 11.21 / 5.46 1.99 / 0.99	0.821 -32.76 / -18.49 -32.91 / -18.26 4.57 / 2.26 1.08 / 0.54	0.858 -26.28 / -14.47 -26.44 / -14.37 6.32 / 3.11 1.67 / 0.83	0.868 -24.67 / -13.64 -24.89 / -13.46 6.17 / 3.04 2.04 / 1.02
10y-10x	0.873 -28.10 / -15.36 -28.73 / -15.74 2.59 / 1.29 -0.48 / -0.24	0.902 -18.76 / -10.34 -18.84 / -10.03 5.85 / 2.88 2.62 / 1.30	0.924 -14.79 / -8.07 -14.67 / -7.74 4.29 / 2.12 1.99 / 0.99	0.930 -13.72 / -7.56 -13.60 / -7.15 3.28 / 1.63 1.50 / 0.75
20y-20x	0.932 -17.63 / -8.97 -19.25 / -10.30 5.97 / 2.94 1.50 / 0.75	0.949 -10.65 / -5.66 -10.37 / -5.42 3.56 / 1.77 1.16 / 0.58	0.960 -9.03 / -4.64 -7.82 / -4.08 2.96 / 1.47 0.85 / 0.42	0.964 -8.66 / -4.50 -7.10 / -3.70 2.82 / 1.40 0.76 / 0.38
30y-30x	0.954 -13.95 / -6.90 -15.79 / -8.37 2.16 / 1.07 -0.96 / -0.48	0.965 -8.70 / -4.61 -7.41 / -3.86 2.24 / 1.12 -0.37 / -0.18	0.973 -7.84 / -4.09 -5.65 / -2.95 2.22 / 1.10 -0.12 / -0.06	0.975 -7.43 / -3.88 -5.08 / -2.66 2.07 / 1.03 -0.20 / -0.10

Note: 1. In each cell, the first number is Cronbach's alpha of the dependent latent variable; the subsequent pairs of numbers represent the %bias in R^2 and %bias in regression coefficient for IRT approach (first pair), CTT approach (second pair), simultaneous IRT-SEM approach (third pair), and constrained IRT-SEM approach (fourth pair). 2. Standard error of the regression estimates ranges from 0.004-0.005 (in IRT approach), from 0.003-0.004 (in CTT approach), from 0.006-0.011 (in simultaneous IRT-SEM approach), and from 0.003-0.006 (in constrained IRT-SEM approach).

Table 9.8: Effect of sample size on bias in the simple latent regression

True $R^2 = 0.5$, True Reg. Coef.= 0.707, $CD=0.6$, Scale length = 5y-5x, Symmetrically distributed items				
Sample size	2 categories	3 categories	5 categories	7 categories
300	(0.774) -43.61 / -25.09 -44.01 / -25.40 11.21 / 5.46 1.99 / 0.99	(0.822) -32.76 / -18.49 -32.91 / -18.26 4.57 / 2.26 1.08 / 0.54	(0.858) -26.28 / -14.47 -26.44 / -14.37 6.32 / 3.11 1.67 / 0.83	(0.869) -24.67 / -13.64 -24.89 / -13.46 6.17 / 3.04 2.04 / 1.02
500	(0.776) -43.11 / -24.56 -43.51 / -24.96 6.26 / 3.08 0.82 / 0.41	(0.823) -33.05 / -18.44 -33.04 / -18.26 1.56 / 0.78 0.03 / 0.01	(0.859) -26.41 / -14.75 -26.42 / -14.31 3.22 / 1.60 0.54 / 0.27	(0.869) -24.73 / -13.90 -24.71 / -13.31 3.28 / 1.63 0.76 / 0.38
800	(0.778) -43.16 / -24.44 -43.44 / -24.87 3.19 / 1.58 -0.23 / -0.11	(0.824) -32.41 / -17.76 -32.38 / -17.82 1.27 / 0.64 0.31 / 0.15	(0.861) -25.57 / -13.76 -25.62 / -13.80 2.10 / 1.05 0.48 / 0.24	(0.871) -24.05 / -12.96 -24.11 / -12.93 2.27 / 1.13 0.48 / 0.24
2000	(0.776) -43.95 / -25.23 -44.22 / -25.35 0.68 / 0.34 -0.28 / -0.14	(0.823) -32.68 / -18.11 -32.71 / -18.00 0.51 / 0.25 0.25 / 0.13	(0.860) -26.32 / -14.35 -26.43 / -14.26 0.71 / 0.35 0.03 / 0.01	(0.870) -24.31 / -13.16 -24.47 / -13.12 0.76 / 0.38 0.17 / 0.08

Note: 1. See the descriptions in Table 9.7. 2. Standard error of the regression estimates ranges from 0.002-0.005 (in IRT approach), from 0.002-0.004 (in CTT approach), from 0.002-0.012 (in simultaneous IRT-SEM approach), and from 0.002-0.006 (in constrained IRT-SEM approach).

Table 9.9: Effect of mmeasurement precision (CD) on bias in the simple latent regression

True $R^2 = 0.5$, True Reg. Coef.= 0.707, Sample size = 300, Scale length = 5y-5x, Symmetrically distributed items				
CD	2 categories	3 categories	5 categories	7 categories
0.3	(0.542) -70.86 / -46.33 -70.85 / -46.38 32.58 / 15.15 2.93 / 1.46	(0.605) -62.27 / -39.26 -62.34 / -38.95 14.91 / 7.20 1.73 / 0.86	(0.651) -57.59 / -35.29 -57.69 / -35.20 12.92 / 6.26 2.99 / 1.48	(0.663) -55.69 / -33.96 -55.82 / -33.81 10.47 / 5.10 1.84 / 0.92
0.6	(0.774) -43.61 / -25.09 -44.01 / -25.40 11.21 / 5.46 1.99 / 0.99	(0.821) -32.76 / -18.49 -32.91 / -18.26 4.57 / 2.26 1.08 / 0.54	(0.858) -26.28 / -14.47 -26.44 / -14.37 6.32 / 3.11 1.67 / 0.83	(0.868) -24.67 / -13.64 -24.89 / -13.46 6.17 / 3.04 2.04 / 1.02

Note: 1. See the descriptions in Table 9.7. 2. Standard error of the regression estimates ranges from 0.004-0.005 (in IRT approach), from 0.003-0.004 (in CTT approach), from 0.007-0.029 (in simultaneous IRT-SEM approach), from 0.004-0.01 (in constrained IRT-SEM approach).

It can be seen from Table 9.8 that with 5 symmetrically distributed items and $CD=0.6$ in each measurement model, the two-step approach is insensitive to sample size as expected but leads to appreciable attenuation in IRT and CTT regression coefficient estimates and in R^2 . On the other hand, the simultaneous IRT-SEM approach produced smaller bias even for the smaller sample sizes and as expected it generates 'consistent' regression estimates for suitably large sample size. Again, bias for constrained IRT-SEM is negligible even for the smallest sample size.

The results in Table 9.9 show that with 5 symmetrically distributed items in each measurement model and a sample of size 300, the higher the CD 's in the measurement model the lower the bias for all approaches. However, even for scales with $CD=0.6$ and 7 categories, the bias in the two-step methods is still appreciable, close to 25% in R^2 . Table 9.10 presents the effect of symmetry of an item's marginal distribution on the bias in the regression estimates for a sample of size 500, when 5 items with $CD=0.6$ in each measurement model are used. It shows that the more skewed the marginal distribution of items, the more bias there is in the two-step approach, reaching 45% even for 5 categories. However, the simultaneous IRT-SEM approach appears not to be sensitive to skewness, and the constrained IRT-SEM approach again yields very little bias.

Table 9.10: Effect of symmetry of item's marginal distribution on bias in the simple latent regression

True R^2 = 0.5, True Reg. Coef. = 0.707, Sample Size = 500, $CD=0.6$, Scale length = 5y-5x				
Distribution	2 categories	3 categories	5 categories	7 categories
Symmetric	(0.776) -43.11 / -24.56 -43.51 / -24.96 6.26 / 3.08 0.82 / 0.41	(0.823) -33.05 / -18.44 -33.04 / -18.26 1.56 / 0.78 0.03 / 0.01	(0.859) -26.41 / -14.75 -26.42 / -14.31 3.22 / 1.60 0.54 / 0.27	(0.869) -24.73 / -13.90 -24.71 / -13.31 3.28 / 1.63 0.76 / 0.38
Skewed (level=2)	(0.724) -57.87 / -35.50 -56.02 / -34.00 4.57 / 2.26 0.20 / 0.10	(0.778) -50.74 / -30.41 -49.29 / -29.08 2.88 / 1.43 0.08 / 0.04	(0.797) -45.53 / -26.60 -46.16 / -26.88 2.10 / 1.05 -0.54 / -0.27	(0.824) -39.04 / -22.04 -41.27 / -23.60 2.82 / 1.40 ---

Note: 1. See the descriptions in Table 9.7. 2. Standard error of the regression estimates ranges from 0.002-0.006 (in IRT approach), from 0.002-0.006 (in CTT approach), from 0.006-0.011 (in simultaneous IRT-SEM approach), from 0.003-0.006 (in constrained IRT-SEM approach). 3. "---" denotes results are not available from the analysis due to too few observations in some categories.

Table 9.11 presents the bias in the regression estimates under a more realistic situation where CD varies uniformly between 0.4 and 0.6 for items in each measurement model, the marginal frequency distributions of the observable items are symmetric for 2/5 of items, uniform for 1/5 of items, skewed with skewness level 1 for 1/5 of items, skewed with level 2 for 1/5 of items, and where the sample size is a moderate 300. When 5 items are used in two-step approach, the downward bias in R^2 increases from the 0.30's for 5 or 7 categories to the 0.50's for 2 categories. It can be seen that there is again appreciable bias in the two-step approach and considerably less bias in the simultaneous IRT-SEM approach. Once again, as predicted, the constrained IRT-SEM approach yields the smallest bias of all, even when sample size is only 300.

Table 9.11: Effect of scale length and number of categories in the simple latent regression under a realistic situation

True R^2 = 0.5, True Reg. Coef. = 0.707, Sample size = 300, CD = 0.4-0.6 (loadings: 0.632-0.775), Mixed items: Symmetric, Uniform, and Skewed				
Scale length	2 categories	3 categories	5 categories	7 categories
5y-5x	(0.679)	(0.735)	(0.768)	(0.783)
	-52.64 / -31.20	-42.36 / -24.37	-36.20 / -20.47	-33.21 / -18.78
	-52.98 / -31.71	-43.51 / -25.07	-39.33 / -22.31	-36.56 / -20.54
	15.82 / 7.62	11.63 / 5.66	9.14 / 4.47	9.11 / 4.45
	3.22 / 1.60	3.54 / 1.75	2.07 / 1.03	2.53 / 1.26
10y-10x	(0.812)	(0.850)	(0.871)	(0.880)
	-34.18 / -18.81	-25.86 / -14.05	-21.03 / -11.34	-19.23 / -10.44
	-34.63 / -19.32	-26.98 / -14.71	-23.41 / -12.62	-21.43 / -11.51
	4.57 / 2.26	4.23 / 2.09	3.22 / 1.60	3.05 / 1.51
	1.39 / 0.69	0.99 / 0.49	0.99 / 0.49	0.96 / 0.48
20y-20x	(0.897)	(0.920)	(0.931)	(0.937)
	-21.16 / -10.96	-14.65 / -7.54	-12.14 / -6.09	-10.80 / -5.38
	-21.78 / -11.72	-15.33 / -8.10	-13.15 / -6.91	-11.75 / -6.17
	4.55 / 2.25	3.22 / 1.60	2.82 / 1.40	3.62 / 1.80
	0.48 / 0.24	0.82 / 0.41	0.45 / 0.23	0.99 / 0.49

Note: 1. See the descriptions in Table 9.7. 2. Standard error of the regression estimates ranges from 0.004-0.005 (in IRT approach), from 0.003-0.005 (in CTT approach), from 0.006-0.012 (in simultaneous IRT-SEM approach), from 0.003-0.006 (in constrained IRT-SEM approach).

As in the assessment of comparability and accuracy, two situations were found where IRT- and CTT-based scores performed differently, with IRT scores showing more accuracy. Tables 9.12 and 9.13 show that the IRT-based two-step approach outperforms the CTT-based approach in these two situations. For the first situation, where 3/5 of items have low precision ($CD = 0.1$) and the rest have good precision ($CD = 0.6$), the discrepancy between the percentage R^2 bias in of IRT and CTT-based scores increases, as the number of categories increases. The greatest difference in R^2 , 15.92%, occurs when 5 items with 5 categories are used. For the second situation where all the items are asymmetrically distributed with skewness level 2, the greatest differences in the bias of R^2 , ranging from 14.35% to 15.40%, between the two-step approaches were found when 20 items with 5 or 7 categories are used. In both situations, however, the biases of two-

step methods are unacceptably large in most cases. The IRT-SEM approach still shows much lower bias than that of two-step approaches, with bias levels of acceptable size for all but the shortest binary scale.

Table 9.12: IRT outperforms CTT (1)

True $R^2 = 0.5$, True Reg. Coef. = 0.707, 2/5 items with $CD = 0.6$, rest with $CD = 0.1$, Sample size = 300, Symmetrically distributed items			
Scale length	2 categories	3 categories	5 categories
5y - 5x	(0.47)	(0.53)	(0.69)
	-65.09 / -41.41	-54.74 / -33.05	-46.44 / -27.03
	-75.75 / -51.19	-67.76 / -43.56	-62.36 / -38.96
10y - 10x	23.34 / 11.06	13.43 / 6.50	10.35 / 5.05
	(0.65)	(0.70)	(0.82)
	-45.22 / -26.41	-34.92 / -19.81	-28.50 / -15.75
	-56.34 / -34.19	-47.42 / -27.69	-42.40 / -24.31
	7.90 / 3.87	8.37 / 4.10	8.69 / 4.26

Note: 1. In each cell, the first number is Cronbach's alpha of the dependent latent variable; the subsequent pairs of numbers represent the %bias in R^2 and %bias in regression coefficient for IRT approach (first pair), CTT approach (second pair), and simultaneous IRT-SEM approach (third pair). 2. Standard error of the regression estimates ranges from 0.004-0.005 (in IRT approach), from 0.003-0.005 (in CTT approach), from 0.007-0.02 (in simultaneous approach).

Table 9.13: IRT outperforms CTT (2)

True $R^2 = 0.1$, True Reg. Coef. = 0.316, $CD = 0.6$, Sample size = 300, Scale length = 20y - 20x, Skewed items (level=2)				
Category	Cronbach's alpha	IRT Regression Estimation	CTT Regression Estimation	IRT-SEM Estimation
2	0.913	-34.46 / -20.87	-41.12 / -25.97	-1.91 / -0.96
3	0.933	-27.14 / -16.22	-37.64 / -23.65	-2.47 / -1.24
5	0.940	-21.48 / -13.06	-35.83 / -22.45	-3.28 / -1.65
7	0.949	-14.26 / -8.52	-29.66 / -18.44	0.74 / 0.37

Note: 1. Cronbach's alpha: for the dependent latent variable. 2. In each cell, the pair of numbers represents the %bias in R^2 and %bias in regression coefficient. 3. Standard error of the regression estimates ranges from 0.004-0.005 (in IRT approach), from 0.003-0.005 (in CTT approach), from 0.006-0.008 (in simultaneous approach).

Based on what we have seen above, we can identify the conditions that lead to the lowest bias in the two-step approach, namely: having a long scale, more item categories, a high CD , and symmetric items. The conditions that lead to the largest biases in the two-step approach are: having short scales, few item categories, lower values of CD , and asymmetric items. The bias in R^2 from the use of IRT- and CTT-scores in regression is very similar in general. Nevertheless, IRT-based latent score regression is a slightly more robust method than the two-step CTT regression approach, since IRT scores have better accuracy than CTT scores when items are all asymmetrically distributed, or part of items have poor measurement precision and the rest have good measurement precision. Despite this slight advantage of IRT scores over CTT scores, however, the percentage bias exhibited by the two-step regression based on IRT scores is still too large in most cases. For simultaneous IRT-SEM, the most critical single parameter appears to be sample size; more than 500 observations are needed for finite sample bias in R^2 to go down to less than 7%, in general. However, constrained IRT-SEM yields smaller biases even with small sample sizes, and is clearly the method to use when IRT parameters are known a priori.

9.2.2 Results on Percentage Bias in the Multiple Latent Regression Model

Tables 9.14-9.19 present the effect of various simulation conditions on the bias in the multiple regression estimates. Each bias entry in these tables was derived based on the same procedures as described in Section 9.2.1. Table 9.14 shows the effect of scale length and number of item categories on bias in the regression estimates, for a sample of size 300 and symmetrically distributed items. As for the simple latent regression case, the parameter biases exhibited by the IRT and CTT two-step methods are very similar for most conditions, and again too large for comfort. The bias with IRT and CTT latent regression decreases as number of items increases, as expected, and is noticeably larger than in simultaneous IRT-SEM. For scales with 5 or 10 items, the bias decreases as the number of categories increases from 2 to 5. There is only slight improvement in bias from 5 to 7 categories. For scales containing 20 or 30 items, a noticeable decrease in bias occurs as the number of categories increases from 2 to 3. The simultaneous IRT-SEM approach is not sensitive to scale length and number of categories. In all cases the bias in IRT-SEM estimates is less than that in either of the two-step methods; appreciable discrepancy in bias between IRT-SEM and the two-step approaches occurs particularly for 10 or fewer items (e.g., a bias of -37% in R^2 dropping to 8% for the case of 5 items and 2 categories). Because the bias in the simultaneous IRT-SEM approach is not as large as for the two-step methods, the relationships of this bias to scale length and number of categories are not as noticeable. It can be seen from Table 9.14, however, that bias does drop with scale length, particularly from 5 items to 10 items, and with number

of categories. This is a finite sample effect; the sample size for this table being only 300. The IRT-SEM result for 30 items appears to reverse the trend displayed for shorter scales. There is no ready explanation for this, and it is assumed to be a statistical / experimental anomaly. For scale lengths up to 20 items and for all number of categories, the bias exhibited by the constrained IRT-SEM method is very small, around 1% or less in most cases.

From Table 9.15, it can be seen that with 5 symmetrically distributed items and $CD=0.6$ in each measurement model, the two-step approach is insensitive to sample size as expected, but leads to appreciable attenuation in regression coefficient estimates and in R^2 . On the contrary, the simultaneous IRT-SEM approach produced smaller bias even for the smaller sample sizes and as expected it generates ‘consistent’ regression estimates for suitably large sample size, and again the bias of the constrained IRT-SEM approach is negligible.

Table 9.16 shows that with 5 symmetrically distributed items in each measurement model and a sample of size 300, the higher the CD ’s in the measurement model the lower the bias for all approaches. However, even for scales with $CD=0.6$ and 7 categories, the R^2 bias in the two-step methods is still appreciable (about 20%); on the other hand, the bias in the simultaneous IRT-SEM approach is only 4%. For the constrained IRT-SEM approach, it is less than 1%.

Table 9.17 presents the effect of symmetry of an item’s marginal distribution on the bias in the regression estimates for a sample of size 500, when 5 items with $CD=0.6$ in each measurement model are used. It shows that the more skewed the marginal

distribution of items, the more bias there is in the two-step approach. However, the simultaneous IRT-SEM approach appears not to be very sensitive to skewness, and the constrained IRT-SEM method again shows the lowest levels of bias and no sensitivity whatsoever to skewness.

Table 9.18 displays the bias in the regression estimates under a more realistic situation where CD varies uniformly between 0.4 and 0.6 for items in each measurement model, the marginal frequency distributions of the observable items are symmetric for 2/5 of items, uniform for 1/5 of items, skewed with skewness level 1 for 1/5 of items, skewed with level 2 for 1/5 of items, and where the sample size is a moderate 300. When 5 items are used in the two-step approach, the downward bias in R^2 increases from the 0.20's or 0.30's for over 5 categories to the 0.40's for 2 categories. It can be seen that there is again appreciable bias in the two-step approach, considerably less bias in the simultaneous IRT-SEM approach, and effectively negligible amounts when constrained IRT-SEM is used.

The results in Table 9.19 show the effect on parameter biases of varying the correlation coefficient between the two independent latent variables. It is clear that the bias of the two-step approach increases as the correlation coefficient decreases from 0.5 to -0.5 (e.g., an attenuated bias of 37% in R^2 increasing to 55% for the case of 2 categories). On the other hand, the simultaneous IRT-SEM approach is quite insensitive to variations in the correlation. It should be noted that the results shown in Tables 9.14-9.18 are all based on the condition where the correlation coefficient between the two independent latent variables is 0.5. In other words, the benefit of using the IRT-SEM approach compared to the two-step approach would be even more apparent for cases

where the correlation coefficient between the two independent latent variables is less than 0.5. A mathematical explanation of the relationship between the correlation coefficient of the two independent latent variables and the magnitude of the parameter bias is given in Appendix C. It assumes continuously measured items but nevertheless serves as an indication of the expected trend when only discrete variables are used. The results based on the constrained IRT-SEM approach are also reported in the highlighted bottom entry in each cell of Tables 9.19, as before, it can be seen that this method yields the smallest bias of all, even when sample size is only 300.

In summary, the simulation results for the multiple latent regression are consistent with those for the simple latent regression shown in Section 9.2.1. The bias in R^2 arising from the use of IRT- and CTT-scores in regression is very similar in general and unacceptably large for the most part. Simultaneous IRT-SEM yields smaller biases compared to the two-step approach, generally reaching acceptable levels. In all cases studied, the constrained IRT-SEM approach is insensitive to all simulation conditions and yields negligible biases in R^2 and in regression parameter estimates.

Table 9.14: Effects of scale length and number of categories on bias in the multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. (γ_1, γ_2) = (0.40825, 0.40825), True Correlation of two latent independent variables (ζ_1, ζ_2) = 0.5, CD=0.6, Symmetrically distributed items, Sample size=300.					
Scale Length	2 categories	3 categories	5 categories	7 categories	
05Y: 05x ₁ -05x ₂	0.882	0.882	0.882	0.882	0.882
	-37.43 / -16.69 / -18.56	-26.84 / -12.01 / -12.46	-21.71 / -9.31 / -10.31	-19.96 / -8.58 / -9.57	
	-37.71 / -17.36 / -18.39	-26.97 / -12.37 / -12.09	-21.82 / -9.53 / -9.85	-20.01 / -8.80 / -8.77	
10Y: 10x ₁ -10x ₂	8.44 / 2.63 / 4.54	6.41 / 2.17 / 3.71	2.64 / 0.26 / 2.09	3.96 / 1.21 / 2.34	
	1.75 / -0.94 / 1.38	1.54 / -0.38 / 1.21	0.02 / -1.29 / 0.77	0.64 / -0.67 / 0.67	
20Y: 20x ₁ -20x ₂	0.938	0.938	0.938	0.938	0.938
	-22.71 / -9.51 / -10.56	-14.40 / -6.03 / -6.47	-11.53 / -4.49 / -5.69	-10.32 / -4.19 / -5.23	
	-23.82 / -10.37 / -10.54	-14.68 / -6.40 / -6.04	-11.51 / -4.55 / -5.16	-10.19 / -4.04 / -4.40	
30Y: 30x ₁ -30x ₂	2.78 / -0.38 / 2.51	1.49 / -0.62 / 1.73	0.99 / -0.67 / 1.41	0.12 / -1.26 / 1.24	
	-0.70 / -1.90 / 0.21	-0.41 / -1.63 / 0.60	-0.78 / -1.46 / 0.21	-1.12 / -1.82 / 0.33	
20Y: 20x ₁ -20x ₂	0.968	0.968	0.968	0.968	0.968
	-13.39 / -5.42 / -5.97	-7.60 / -2.77 / -3.98	-6.36 / -2.22 / -3.38	-5.82 / -2.05 / -3.26	
	-15.45 / -6.64 / -6.35	-7.42 / -2.96 / -3.06	-5.37 / -1.91 / -2.43	-4.50 / -1.62 / -1.89	
30Y: 30x ₁ -30x ₂	1.88 / 0.87 / 1.29	-0.36 / 0.55 / -0.70	-0.67 / 0.01 / -0.50	-0.26 / 0.13 / -0.13	
	-0.07 / -0.13 / 0.38	-0.95 / 0.18 / -0.94	-1.09 / -0.21 / -0.65	-0.98 / -0.26 / -0.40	
30Y: 30x ₁ -30x ₂	0.978	0.978	0.978	0.978	0.978
	-10.46 / -3.99 / -4.96	-5.83 / -2.27 / -3.40	-5.18 / -2.04 / -2.83	-4.87 / -1.94 / -2.99	
	-12.84 / -5.46 / -5.27	-4.96 / -2.13 / -1.87	-3.65 / -1.53 / -1.44	-2.90 / -1.22 / -1.04	
30Y: 30x ₁ -30x ₂	4.76 / 1.38 / 3.64	5.01 / 1.36 / 3.83	4.02 / 1.09 / 3.25	3.66 / 1.16 / 2.76	

Note: 1. In each cell, the first number is Cronbach's alpha of the dependent latent variable; the subsequent three numbers are the %bias in R^2 and %bias in regression coefficients; these three statistics are reported for IRT approach, CTT approach, simultaneous IRT-SEM approach, and constrained IRT-SEM approach. 2. Standard error of the regression estimates ranges from 0.002-0.005 (in IRT approach), from 0.002-0.005 (in CTT approach), from 0.005-0.01 (in simultaneous IRT-SEM approach), and from 0.004-0.009 (in constrained IRT-SEM approach).

Table 9.15: Effect of sample size on bias in the multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\xi_1, \xi_2) = 0.5$, Scale length = 0.5, $CD = 0.6$, Symmetrically distributed items.					
Sample size	2 categories	3 categories	5 categories	7 categories	
300	0.775 -37.43 / -16.69 / -18.56 -37.71 / -17.36 / -18.39 8.44 / 2.63 / 4.54 1.75 / -0.94 / 1.38	0.823 -26.84 / -12.01 / -12.46 -26.97 / -12.37 / -12.09 6.41 / 2.17 / 3.71 1.54 / -0.38 / 1.21	0.861 -21.71 / -9.31 / -10.31 -21.82 / -9.53 / -9.85 2.64 / 0.26 / 2.09 0.02 / -1.29 / 0.77	0.871 -19.96 / -8.58 / -9.57 -20.01 / -8.80 / -8.77 3.96 / 1.21 / 2.34 0.64 / -0.67 / 0.67	
500	0.777 -37.77 / -16.72 / -18.95 -38.16 / -16.76 / -19.49 6.41 / 3.54 / 2.02 1.77 / 0.99 / -0.11	0.824 -27.79 / -10.76 / -15.51 -27.80 / -10.15 / -15.42 4.06 / 2.31 / 1.41 0.89 / 0.67 / -0.33	0.861 -21.88 / -7.93 / -12.45 -21.86 / -7.26 / -12.47 1.94 / 0.97 / 0.85 0.06 / -0.26 / -0.04	0.871 -20.08 / -7.57 / -11.06 -20.19 / -6.93 / -11.14 2.94 / 1.43 / 1.21 0.72 / 0.09 / 0.13	
800	0.777 -38.21 / -18.18 / -17.72 -38.65 / -18.69 / -18.03 3.10 / 2.05 / 0.80 0.40 / 0.62 / -0.48	0.824 -28.29 / -12.81 / -13.01 -28.25 / -12.56 / -13.13 2.75 / 2.31 / 0.48 0.60 / 1.16 / -0.62	0.861 -22.30 / -10.52 / -9.55 -22.30 / -10.19 / -9.68 1.05 / 1.02 / 0.28 0.00 / 0.38 / -0.18	0.872 -20.43 / -9.51 / -8.75 -20.43 / -9.24 / -8.79 1.38 / 1.02 / 0.40 0.32 / 0.33 / -0.04	
2000	0.776 -38.99 / -19.11 / -17.46 -39.44 / -19.39 / -17.65 -1.08 / -0.38 / -0.58 -0.87 / -0.26 / -0.45	0.824 -28.50 / -12.96 / -12.32 -28.51 / -12.98 / -12.55 0.14 / -0.09 / 0.38 -0.35 / -0.43 / 0.21	0.860 -22.53 / -10.28 / -9.35 -22.68 / -10.25 / -9.52 -0.37 / -0.31 / 0.16 -0.43 / -0.38 / 0.18	0.871 -20.87 / -9.46 / -8.60 -21.02 / -9.42 / -8.77 -0.48 / -0.50 / 0.18 -0.42 / -0.53 / 0.28	

Note: 1. See the descriptions in Table 9.14. 2. Standard error of the regression estimates ranges from 0.002-0.006 (in IRT approach), from 0.002-0.004 (in CTT approach), from 0.002-0.005 (in simultaneous IRT-SEM approach), and from 0.002-0.009 (in constrained IRT-SEM approach).

Table 9.16: Effect of measurement precision (CD) on bias in the multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\xi_1, \xi_2) = 0.5$, Scale length = 05y:05x ₁ -05x ₂ , Symmetrically distributed items, Sample size=300.					
CD	2 categories	3 categories	5 categories	7 categories	
0.3	0.548 -64.01 / -34.91 / -35.41 -64.00 / -35.37 / -35.71 28.24 / 5.52 / 19.36 4.64 / -4.67 / 7.85	0.611 -56.17 / -29.09 / -30.42 -56.17 / -29.27 / -30.10 14.26 / 7.83 / 5.30 2.09 / 1.63 / -0.40	0.653 -50.98 / -25.52 / -26.85 -51.05 / -25.84 / -26.36 7.55 / 2.83 / 4.45 0.89 / -0.97 / 1.41	0.665 -49.06 / -24.54 / -25.49 -49.16 / -24.73 / -24.97 7.26 / 3.27 / 3.86 0.47 / -0.53 / 0.72	
	0.775 -37.43 / -16.69 / -18.56 -37.71 / -17.36 / -18.39 8.44 / 2.63 / 4.54 1.75 / -0.94 / 1.38	0.823 -26.84 / -12.01 / -12.46 -26.97 / -12.37 / -12.09 6.41 / 2.17 / 3.71 1.54 / -0.38 / 1.21	0.861 -21.71 / -9.31 / -10.31 -21.82 / -9.53 / -9.85 2.64 / 0.26 / 2.09 0.02 / -1.29 / 0.77	0.871 -19.96 / -8.58 / -9.57 -20.01 / -8.80 / -8.77 3.96 / 1.21 / 2.34 0.64 / -0.67 / 0.67	
0.6					

Note: 1. See the descriptions in Table 9.14. 2. Standard error of the regression estimates is 0.005 (in both IRT and CTT approaches), ranges from 0.005-0.02 (in simultaneous IRT-SEM approach), and 0.001 (in constrained IRT-SEM approach).

Table 9.17: Effect of symmetry of item's marginal distribution on bias in the multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\xi_1, \xi_2) = 0.5$, Scale length = 0.5, $\gamma = 0.05x_1 - 0.05x_2$, $CD = 0.6$, Sample size = 500.					
Distribution	2 categories	3 categories	5 categories	7 categories	
Symmetric	0.777 -37.77 / -16.72 / -18.95 -38.16 / -16.76 / -19.49 6.41 / 3.54 / 2.02 1.77 / 0.99 / -0.11	0.824 -27.79 / -10.76 / -15.51 -27.80 / -10.15 / -15.42 4.06 / 2.31 / 1.41 0.89 / 0.67 / -0.33	0.861 -21.88 / -7.93 / -12.45 -21.86 / -7.26 / -12.47 1.94 / 0.97 / 0.85 0.06 / -0.26 / -0.04	0.871 -20.08 / -7.57 / -11.06 -20.19 / -6.93 / -11.14 2.94 / 1.43 / 1.21 0.72 / 0.09 / 0.13	
Skewed (level = 2)	0.731 -52.26 / -26.55 / -27.32 -50.97 / -24.10 / -27.59 7.85 / 3.42 / 4.05 1.65 / 0.11 / 1.11	0.783 -44.93 / -21.42 / -22.72 -43.73 / -18.93 / -23.42 5.42 / 2.95 / 2.05 1.51 / 1.07 / 0.09	0.803 -39.63 / -18.01 / -20.06 -41.17 / -17.70 / -21.71 3.53 / 2.39 / 0.99 0.09 / 0.70 / -0.82	0.829 -33.22 / -13.81 / -16.88 -35.85 / -14.21 / -19.18 4.46 / 3.15 / 0.94 1.01 / 1.43 / -0.85	

Note: 1. See the descriptions in Table 9.14. 2. Standard error of the regression estimates ranges from 0.002-0.004 (in IRT approach), from 0.002-0.005 (in CTT approach), from 0.007-0.01 (in simultaneous IRT-SEM approach), and from 0.006-0.008 (in constrained IRT-SEM approach).

Table 9.18: Effect of scale length and number of categories on bias in the multiple latent regression under a realistic situation

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\xi_1, \xi_2) = 0.5$, $CD=0.4-0.6$ (loadings: $0.632-0.775$), Mixed items: Symmetric, Uniform, and Skewed, Sample size = 300.					
Scale Length	2 categories	3 categories	5 categories	7 categories	
05y: 05x ₁ -05x ₂	0.689	0.740	0.774	0.787	
	-45.88 / -22.62 / -22.25	-36.68 / -17.74 / -16.96	-30.42 / -14.64 / -14.38	-27.73 / -12.33 / -13.80	
	-46.22 / -23.90 / -22.02	-37.99 / -18.79 / -17.39	-33.67 / -16.49 / -15.01	-31.03 / -14.04 / -14.43	
	11.01 / 4.13 / 5.50	9.95 / 4.54 / 4.74	6.17 / 3.69 / 2.22	6.12 / 2.80 / 3.05	
	1.28 / -0.75 / 0.55	2.06 / 0.62 / 0.87	0.63 / 0.50 / -0.36	1.20 / 0.11 / 0.67	
10y: 10x ₁ -10x ₂	0.817	0.853	0.875	0.883	
	-29.04 / -13.22 / -13.45	-20.99 / -8.99 / -9.48	-16.94 / -7.33 / -7.80	-15.37 / -6.16 / -7.52	
	-29.69 / -13.86 / -13.08	-21.89 / -9.61 / -9.54	-19.09 / -8.57 / -8.01	-17.39 / -7.21 / -7.52	
	4.72 / 0.67 / 3.29	1.86 / -0.94 / 2.68	1.02 / -0.62 / 1.56	0.10 / -0.60 / 0.65	
	0.25 / -1.29 / 0.50	-0.52 / -2.05 / 1.14	-1.08 / -1.58 / 0.23	-1.55 / -1.21 / -0.55	
20y: 20x ₁ -20x ₂	0.899	0.922	0.933	0.938	
	-16.40 / -6.53 / -7.37	-11.60 / -4.19 / -5.48	-8.85 / -3.39 / -4.06	-7.90 / -2.72 / -3.91	
	-17.35 / -7.33 / -7.39	-12.30 / -4.88 / -5.33	-9.88 / -4.13 / -3.92	-8.81 / -3.50 / -3.58	
	3.63 / 2.34 / 1.07	0.42 / 1.51 / -1.24	0.52 / 1.56 / -1.21	---	
	0.62 / 0.65 / -0.23	-0.96 / 0.80 / -1.85	-0.82 / 0.72 / -1.68	---	

Note: 1. See the descriptions in Table 9.14. 2. Standard error of the regression estimates ranges from 0.004-0.006 (in IRT approach), from 0.004-0.005 (in CTT approach), from 0.006-0.01 (in simultaneous IRT-SEM approach), and from 0.005-0.009 (in constrained IRT-SEM approach). 3. “---” denotes results are not available from the analysis due to too few observations in some categories.

Table 9.19: Effect of correlation between two independent latent variables on bias in the multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, Scale length = $05y:05x_1-05x_2$, $CD=0.6$, Symmetrically distributed items, Sample size=300.					
Correlation	2 categories	3 categories	5 categories	7 categories	
0.5	0.775	0.823	0.861	0.871	
	-37.43 / -16.69 / -18.56	-26.84 / -12.01 / -12.46	-21.71 / -9.31 / -10.31	-19.96 / -8.58 / -9.57	
	-37.71 / -17.36 / -18.39	-26.97 / -12.37 / -12.09	-21.82 / -9.53 / -9.85	-20.01 / -8.80 / -8.77	
	8.44 / 2.63 / 4.54	6.41 / 2.17 / 3.71	2.64 / 0.26 / 2.09	3.96 / 1.21 / 2.34	
	1.75 / -0.94 / 1.38	1.54 / -0.38 / 1.21	0.02 / -1.29 / 0.77	0.64 / -0.67 / 0.67	
0.3	0.777	0.823	0.860	0.870	
	-39.36 / -19.54 / -20.58	-27.91 / -13.75 / -13.94	-23.00 / -11.12 / -11.59	-21.33 / -10.19 / -10.82	
	-39.56 / -20.17 / -20.44	-28.01 / -14.11 / -13.30	-23.12 / -11.19 / -11.04	-21.37 / -10.35 / -10.10	
	7.19 / 2.71 / 3.05	6.08 / 2.39 / 3.07	3.90 / 0.97 / 2.62	4.16 / 1.66 / 1.98	
	1.06 / -0.33 / -0.01	1.48 / -0.05 / 0.95	0.29 / -0.94 / 0.84	0.91 / -0.14 / 0.43	
0.1	0.777	0.823	0.860	0.870	
	-42.02 / -22.92 / -23.78	-30.37 / -16.30 / -16.61	-24.17 / -12.97 / -12.90	-22.57 / -11.82 / -12.37	
	-42.33 / -23.49 / -24.10	-30.52 / -16.61 / -16.13	-24.32 / -13.16 / -12.44	-22.63 / -12.02 / -11.54	
	8.51 / 3.69 / 3.35	7.64 / 3.18 / 3.92	4.61 / 1.48 / 2.64	4.35 / 1.36 / 2.20	
	1.05 / -0.03 / -0.11	1.99 / 0.31 / 1.34	0.44 / -0.64 / 0.64	0.65 / -0.64 / 0.54	
0	0.778	0.823	0.860	0.870	
	-43.54 / -24.84 / -25.63	-31.53 / -17.77 / -17.89	-25.35 / -14.04 / -14.39	-23.33 / -12.83 / -13.25	
	-43.88 / -25.45 / -26.18	-31.68 / -18.09 / -17.58	-25.52 / -14.27 / -13.90	-23.41 / -13.08 / -12.50	
	8.68 / 3.86 / 3.44	8.03 / 3.18 / 4.22	4.42 / 1.48 / 2.38	4.34 / 1.50 / 2.08	
	1.12 / 0.14 / -0.02	1.60 / 0.06 / 1.20	0.35 / -0.58 / 0.50	0.65 / -0.46 / 0.48	

Note: 1. See the descriptions in Table 9.14. 2. Standard error of the regression estimates ranges from 0.004-0.006 (in IRT approach), from 0.004-0.005 (in CTT approach), from 0.006-0.015 (in simultaneous IRT-SEM approach), and from 0.004-0.009 (in constrained IRT-SEM approach).

Table 9.19: Effect of correlation between two independent latent variables on bias in the multiple latent regression (continued)

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, Scale length = $0.5\gamma_1 - 0.5\gamma_2$, $CD=0.6$, Symmetrically distributed items, Sample size=300.					
Correlation	2 categories	3 categories	5 categories	7 categories	
-0.1	0.778	0.823	0.861	0.870	
	-45.08 / -26.92 / -27.87	-33.01 / -19.25 / -19.68	-26.54 / -15.55 / -15.75	-24.25 / -14.00 / -14.41	
	-45.47 / -27.58 / -28.32	-33.17 / -19.63 / -19.47	-26.72 / -15.80 / -15.22	-24.38 / -14.28 / -13.73	
	10.19 / 4.64 / 4.24	8.16 / 3.10 / 4.24	4.59 / 1.74 / 2.31	4.75 / 1.81 / 2.40	
	1.63 / 0.29 / 0.52	1.33 / -0.14 / 1.02	0.32 / -0.41 / 0.37	0.80 / -0.26 / 0.71	
-0.3	0.778	0.823	0.861	0.871	
	-49.40 / -32.70 / -33.47	-36.38 / -23.70 / -23.73	-29.40 / -19.05 / -19.20	-27.00 / -17.37 / -17.62	
	-49.88 / -33.49 / -33.95	-36.64 / -24.07 / -23.68	-29.61 / -19.33 / -18.92	-27.17 / -17.78 / -17.13	
	13.54 / 6.11 / 6.16	10.08 / 4.42 / 5.25	5.41 / 2.41 / 2.76	5.63 / 2.36 / 2.83	
	1.96 / 0.63 / 0.90	1.73 / 0.53 / 1.37	0.26 / -0.07 / 0.48	0.90 / -0.02 / 0.82	
-0.5	0.778	0.823	0.861	0.871	
	-55.40 / -40.76 / -41.31	-41.65 / -30.55 / -30.48	-34.26 / -24.89 / -25.03	-32.02 / -22.81 / -23.26	
	-55.93 / -41.52 / -41.90	-41.90 / -30.89 / -30.42	-34.50 / -25.13 / -24.83	-32.23 / -23.23 / -22.95	
	13.68 / 5.99 / 5.99	12.42 / 5.73 / 6.60	7.63 / 3.32 / 3.96	6.32 / 2.94 / 3.05	
	1.14 / 0.38 / 0.45	2.27 / 1.23 / 2.11	0.56 / 0.14 / 0.97	0.59 / 0.25 / 0.73	

Note: 1. See the descriptions in Table 9.14. 2. Standard error of the regression estimates ranges from 0.004-0.006 (in IRT approach), from 0.004-0.005 (in CTT approach), from 0.006-0.015 (in simultaneous IRT-SEM approach), and from 0.004-0.009 (in constrained IRT-SEM approach).

9.3 Meta-Models for the Simulation Results from the Two-Step and IRT-SEM Approaches

This section serves as an aid to summarizing and understanding the complete set of simulation results for the second research question. Linear regression models (ANOVA) are used as meta-models to explicitly describe the relationships between the performance criterion (i.e., the percentage bias of R^2 in latent variable regression) and the experimental factors (e.g., number of items, sample size, number of categories, etc). Illustrations of how to use the meta-models to predict bias in R^2 when performing latent regression analysis will also be presented, and plots of the main trends from the meta-models will be shown. The statistical software SPSS 11.0 for windows was used for the analysis in this section.

9.3.1 Specifications of the Meta-Models

The raw data for the meta-models were the values of the percentage bias in R^2 obtained from the simulation runs for each setting studied. For each setting, these bias estimates are aggregates of individual Monte Carlo replications, so that the modeling exercises did not include a between-replications component of error. All error estimates thus consisted of higher-order interactions not explicitly included in the meta-model. This approach does not affect the estimation of meta-model regression parameters, only their standard errors. Provided the omitted interactions are indeed zero, the resulting error estimates are still valid. Even if this assumption is not strictly correct, it will not affect the meta-modeling effort as the objective was to include only those terms in the model that have an appreciable (and visible) effect on trends, i.e., statistical significance of specific terms was not the primary criterion for their inclusions.

The meta-models were in fact derived using two primary criteria. First, graphs based on both raw data and fitted values were examined to identify trends. Then, the effect sizes of the main effects, two-factor, and three-factor interactions were examined in order to include those features of the relationships that were important from a practical point of view. Four-factor or higher interactions were ignored as indicated above due to the principle of parsimony. Partial eta-squared was used as a measure of effect size, defined as the ratio of the variation accounted for by the individual factor (or effect) to the sum of the variation accounted for by the factor (or effect) and the variation unaccounted for by the model as a whole (SPSS 11.0, 2001).

The number of observations in each meta-model is equal to the number of simulation settings (experimental treatments). Each observation of the performance criterion (treated as the dependent variable) in a meta-model is generated through the following procedures: (a) obtain the $\hat{R}^2$ from the latent variable regression for each simulated sample, (b) average the percentage biases (defined by $\{[(\hat{R}^2 - R^2)/R^2] 100\}$), across all simulated samples under the same simulation setting. In other words, each observation is the average of R^2 biases obtained from 100 randomly simulated samples of N respondents.

Four meta-models were fitted using the simulation results. Results for the IRT and CTT two-step methods are included in the same meta-model, and a separate meta-model was set up for the simultaneous IRT-SEM approach. In both cases, separate meta-models were then constructed for the simple and multiple latent regressions. The meta-models of percentage R^2 bias derived from the combined IRT and CTT two-step approaches under the simple and multiple latent regressions are specified, respectively, as follows:

$$B_S(R^2 : 2 \text{ step}) = \beta_0 + \beta_1 LM + \beta_2 MD_1 + \beta_3 MD_2 + \beta_4 MD_3 + \beta_5 RQ + \beta_6 CD + \beta_7 CT + \beta_8 IM + \beta_9 (1/IM) + \beta_{10} (IM * CD) + \text{error}$$

and

$$B_M(R^2 : 2 \text{ step}) = \beta_0 + \beta_1 LM + \beta_2 MD_1 + \beta_3 MD_2 + \beta_4 MD_3 + \beta_5 RQ + \beta_6 CD + \beta_7 CT + \beta_8 IM + \beta_9 (1/IM) + \beta_{10} (IM * CD) + \beta_{11} CR + \text{error}$$

These are ANOVA models on the percentage *downward* bias of R^2 , based on 480 and 2880 observations for simple and multiple latent regression cases, respectively. The descriptions of the variables in the model and the detailed simulation settings used for generating the models are reported in Tables 9.20, 9.21 and 9.22.

Table 9.20: Specifications of the meta-model for the two-step approach

Dependent variable: Percentage downward bias of R^2 in structural regression model		
Exploratory Variable	Description	Value / Level
LM	The latent regression method	1: CTT-based 0: IRT-based
MD ₁ MD ₂ MD ₃	The Marginal frequency distribution of the observable discrete indicators*	Uniform: MD ₁ =1, MD ₂ =0, MD ₃ =0 Mixed: MD ₁ =0, MD ₂ =1, MD ₃ =0 Skewed: MD ₁ =0, MD ₂ =0, MD ₃ =1 Symmetric: MD ₁ =0, MD ₂ =0, MD ₃ =0
RQ	The coefficient of determination, R^2 , of the structural regression model	Simple latent regression: 0.1, 0.5 Multiple latent regression: 0.3, 0.5
CD	The coefficient of determination (CD) of the measurement model	0.3, 0.6
CT	The number of item response options (categories)	2, 3, 5, 7
IM	The number of items in the scale	5, 10, 20, 30
CR	The correlation between two latent explanatory variables in the multiple latent regression model	0.5, -0.5, 0.3, -0.3, 0.1, -0.1

* Uniform: all the items are uniformly distributed; Mixed: the marginal frequency distributions of the observable items are symmetric for 2/5 of items, uniform for 1/5 of items, skewed with skewness level 1 for 1/5 of items, skewed with level 2 for 1/5 of items; Skewed: all the items are asymmetrically distributed with skewness level 2; Symmetric: all the items are symmetrically distributed.

Table 9.21: Simulation settings of the meta-model for the two-step approach in the simple latent regression

Marginal frequency Distribution*	R^2	CD	# of categories	# of items	Settings
Uniform	0.1, 0.5	0.3, 0.6	3, 5, 7	5, 10, 20, 30	2×2×3×4 =48
Mixed	0.1, 0.5	0.3, 0.6	2, 3, 5, 7	5, 10, 20, 30	2×2×4×4 =64
Skewed	0.1, 0.5	0.3, 0.6	2, 3, 5, 7	5, 10, 20, 30	2×2×4×4 =64
Symmetric	0.1, 0.5	0.3, 0.6	2, 3, 5, 7	5, 10, 20, 30	2×2×4×4 =64
Total number of simulation settings for IRT- and CTT-based two-step approaches, i.e., total number of observations used in the meta-model					240×2 =480

* See the description in Table 9.20

Table 9.22: Simulation settings of the regression summary for the two-step approach in the multiple latent regression

Marginal frequency Distribution*	R^2	CD	# of categories	# of items	Corr.**	Settings
Uniform	0.3	0.3	3	5	0.5, -0.5	2×2×3×4×6 =288
	0.5	0.6	5	10	0.3, -0.3	
			7	20	0.1, -0.1	
				30		
Mixed	0.3	0.3	2	5	0.5, -0.5	2×2×4×4×6 =384
	0.5	0.6	3	10	0.3, -0.3	
			5	20	0.1, -0.1	
			7	30		
Skewed	0.3	0.3	2	5	0.5, -0.5	2×2×4×4×6 =384
	0.5	0.6	3	10	0.3, -0.3	
			5	20	0.1, -0.1	
			7	30		
Symmetric	0.3	0.3	2	5	0.5, -0.5	2×2×4×4×6 =384
	0.5	0.6	3	10	0.3, -0.3	
			5	20	0.1, -0.1	
			7	30		
Total number of simulation settings for IRT- and CTT-based two-step approaches, i.e., total number of observations used in the meta-model						1440×2 =2880

* See the description in Table 9.20.

** Corr. denoted correlation coefficient between the exploratory latent variables

The meta-models for the simulation results derived from the simultaneous IRT-SEM approach under the simple and multiple latent regressions are specified, respectively, as follows:

$$\begin{aligned} B_S(R^2 : \text{IRT} - \text{SEM}) = & \beta_0 + \beta_1 CD_1 + \beta_2 CD_2 + \beta_3 CT + \beta_4 IM + \beta_5 (1/IM) \\ & + \beta_6 SS + \beta_7 (1/SS) + \beta_8 [(1/IM) * (1/SS)] + \beta_9 (CD_1 * CT) \\ & + \beta_{10} (CD_2 * CT) + \beta_{11} (CD_1 * IM) + \beta_{12} (CD_2 * IM) \\ & + \beta_{13} [(CD_1 * (1/SS))] + \beta_{14} [CD_2 * (1/SS)] + \beta_{15} [CT * (1/IM)] \\ & + \beta_{16} [CT * (1/SS)] + \text{error} \end{aligned}$$

and

$$\begin{aligned} B_M(R^2 : \text{IRT} - \text{SEM}) = & \beta_0 + \beta_1 CD_1 + \beta_2 CD_2 + \beta_3 CT + \beta_4 IM + \beta_5 (1/IM) + \beta_6 SS \\ & + \beta_7 (1/SS) + \beta_8 [(1/IM) * (1/SS)] + \beta_9 (CD_1 * CT) \\ & + \beta_{10} (CD_2 * CT) + \beta_{11} (CD_1 * IM) + \beta_{12} (CD_2 * IM) \\ & + \beta_{13} [(CD_1 * (1/SS))] + \beta_{14} [CD_2 * (1/SS)] + \beta_{15} [CT * (1/IM)] \\ & + \beta_{16} [CT * (1/SS)] + \beta_{17} CR + \beta_{18} (CR * CD_1) + \beta_{19} (CR * CD_2) \\ & + \beta_{20} [(CR * (1/SS))] + \text{error}. \end{aligned}$$

These models are ANOVA's with percentage *upward* bias of R^2 as the dependent variables, based on 192 and 324 observations for the single and multiple cases, respectively. The descriptions of the variables in the model and the detailed simulation settings used for generating the models are reported in Tables 9.23, 9.24 and 9.25.

Table 9.23: Specification of the meta-model for the simultaneous IRT-SEM approach

Dependent variable: Percentage upward bias of R^2 in structural regression model		
Exploratory Variable	Description	Value / Level
CD ₁ CD ₂	The coefficient of determination (CD) of the measurement model and the marginal frequency distribution of the observable discrete indicators*	CD=0.3 and symmetric items: CD ₁ =1, CD ₂ =0 CD varying uniformly between 0.4 and 0.6 and mixed items: CD ₁ =0, CD ₂ =1 CD=0.6 and symmetric items: CD ₁ =0, CD ₂ =0
CT	The number of item response options (categories)	2, 3, 5, 7
IM	The number of items in the scale	Simple latent regression: 5, 10, 20, 30 Multiple latent regression: 5, 10, 20
SS	Sample size	Simple latent regression: 300, 500, 800, 2000 Multiple latent regression: 300, 800, 2000
CR	The correlation between two latent explanatory variables in the multiple latent regression model	0.5, -0.5, 0.3, -0.3, 0.1, -0.1

* See the description in Table 9.20.

Table 9.24: Simulation settings of the regression summary for the IRT-SEM approach in the simple latent regression

Marginal frequency Distribution*	R^2	CD	# of categories	# of items	Sample size	Settings
Mixed	0.5	(Uniform from 0.4 to 0.6)	2	5	300	1×1×4×4×4 =64
			3	10	500	
			5	20	800	
			7	30	2000	
Symmetric	0.5	0.3 0.6	2	5	300	1×2×4×4×4 =128
			3	10	500	
			5	20	800	
			7	30	2000	
Total number of simulation settings for simultaneous IRT-SEM approach, i.e., total number of observations used in the meta-model						192

* See the description in Table 9.20.

Table 9.25: Simulation settings of the regression summary for the simultaneous IRT-SEM approach in the multiple latent regression

Marginal frequency Distribution*	R ²	CD	# of categories	# of items	Sample size	Corr. **	Settings
Mixed	0.5	(Uniform from 0.4 to 0.6)	2, 3, 5, 7	5 10 20	300 800 2000	0.5 0 -0.5	1×1×1×4×3×3×3 =108
Symmetric	0.5	0.3 0.6	2, 3, 5, 7	5 10 20	300 800 2000	0.5 0 -0.5	1×2×1×4×3×3×3 =216
Total number of simulation settings for simultaneous IRT-SEM approach, i.e., total number of observations used in the meta-model							216-4 =324***

* See the description in Table 9.20.

** Corr. denoted correlation coefficient between the exploratory latent variables.

*** There are 4 setting where Mplus failed to generate the results due to one or more categories having too few observations.

Four fitted meta-models are presented in Table 9.26. All four models show good fit to the data, with $R^2 > 0.950$ for the two-step approach, and $R^2 > 0.730$ for the IRT-SEM approach. It is clear that the effect of the true R^2 in the structural regression model in the two-step approach (variable RQ) is negligible, with partial eta-squared < 0.001 for the simple latent regression and 0.007 for the multiple latent regression. The effect of the sample size is not included in the fitted meta-models for the two-step method because here sample size was not considered as a simulation condition for the meta-modeling, as the simulation results in section 9.2 show that the two-step approach is not sensitive to sample size.

Table 9.26: Meta-models for the two-step and simultaneous IRT-SEM approaches

Two-Step Approach					Simultaneous IRT-SEM Approach				
Dependent Variable: Percentage downward bias in R^2	Simple Latent Regression		Multiple Latent Regression		Dependent Variable: Percentage upward bias in R^2	Simple Latent Regression		Multiple Latent Regression	
Adjusted R^2	0.966		0.959		Adjusted R^2	0.759		0.732	
Independent Variable	$\hat{\beta}$	Partial Eta-squared	$\hat{\beta}$	Partial Eta-squared	Independent Variable	$\hat{\beta}$	Partial Eta-squared	$\hat{\beta}$	Partial Eta-squared
Intercept	70.215	0.820	69.328	0.737	Intercept	-4.973	0.057	-9.074	0.024
LM	1.493	0.048	1.951	0.058	CD1	-0.466	0.001	8.474	0.058
MD1	1.922	0.036	2.226	0.035	CD2	-0.415	0.001	2.068	0.004
MD2	3.327	0.101	3.680	0.090	CT	0.717	0.074	1.358	0.044
MD3	15.990	0.721	18.498	0.713	IM	0.09291	0.052	0.146	0.007
RQ	0.320	0.000	3.296	0.007	1/IM	16.534	0.026	22.814	0.007
CD	-99.030	0.837	-100.002	0.791	SS	0.0001664	0.001	0.0006311	0.003
CT	-2.220	0.608	-2.252	0.534	1/SS	775.501	0.017	651.522	0.002
IM	-1.542	0.586	-1.528	0.500	(1/IM) $\times$ (1/SS)	16379.652	0.296	27524.381	0.195
1/IM	151.307	0.637	151.068	0.579	CD ₁ $\times$ CT	-0.447	0.040	-0.959	0.032
IM $\times$ CD	2.593	0.557	2.561	0.557	CD ₂ $\times$ CT	-0.0946	0.002	-0.327	0.004
CR			-13.495	0.469	CD ₁ $\times$ IM	25.122	0.131	-1.960	0.000
					CD ₂ $\times$ IM	7.155	0.012	3.296	0.000
					CD ₁ $\times$ (1/SS)	1202.670	0.082	2371.760	0.074
					CD ₂ $\times$ (1/SS)	415.070	0.011	402.389	0.002
					CT $\times$ (1/IM)	-4.537	0.098	-7.510	0.044
					CT $\times$ (1/SS)	-212.855	0.058	-501.546	0.069
					CR			4.550	0.039
					CR $\times$ CD ₁			-17.635	0.337
					CR $\times$ CD ₂			-3.378	0.018
					CR $\times$ (1/SS)			-2237.112	0.065

9.3.2 An Application of the Meta-Models

One of the applications of the fitted meta-models in Table 9.26 is to predict the bias in R^2 due to the two-step and IRT-SEM methods. From the point of view of the researcher, after a particular latent variable regression has been carried out and a R^2 has been obtained, it is extremely useful to be able to get a sense of how different could this estimate be from the true population R^2 (i.e., bias) as a result of the inherent inadequacies of the method used. In this section, we demonstrate how a researcher can use characteristics of their analysis (e.g., method used, sample size, CD , etc) to predict how much bias might have resulted in their analysis.

Suppose a researcher is planning to conduct multiple latent regression analysis with two exploratory latent variables and is considering using either a CTT-based two-step method or an IRT-SEM approach. The performance criterion of the methods is the bias of R^2 in the structural latent regression. To use the meta-models as a tool to evaluate the performance of each method, the researcher needs to use Tables 9.20 and 9.23 (i.e., the descriptions of the variables in the meta-models for the two-step and IRT-SEM methods, respectively), Table 9.26 (i.e., the fitted meta-models), or a spreadsheet like Table 9.27, and to input the values for the following variables into the fitted meta-models:

- (1) Intercept. Input 1 in both columns B and D for the intercept.

(2) Latent regression method. Since the CTT-based method is chosen for the two-step method, input 1 in column B for LM; otherwise, input 0 if IRT-based method is selected.

(3) True coefficient of determination, R^2 , of the structural regression model. This variable appears in the meta-models for the two-step method; however, the effect of this variable on bias is negligible with partial eta-squared's of less than 0.001 for simple latent regression and 0.007 for multiple latent regression in Table 9.26. And, of course, the true R^2 is unknown; therefore, the researcher can simply use the expected R^2 for the structural latent regression as input into the fitted meta-models. (Suppose expected R^2 is 0.5, then input 0.5 in column B of Table 9.27 as RQ).

(4) Sample size. This variable only appears in the IRT-SEM method in that the two-step method is not sensitive to sample size. (Suppose the sample size is 500, then input 500 for SS in column D.)

(5) Coefficient of determination (CD) of the measurement model. The researcher can run confirmatory factor analysis for each measurement model separately in order to obtain the CD for each indicator (i.e., square of indicator's standardized loading). Then, calculate the average of CD 's as the input in the meta-models for the two-step method. (Suppose the average of CD 's is 0.6, then input 0.6 in column B for CD .) For the IRT-SEM approach, the variable CD was treated as a dummy variable (see Table 9.23). (Input 0 in column D for both CD_1 and CD_2 .)

(6) Number of items in the scale. (Suppose each latent variable is measured by 5 items, then input 5 in both columns B and D for IM.)

(7) Number of item response options (categories). (Suppose 2 categories were used for each item, then input 2 in both columns B and D for CT.)

(8) Marginal frequency distribution of the discrete observable item indicators. This variable only appears in the meta-models for the two-step method. The meta-models for the IRT-SEM approach assume symmetric items. (Suppose all the items are symmetrically distributed, then input 0 in column B for all MD_1 , MD_2 and MD_3 .)

(9) Correlation between two latent explanatory variables in the multiple latent regression model. (Suppose the correlation is -0.2, then input -0.2 in both columns B and D for CR.)

Then, the researcher can predict the percentage of bias in R^2 by summing the column $A \times B$ for the CTT-based two-step method and the column $C \times D$ for the IRT-SEM method. It is shown in Table 9.27 that the researcher will get a downward bias of 41.38% in R^2 for the CTT-based method, but only a 6.54% upward bias for the IRT-SEM method. Thus if the researcher's CTT-based regression (with above characteristics) finds $R^2 = 50\%$, the true R^2 in the population may be as high as 85%; if IRT-SEM based regression had been used (with above characteristics) and again say $R^2 = 50\%$ from the regression, then the true population R^2 may really be 47%. These two calculations are based on the definition of percentage bias given in Equation (9.1), where, we may now input values of $\hat{R}^2$ and bias, and then solve for R^2 , as,

$$-41.38 = \left(\frac{0.50 - R^2}{R^2} \right) \times 100 \quad \text{and} \quad 6.54 = \left(\frac{0.50 - R^2}{R^2} \right) \times 100.$$

In terms of statistical decisions, this demonstrates a much greater risk of Type II errors when the CTT-based regression is used (i.e., decide there is no significant effect, only because our estimate is too low) compared to the risk of Type I errors from using the IRT-SEM regression (i.e., decide there is a significant effect only because our estimate is too high). Two more examples of predicting bias in R^2 are provided in Tables 9.28 and 9.29 for the IRT-based two-step method and the IRT-SEM approach, respectively, under different research scenarios.

Table 9.27: Illustrations of the application of meta-models (1)

Research scenarios:							
Symmetrically distributed items, CD=0.6, Sample size=500, 5 items in each scale, 2 categories in each item, Correlation coefficient between the two exploratory variables = -0.2.							
CTT-based two-step approach in multiple latent regression				Simultaneous IRT-SEM approach in multiple latent regression			
Independent variable	Reg. coef. (A)	Value for indep. variable (B)	A X B	Independent variable	Reg. Coef. (C)	Value for indep. variable (D)	C X D
Intercept	69.328	1	69.328	Intercept	-9.074	1	-9.074
LM	1.951	1	1.951	CD1	8.474	0	0
MD1	2.226	0	0	CD2	2.068	0	0
MD2	3.680	0	0	CT	1.358	2	2.716
MD3	18.498	0	0	IM	0.146	5	0.730
RQ	3.296	0.5	1.648	1/IM	22.814	1/5	4.563
CD	-100.002	0.6	-60.012	SS	0.0006311	500	0.316
CT	-2.252	2	-4.504	1/SS	651.522	1/500	1.303
IM	-1.528	5	-7.640	(1/IM) x (1/SS)	27524.381	(1/5) x (1/500)	11.010
1/IM	151.068	1/5	30.214	CD ₁ xCT	-0.959	0x2	0
IMxCD	2.561	5x0.6	7.683	CD ₂ xCT	-0.327	0x2	0
CR	-13.495	-0.2	2.699	CD ₁ xIM	-1.960	0x5	0
				CD ₂ xIM	3.296	0x5	0
				CD ₁ x (1/SS)	2371.760	0x (1/500)	0
				CD ₂ x (1/SS)	402.389	0x (1/500)	0
				CTX (1/IM)	-7.510	2x (1/5)	-3.004
				CTX (1/SS)	-501.546	2x (1/500)	-2.006
				CR	4.550	-0.2	-0.910
				CRxCD ₁	-17.635	0	0
				CRxCD ₂	-3.378	0	0
				CRx (1/SS)	-2237.112	-0.2x (1/500)	0.895
Predicted downward %bias in R ²			Sum: 41.38 (%)	Predicted upward %bias in R ²			Sum: 6.54 (%)

Table 9.28: Illustrations of the application of meta-models (2)

Research scenarios:			
Mixed items: symmetric, uniform, and skewed. Expected $R^2=0.5$, $CD=0.5$, 6 items in each scale, 7 categories in each item, correlation coefficient between the two exploratory variables = 0.3.			
IRT-based two-Step Approach			
Independent variable	Regression coefficient (A)	Value for independent variable (B)	A × B
Intercept	69.328	1	69.328
LM	1.951	0	0
MD1	2.226	0	0
MD2	3.680	1	3.680
MD3	18.498	0	0
RQ	3.296	0.5	1.648
CD	-100.002	0.5	-50.001
CT	-2.252	7	-15.764
IM	-1.528	6	-9.168
1/IM	151.068	1/6	25.178
IM×CD	2.561	6×0.5	7.683
CR	-13.495	0.3	-2.699
Predicted downward bias in R^2			Sum: 29.89 (%)

Table 9.29: Illustrations of the application of meta-models (3)

Research scenarios:			
Mixed items: symmetric, uniform, and skewed. CD=0.4-0.6, 6 items in each scale, sample size=200, 5 categories in each item, correlation coefficient between the two exploratory variables = 0.35.			
Simultaneous IRT-SEM Approach			
Independent variable	Regression coefficient (C)	Value for independent variable (D)	C × D
Intercept	-9.074	1	-9.074
CD1	8.474	0	0
CD2	2.068	1	2.068
CT	1.358	5	6.790
IM	0.146	6	0.876
1/IM	22.814	1/6	3.802
SS	0.0006311	200	0.126
1/SS	651.522	1/200	3.258
(1/IM) × (1/SS)	27524.381	(1/6) × (1/200)	22.937
CD ₁ × CT	-0.959	0 × 5	0
CD ₂ × CT	-0.327	1 × 5	-1.635
CD ₁ × IM	-1.960	0 × 6	0
CD ₂ × IM	3.296	1 × 6	0.549
CD ₁ × (1/SS)	2371.760	0 × (1/200)	0
CD ₂ × (1/SS)	402.389	1 × (1/200)	2.012
CT × (1/IM)	-7.510	5 × (1/6)	-6.258
CT × (1/SS)	-501.546	5 × (1/200)	-12.539
CR	4.550	0.35	1.593
CR × CD ₁	-17.635	0.35 × 0	0
CR × CD ₂	-3.378	0.35 × 1	-1.182
CR × (1/SS)	-2237.112	-0.35 × (1/200)	-3.915
Predicted upward bias in R ²			Sum: 9.41 (%)

9.3.3 Graphical Presentation of the Simulation Results based on the Meta-Models

In section 9.2, we have presented the simulation results in tables based on the actual values of biases found in simulation. Rather than using the actual values of biases, this section illustrates the simulation results using plots based on the predicted biases derived from the fitted meta-models. The purpose of this section is to show general trends of the effects of various simulation conditions, and to give an overall picture of how the two-step method compares with the IRT-SEM approach.

Plots 9.1-9.7 display the effect of various simulation conditions on the downward bias of R^2 in the two-step latent variable regression and the upward bias of R^2 in the IRT-SEM approach. In particular, Plot 9.1 demonstrates the effect of scale length for a sample of size 300. It is clear that the bias lines for the IRT and CTT two-step methods stay very close to each other for all conditions. As predicted, the absolute parameter bias with IRT and CTT regression decreases as number of items increases. The simultaneous IRT-SEM approach is sensitive to the number of items only when number of items is less than 10. In all cases the bias in the IRT-SEM estimates is less than that in either of the two-step methods, the relative benefits of IRT-SEM being particularly noticeable for 10 or fewer item. After 10 items, the absolute difference in bias between the two-step and IRT-SEM approaches decreases as the number of items increases. Trends are similar for both simple and multiple regressions. Note that the multiple regression case shown corresponds to the correlation coefficient of 0.5 between the

exploratory latent variables, so that the differences between the 2-step and IRT-SEM approaches are minimized in this situation.

Plot 9.2 shows that with 5 or 10 symmetric items, for both two-step approaches, the bias decreases as the number of categories increases, for a sample of size 300. The simultaneous IRT-SEM approach is not as sensitive to number of categories as are the two-step methods, especially for 10 items. The bias in the two-step methods is appreciably larger than in simultaneous IRT-SEM for a small number of items, for both simple and multiple regression cases.

It can be seen from Plot 9.3 that with 5 symmetrically distributed items and $CD=0.6$ in each measurement model, the bias of the simultaneous IRT-SEM method approaches zero for suitably large sample size, as expected. For 10 symmetric items, the bias lines are very close to the x-axis even for small sample size (e.g., 300). The effect of the sample size on bias for the two-step approach is not presented in this plot sample size was not included in the meta-models for the two-step method. Note also that no meta-models were constructed for constrained IRT-SEM. As can be seen from the tables in Section 9.2, this method results in negligible bias across all conditions. There is nothing of interest to model!

The results in Plot 9.4 show that with 5 symmetrically distributed items in each measurement model and a sample of size 300, the CTT and IRT two-step methods are much more sensitive to the CD 's in the measurement model than is the IRT-SEM approach. Again, the bias in the two-step method is much greater than that in IRT-SEM. Results are similar for both simple and multiple regressions. Plots 9.5 and 9.6 present

the effects of symmetry of an item's marginal distribution and the population R-square on the R^2 bias of the two-step methods for a sample of size 300, when 5 items with $CD=0.6$ in each measurement model are used. Plot 9.5 shows that the more skewed the marginal distribution of items, the more bias there is in the two-step approach. However, the biases are very close for all the symmetric, uniform, and mixed cases. It also can be seen in Plot 9.6 that the bias for the two-step methods are not related to the population R-square. The presentation of the effects of an item's marginal distribution and the population R^2 on biases in the IRT-SEM method is not available in the plot since these two factors were omitted from the meta-models for the IRT-SEM approach owing to their small effect sizes, i.e., neither the shape of the marginal distribution nor the value of population R^2 has any effect on R^2 bias for IRT-SEM.

The results in Plot 9.7 show the effect of correlation coefficient between the two exploratory latent variables on bias in R^2 . It is clear that with 5 symmetric distributed items and sample size of 300, the bias with the two-step approach increases as the correlation coefficient decreases from 0.5 to -0.5 (please see Appendix C for a mathematical explanation). On the other hand, the simultaneous IRT-SEM approach appears not to be as sensitive to the correlation as the two-step approach. It can be seen that there is again appreciable bias in the two-step approach and considerably less bias in the simultaneous IRT-SEM approach.

It is worth mentioning again that the results shown in Plots 9.1-9.5 are all based on the condition where the correlation coefficient between the two exploratory latent variables is 0.5. In other words, the difference in the absolute bias in R^2 between the

two-step and IRT-SEM approaches would be even greater when the correlation coefficient between the two exploratory latent variables is less than 0.5.

In summary, the bias in R^2 attributed to using IRT- and CTT-scores in regression is very similar in general and is generally of a magnitude that might cause serious concern to analysts. Nevertheless, the IRT-based latent regression method consistently does show slightly lower bias than the CTT-based method. The IRT-SEM appears much less biased than both two-step methods in all the situations considered. Recall, however, that the constrained IRT-SEM method (not shown in the plots) shows the lowest biases of all, across all conditions studied. We can also conclude that bias in the two-step approaches decreases as the scale length, number of categories, CD , extent of symmetry of items, and correlation coefficient between the two exploratory latent variables increase; on the other hand, the simultaneous IRT-SEM does not show as strong relationships with these variables, in general.

Plot 9. 1: Effect of scale length on R^2 bias in latent variable regression (for the two-step and simultaneous IRT-SEM methods)

A: Simple latent regression

True $R^2 = 0.5$, True Reg. Coef. = 0.707, Sample size = 300, $CD = 0.6$, Symmetrically distributed items.

B: Multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\zeta_1, \zeta_2) = 0.5$, $CD=0.6$, Symmetrically distributed items, Sample size=300.

Plot 9. 2: Effect of number of categories on R^2 bias in latent variable regression (for the two-step and simultaneous IRT-SEM methods)

A: Simple latent regression

True $R^2 = 0.5$, True Reg. Coef. = 0.707, Sample size = 300, $CD = 0.6$, Symmetrically distributed items.

B: Multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\zeta_1, \zeta_2) = 0.5$, $CD=0.6$, Symmetrically distributed items, Sample size=300.

Plot 9. 3: Effect of sample size on R^2 bias in latent variable regression (for the simultaneous IRT-SEM method)

A: Simple latent regression

True $R^2 = 0.5$, True Reg. Coef. = 0.707, $CD = 0.6$, Symmetrically distributed items.

B: Multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\zeta_1, \zeta_2) = 0.5$, $CD=0.6$, Symmetrically distributed items.

**Plot 9. 4: Effect of measurement precision (CD) on R^2 bias in latent variable regression
(for the two-step and simultaneous IRT-SEM methods)**

A: Simple latent regression

True $R^2 = 0.5$, True Reg. Coef. = 0.707, Sample size = 300, Scale length = 5y-5x, Symmetrically distributed items.

B: Multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\xi_1, \xi_2) = 0.5$, Scale length = 05y:05x₁-05x₂, Symmetrically distributed items, Sample size=300.

Note: For the IRT-SEM approach, the trends represent the fitted values of $CD=0.3$ and 0.6 only, in the predicted meta-models.

Plot 9. 5: Effect of symmetry of item's marginal distribution on R^2 bias in latent variable regression (for the two-step method)

A: Simple latent regression

True $R^2 = 0.5$, True Reg. Coef. = 0.707, Sample size = 300, $CD=0.6$, Scale length = 5y:5x

B: Multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\xi_1, \xi_2) = 0.5$, Scale length = 05y:05x₁-05x₂, $CD=0.6$, Sample size = 300.

Plot 9. 6: Effect of population R-square on R^2 bias in latent variable regression (for the two-step method)

A: Simple latent regression

True $R^2 = 0.5$, True Reg. Coef. = 0.707, Sample size = 300, $CD=0.6$, Scale length = 5y:5x

B: Multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\xi_1, \xi_2) = 0.5$, Scale length = 05y:05x₁-05x₂, $CD=0.6$, Sample size = 300.

Plot 9.7: Effect of correlation between two exploratory latent variables on R^2 bias in multiple latent regression (for the two-step and simultaneous IRT-SEM methods)

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, Scale length = $0.5\gamma_1 - 0.5\gamma_2$, $CD=0.6$, Symmetrically distributed items, Sample size=300.

9.4 Simulation Results for Partial Least Squares Approach

For the final research question (*How does the finite item bias of PLS compare with that of the two-step and IRT-SEM approaches?*), the results of some computer simulations under different simulation conditions are reported in this section. It should be noted that this aspect of the investigation is preliminary, being motivated by the findings related to the two-step methods. Table 9.30 shows the effect of scale length on bias in the regression estimates, for 5 item categories and a sample of size 300 in the simple latent regression case. It can be seen that the bias in the PLS approach (bold type) is very close to that in IRT and CTT two-step methods, being slightly less than that of the IRT approach, for conditions of symmetrically distributed items with $CD=0.6$ in each measurement model, and for mixed items (frequency distributions of the items are symmetric, uniform, or skewed) with CD ranging from 0.4 to 0.6. The parameter bias with PLS regression decreases as number of items increases and is appreciably larger than that of the IRT-SEM approaches. Thus, PLS parameter estimates do suffer from finite item bias, as is known from the theory. What is striking about these results is that this PLS bias is very close in magnitude to that observed for the CTT and IRT two-step methods. It can be seen from Table 9.31 that the findings for multiple latent regression are very similar to those for the simple latent regression.

As noted, this comparison of PLS with the two-step and SEM-based approaches is very preliminary. A more detailed examination of this issue under different conditions,

such as number of categories, sample size, etc., is needed before definitive conclusions can be drawn.

Table 9.30: Comparison of bias in PLS approach with the two-step and RT-SEM approaches for the simple latent regression

True R^2 = 0.5, True Reg. Coef.= 0.707, Sample size = 300, 5 categories.		
Scale length	CD=0.6 Symmetrically distributed items	CD=0.4-0.6 Mixed items: Symmetric, uniform, and skewed
05y-05x	0.858 -26.28 / -14.47 -26.44 / -14.37 6.32 / 3.11 1.67 / 0.83 -25.89 / -13.91	0.768 -36.20 / -20.47 -39.33 / -22.31 9.14 / 4.47 2.07 / 1.03 -35.96 / -19.97
10y-10x	0.924 -14.79 / -8.07 -14.67 / -7.74 4.29 / 2.12 1.99 / 0.99 -13.95 / -7.24	0.871 -21.03 / -11.34 -23.41 / -12.62 3.22 / 1.60 0.99 / 0.49 -20.34 / -10.75

Note: 1. In each cell, the first number is Cronbach's alpha of the dependent variable; the subsequent pairs of numbers represent the %bias in R^2 and %bias in regression coefficient for IRT approach (first pair), CTT approach (second pair), simultaneous IRT-SEM approach (third pair), constrained IRT-SEM approach (fourth pair), and PLS approach (fifth pair). 2. Standard error of the regression estimates ranges from 0.004-0.005 (in IRT approach), from 0.003-0.004 (in CTT approach), from 0.008-0.009 (in simultaneous IRT-SEM approach), from 0.004-0.005 (in constrained IRT-SEM approach), and from 0.003-0.004 (in PLS approach). 3. Mixed items: marginal frequency distributions of the observable items are symmetric for 2/5 of items, uniform for 1/5 of items, skewed with skewness level 1 for 1/5 of items, skewed with level 2 for 1/5 of items.

Table 9.31: Comparison of bias in PLS approach with the two-step and IRT-SEM approaches for the multiple latent regression

True $R^2 = 0.5$, True Reg. Coef. $(\gamma_1, \gamma_2) = (0.40825, 0.40825)$, True Correlation of two latent independent variables $(\xi_1, \xi_2) = 0.5$, Sample size=300, 5 categories.		
Scale length	CD=0.6 Symmetrically distributed items	CD=0.4-0.6 Mixed items: Symmetric, uniform, and skewed
05y-05x	0.882 -21.71 / -9.31 / -10.31 -21.82 / -9.53 / -9.85 2.64 / 0.26 / 2.09 0.02 / -1.29 / 0.77 -20.69 / -8.96 / -9.22	0.774 -30.42 / -14.64 / -14.38 -33.67 / -16.49 / -15.01 6.17 / 3.69 / 2.22 0.63 / 0.50 / -0.36 -29.50 / -14.30 / -14.50
10y-10x	0.938 -11.53 / -4.49 / -5.69 -11.51 / -4.55 / -5.16 0.99 / -0.67 / 1.41 -0.78 / -1.46 / 0.21 -10.27 / -3.90 / -4.60	0.875 -16.94 / -7.33 / -7.80 -19.09 / -8.57 / -8.01 1.02 / -0.62 / 1.56 -1.08 / -1.58 / 0.23 -15.88 / -6.67 / -6.36

Note: 1. In each cell, the first number is Cronbach's alpha of the dependent latent variable; the subsequent three numbers are the %bias in R^2 and %bias in regression coefficients; these three statistics are reported for IRT approach, CTT approach, simultaneous IRT-SEM approach, constrained IRT-SEM approach, and PLS approach. 2. Standard error of the regression estimates ranges from 0.003-0.004 (in IRT approach), from 0.003-0.004 (in CTT approach), from 0.008-0.009 (in simultaneous IRT-SEM approach), from 0.005-0.007 (in constrained IRT-SEM approach), and from 0.003-0.004 (in PLS approach).

9.5 Summary

This chapter has presented a series of simulation results in a search for answers to the research questions. These results can be summarized as follows. First, regardless of the measurement framework, IRT and CTT perform similarly in terms of latent variable prediction. Nevertheless, IRT is a slightly more robust method in that it can tolerate skewness and poor measurement precision of individual items better than CTT.

Secondly, the bias in R^2 attributed to using IRT- and CTT-scores in regression is very similar in general. Nevertheless, IRT-based latent score regression is a slightly more robust method than the two-step CTT regression approach, since IRT scores have better accuracy than CTT scores when items are all asymmetrically distributed, or when some items in the scale have poor measurement precision. The conditions that can lead to the lowest bias in the two-step approach are having a long scale, more item categories, a high measurement model CD , symmetric items, and a high positive correlation coefficient between latent variables in multiple latent regression. The conditions that lead to the largest bias in the two-step approach are short scales, few item categories, low values of CD , asymmetric items, and high negative correlation coefficient between the exploratory latent variables. Results on bias for the IRT-SEM approach differ markedly from those of the two-step methods. Both simultaneous and constrained IRT-SEM approaches yield smaller bias than either of the two-step approach. As a result, the overall sensitivity of both IRT-SEM approaches to changes in CD of the measurement model, scale length, and number of categories, is less than for the two-step approaches. Detailed bias trends for IRT-SEM method are displayed in

Plots 9.1-9.7. It is worth noting specially that constrained IRT-SEM method yields very small values of bias even with the smallest sample sizes studied (300). In all cases studied, simulation results for the multiple latent regression are consistent with that for simple latent regression.

Third, the R^2 bias for the PLS approach appears to be slightly lower than that in IRT approach, but again is appreciably larger than the bias exhibited by the IRT-SEM approaches. As was known from theory (Chin, 1998; Barclay, et al., 1995; Schneeweiss, 1993; Wold, 1982), PLS parameter estimates do suffer from finite item bias. What is new in these preliminary results is that the bias may be no less than that exhibited by the simple two-step methods, though further investigation is needed before definitive conclusions can be drawn. Again, similar PLS simulation results were found for both simple and multiple latent regression.

Four fitted meta-models for the two-step and IRT-SEM approaches were provided to summarize the relationships between the percentage bias of R^2 and the simulation conditions. Applications of the fitted meta-models for predicting parameter bias in latent variable regressions were illustrated.

In the next chapter, some applications of the study results to research in entrepreneurship and organizational behaviour / work-family balance will be discussed. The applications will illustrate the finite item bias attributable to using predicted scores in both two-step and PLS approaches, as compared to the bias in the SEM-based analysis.

Chapter 10

Applications to Two Business Research Topics: Entrepreneurship and Work-Family Conflict

In this chapter, we will examine the application of the study results to two research topics in business. The first, featuring binary data, relates to a study of managerial competence carried out by Orser, Madill, & Riding (2002). An objective of their study was to create an index of Management Competence for use in research, in particular, the study of business owners' gender difference in management competence. Managerial competence is an important issue in general, since entrepreneurship theory argues that business owners' management competencies have an impact on firm performance. A good understanding of the influences of management competencies can assist public policy makers in developing entrepreneurial and small business training or other public policy initiatives (Orser & Riding, 2003). Newton (2001) argues that management competencies are a central component of a firm's innovation strategy which is the key to business survival and growth. He also stresses that since SMEs account for large proportions of economic activity and are an important source of innovation, small business management skills should be a primary focus for economic policy in general and for innovation strategies in particular. Furthermore, banks and other lending

organizations have adopted managerial competences diagnostics to evaluate new business proposals (Martin & Staines, 1994). The application studied here will investigate how the different methodologies examined in this thesis perform for a simple regression featuring two sub-scales drawn from the index developed by Orser, et al. (2002). In a broader sense, it will explore use of binary scales in research in an important application area, where competency scales are typically binary in nature.

The second application, featuring polytomous data, relates to a series of studies on work and family conflict, conducted by Duxbury & Higgins (2001, 2002, 2003). An objective of their studies was to provide business and labour leaders, policy makers and academics with an overall view of the current dynamics between work and family domains. Work-family balance is an important issue for the following reasons: (a) from the employer's perspective, work-family conflict has been linked to reduced work performance, lower productivity, increased absenteeism, lower commitment, higher turnover, and poor morale, (b) from the perspectives of employees and their families, work-family conflict are related to marital problems, lower family and life satisfaction, increased perceived stress, burnout, depression, and stress-related illness, (c) society will benefit if people achieve balance and are able to devote more time and effort to their roles of neighbour and volunteer (Duxbury & Higgins, 2003).

In the following, empirical data from the entrepreneurship and work-family conflict studies will be examined. These two case studies examined in the chapter will allow for a demonstration of the usefulness of the previous meta-models, which will enable us to assess the accuracy of the parameter estimates obtained using the various methodologies. Such applications of the meta-models are potentially very useful to

primary researchers, who can use them to assess the likely influence of their methodology on consequent parameter estimates. Appropriate meta-models can also be used *prospectively* to help decide on the best methodology for performing specific analyses. On the other hand, where the relevant information is available, meta-models can be used *retrospectively* (as in the present investigation), to gauge the level of bias in previous research results. For the entrepreneurial case study, with binary data, CTT and IRT-SEM approaches will be illustrated, while for the work-family study, with polytomous data, CTT, PLS, and IRT-SEM approaches will be illustrated.

10.1 The Entrepreneurship Case Study (Binary Indicators)

As described, these data were collected as part of the Managerial Capacity Index study reported in Orser, et al. (2002), and made available for this exploration courtesy of Dr. Orser. A questionnaire was administered to business owners in the provinces of Ontario and Manitoba, and 214 cases were available for analysis after deletion of cases containing missing data on the variables of interest. Two sub-scales were selected, namely business owners' "prior financial management experience" and "prior capital formation experience". The former was treated as a dependent variable to be predicted by the latter, i.e., a simple latent regression simulation. Both scales contain 5 binary items with Cronbach's alpha of 0.80 and 0.73, for prior financial management experience and prior capital formation experience, respectively. The data reveal that 6 of the 10 indicators are asymmetrically distributed and that the average standardized loading for the 10 indicators is 0.793. Hence, the average *CD* for the measurement models is 0.629.

From Table 10.1, it can be seen that the values of both $\hat{R}^2$ and regression coefficient, 0.392 and 0.626, respectively, corresponding to the CTT-based score regression are much lower than the corresponding estimates produced by the simultaneous IRT-SEM approach, namely 0.637 and 0.820. We can estimate the true R^2 in this latent regression by making a bias correction based on the fitted meta-model shown in Table 9.26. It turns out that the predicted downward bias is 39% for the CTT-based simple latent regression with 5 binary mixed items, a *CD* of 0.629, and an

assumed true R^2 of 0.5 (refer to Section 9.3.2 for an illustration of the use of the meta-models). Thus, the estimate of the bias corrected R^2 in this example is 0.643. This value is obtained using the definition of bias in Equation (9.1), entering the value of $\hat{R}^2 = 0.392$, and bias = -0.39, and then solving for the bias corrected value of R^2 .

Table 10.1: Entrepreneurship case study with binary indicators in the simple latent regression

Sample Size	CTT Regression Estimation		Simultaneous IRT-SEM Regression Estimation	
	R^2	Reg. Coef.	R^2	Reg. Coef.
214	0.392	0.626	0.637	0.820

Note: Bias-corrected R^2 based on the meta-model is 0.643.

In this case, even the attenuated $\hat{R}^2$ (or, equivalently, the attenuated regression coefficient) is statistically significant at standard α -levels. However, statistical significance is not necessarily the primary criterion of interest in a policy study. Rather, it is the strength of the relationship that is (or should be) of primary concern. In this case, the difference between an R^2 (i.e., a measure of predictive strength) of nearly 65% and an attenuated CTT value of 39% is of considerable substantive importance. In other more exploratory studies, where the existence of an effect would have important research significance, a negative attenuation of nearly 40% in the regression R^2 could turn what should be a significant finding into a non-significant one. For example, an attenuated regression coefficient of 0.13 for a sample of size 214 could just fail to make significance at the 5% level. However, the attenuation corrected coefficient of 0.17

would have a p-value of approximately 0.016, clearly significant (corresponding to an R^2 attenuation of 39%).

It is clear from Table 10.1 that simultaneous IRT-SEM estimation does a far better job of estimating the strength of the relationship than does the two-step method. It yields an $\hat{R}^2$ of 0.637 compared to the attenuation corrected value of 0.643.

10.2 The Work-Family Conflict Case Study (Polytomous Indicators)

The second dataset, made available courtesy of Drs. Duxbury and Higgins, consists of a subset of responses from a study of Canadian employees who work for public, private and non-profit sector organizations with 500 or more employees. A complete account of the findings of the study has been published in *The 2001 National Work-Life Conflict Study* (Duxbury & Higgins, 2003). Two sub-samples of size 500 and 2000, free of missing data, were randomly selected from the dataset for the purposes of this application. In their study, one of the topics of interest was to examine the impact of “role overload” on “burnout”, both of which were measured using 5-point Likert scales containing 5 items each, with Cronbach’s alpha of 0.88 and 0.91 respectively. In their report, Duxbury & Higgins (2003) explored the relationships between these scales by first recoding the CTT mean score into binary measures (high/low) and then using an odds ratio as an association measure. In this application, simple linear regression will be used to study the pairwise relationship. In the sub-sample with 2000 cases, all 10 indicators were roughly symmetrically distributed with an average standardized loading of 0.827 (i.e., an average *CD* of 0.684).

As in the previous example, Table 10.2 shows that the estimated R^2 and the corresponding regression coefficient in the CTT regression are appreciably lower than the simultaneous IRT-SEM estimates, but very close to those obtained using PLS approach. The predicted downward bias in $\hat{R}^2$ is estimated to be 24.45% in the CTT latent regression, based on the fitted meta-model with 5 symmetric items containing 5

categories, a CD of 0.684, and an assumed true R^2 of 0.5, bias corrected R^2 in this application then being 0.531, quite close to the IRT-SEM estimate of 0.552, based on a meta-model for the sub-sample of size 500. Without taking this bias into account, a researcher would thus be misled by the attenuated relationship. The results in Table 10.2 again show that IRT-SEM methods provide worthwhile improvements in estimation accuracy, i.e., the corresponding regression gives a much more accurate description of the actual relationship in the population.

Table 10.2: Work-family conflict case study with polytomous indicators in the simple latent regression

Sample Size	PLS Estimation		CTT Regression Estimation		Simultaneous IRT-SEM Regression Estimation	
	R^2	Reg. Coef.	R^2	Reg. Coef.	R^2	Reg. Coef.
500	0.427	0.635	0.401	0.633	0.556	0.746
2000	--	--	0.390	0.625	0.549	0.741

Note: 1. "--" denotes the PLS processing failed due to unknown reason. 2. Bias-corrected R^2 based on the meta-model for the sub-sample of size 500 is 0.531.

The investigation of the Duxbury and Higgins (2003) dataset was extended by adding "work-to-family conflict" as another exploratory variable to examine its joint impact with "role overload" on "burnout". "Work-to-family conflict" was measured using a 5-point Likert scale containing 6 items, and has a Cronbach's alpha over 0.90. Inspection of the sub-sample of 2000 showed that all the indicators of "work-to-family conflict" were approximately symmetrically distributed. The correlation coefficient

between the two explanatory latent variables was found to be 0.451, based on confirmatory factor analysis, and the average standardized loading for all 16 indicators of the dependent and explanatory latent variables was 0.827 (i.e., an average *CD* of 0.684).

Table 10.3 shows that $\hat{R}^2$ in the CTT regression is appreciably lower than for the simultaneous IRT-SEM approach, but again quite close to that in PLS approach. We can again estimate that bias corrected R^2 in this multiple latent regression, which turns out to be 0.546 (close to the IRT-SEM value of 0.574), after making a correction for a predicted downward bias of 18.51%. Thus, the IRT-SEM regression method again generates better R^2 estimates than does either the CTT regression method or the PLS approach.

Table 10.3: Work-family conflict case study with polytomous indicators in the multiple latent regression

Sample Size	PLS Estimation	CTT Regression Estimation	IRT-SEM Regression Estimation
	R^2	R^2	R^2
500	0.470	0.445	0.574
2000	--	0.431	0.563

Note: 1. "--" denotes the PLS processing failed due to unknown reason. 2. Bias-corrected R^2 based on the meta-model for the sub-sample of size 500 is 0.546.

Data from another study of work-life balance by Duxbury & Higgins (1998) has been investigated in detail by Power (2004), who examined the relationship between care-giving and role interference, and a number of other moderating and potentially

confounding variables. One of the objectives of her study was to assess the overall performance of a “family interferes with work” regression model containing 9 observed explanatory variables and 4 explanatory latent variables. All the latent variables, including the dependent variable, were measured using 5-point Likert scales containing from 4 to 14 items. The CTT mean scores for all 5 latent variables were used in the regression analysis, i.e., her analysis comprised a classic “two-step” CTT approach. Power’s (2004) analysis was successful in that most of her prior expectations were supported by her data, only one of the four latent variables in the “family interference with work” model being non-significant. The reported $\hat{R}^2$ was 0.201. Nevertheless, it is reasonable to suppose based on the results of this thesis, that her overall $\hat{R}^2$ value may be lower than it should be, and that some of the regression coefficients pertaining to the latent variables may be somewhat biased. Unfortunately, the size of the model exceeds the scope of the meta-models discussed earlier, so no indirect estimates of the biases can be provided.

Power’s (2004) work is noteworthy that she avoided the trap of only looking at the significance levels of the regression coefficients, but instead went on to assess the relative importance of different variables. She based her importance assessment on measures of the Pratt index, as described by Thomas, Hughes, & Zumbo (1998). These indices depend on both the size of individual standardized regression coefficient and on the correlation between each predictor variable and the dependent variable. Thus the biases described in this thesis, arising from the use of latent variable scores, can affect these importance indices in two ways, namely through the bias in the regression coefficients and through the bias in the dependent-predictor correlation. Further,

Power's regression contained 4 latent predictors, so the biases could in fact vary in direction. Thus it would be worthwhile to rerun her study using the discrete IRT-SEM approach to determine whether or not the measurement error in the latent variable scores affected the relative importance ordering of the variables. This has not been done, but it is an interesting possibility that has been flagged for future study.

10.3 Summary

The application of the study results to data from two separate research studies (entrepreneurship and work-family conflict) was examined. These applications illustrated and confirmed the bias arising from the two-step CTT regression approach and the PLS approach. In the entrepreneurship case, featuring binary data, the R^2 and regression parameter estimates in the CTT regression were much lower than for IRT-SEM approach. For the polytomous example, the R^2 and regression parameter estimates in the CTT regression were very close to that obtained using the PLS approach, but both methods led to appreciably lower parameter estimates than the IRT-SEM approach. When bias corrections based on the CTT meta-models were applied to the CTT results, the bias adjusted results were much closer to the values predicted by the IRT-SEM methodology. It is clear that different approaches using the same data will generate contrasting estimates of the structural latent variable relationships. This raises the spectre of different researchers drawing different conclusions from the same data due solely to the type of the latent variable structural approach used to test the relationships.

Chapter 11

Concluding Comments

11.1 Review and Major Conclusions

Classical test theory (CTT) has many well-documented drawbacks. The most telling criticisms are item- or scale-dependent scores, respondent-dependent scale characteristics, and constant standard error of measurement across entire latent variable levels, which ultimately limit the utility of the respondent and item statistics in practical scale construction and application. Nevertheless, CTT continues to be the dominant paradigm for measurement research in business. *Item response theory* (IRT) was designed to address the various perceived shortcomings of CTT and offers some of the best alternatives for devising and optimizing scales and tests, and for performing item analysis. Also, IRT models are applicable to both dichotomous and polytomous scored data, which are commonly seen in business research. However, it is surprising to note that IRT has been slow to diffuse into the business field. Most of the existing IRT applications in business measurement appeared over the last five years, and are still relatively scarce. IRT models have recently been recommended for use in business by McCollam (1998) and Singh (2004).

Given this trend, it is natural to enquire whether or not IRT methods yield substantially better measures than does CTT when the goal of measurement is to obtain latent variable scores for individuals. That is, how comparable and how accurate are the CTT- and IRT-based scores (Research Question 1)? From a theoretical point of view, the comparison of IRT and CTT scores poses some difficult mathematical challenges, with the result that the comparative information is fragmentary. Our Monte Carlo simulation has filled the gaps in the coverage of the various design characteristics of scales in the literature. The simulation results show that IRT and CTT perform similarly in terms of latent variable prediction, which is consistent with the empirical literature. Nevertheless, we found that IRT is a slightly more robust method in that it can tolerate skewness and poor measurement precision of individual items better than CTT.

The study of the relationships among latent variables is also very common in business research, where many researchers use a *two-step approach*, i.e., obtain latent variable predicted scores and then use these scores directly in subsequent analyses such as regression. While this approach is very common in business, it has serious drawbacks, regardless of the individual optimality of the measurement technique used to derive the scores. The major drawback is that analyses based on the predicted scores are subject to *finite item bias*, regardless of sample size. As a result, in a simple regression analysis, the consequence of ignoring measurement error (i.e., directly using an observed single-indicator measure or composite score from a multiple scale and treating it as true latent variable value in analysis) is the attenuation of parameter estimates; but in multiple regression or multi-equation systems, ignoring measurement error can lead both to attenuated and inflated parameter estimates. Since it is difficult to predict the

consequence of measurement error, the need to take into account the measurement error in modeling becomes extremely important. Nevertheless, given some of the potential advantages of IRT over CTT, one might expect there to be an appreciable reduction in finite item bias from using IRT-based versus CTT-based scores. Surprisingly, little attention has been focused on this issue in the psychometric or other literature, despite the long history of IRT.

Among the methods for bias reduction, *structural equation modeling* (SEM) is the most appealing one to us, due to its applicability to discrete indicator variables and its statistical property of consistent parameter estimates, irrespective of the number of items defining an individual construct. It is not well known that, in fact, in the case of discrete indicators, the IRT model is embedded in the SEM model. In other words, parameter estimation in the discrete SEM is equivalent to simultaneous estimation of IRT item parameters and the latent regression parameters, referred to as *simultaneous IRT-SEM* estimation. When IRT parameters are known a priori and fixed in the IRT-SEM estimation, we call this *constrained IRT-SEM* approach, which again yields consistent estimates. Another technique known as *partial least squares* (PLS) has been used by business researchers to estimate the parameters of a structural model. PLS parameter estimates, however, are not consistent as the PLS technique generates latent variable scores as an intermediate step and so suffers from finite item bias unless the number of items is very large.

Given the above alternatives to latent variable modeling, i.e., CTT-and IRT-based two-step approaches, simultaneous and constrained IRT-SEM approaches, and PLS, are there appreciable differences in the biases of their structural regression coefficient

estimates or the corresponding R^2 for discrete indicators (covering Research Questions 2 & 3)? To the author's knowledge, this line of study has not been seen in the literature. A Monte Carlo study was carried out to assess the properties of the R^2 and regression coefficients in a latent structural equation generated by various methods, when the indicators are discrete.

The simulation results showed that the biases in R^2 and regression coefficients attributed to using IRT- and CTT-scores in regression are very similar in general. Nevertheless, IRT-based latent score regression is a slightly more robust method than the two-step CTT regression approach, since IRT scores have better accuracy than CTT scores when items are all asymmetrically distributed, or when some items in the scale have poor measurement precision. The two-step approach is sensitive to scale length, number of categories, measurement precision, and symmetry of item distribution.

Results on bias for the IRT-SEM approach differ markedly from those of the two-step methods. Both simultaneous and constrained IRT-SEM approaches yield smaller bias than either of the two-step approaches. As a result, the overall sensitivity of both IRT-SEM approaches to changes in scale length, number of categories, measurement precision, and symmetry of item distribution, is less than for the two-step approaches. It is worth noting especially that constrained IRT-SEM method yields very small values of bias even with the smallest sample sizes studied (300). In all cases studied, simulation results for the multiple latent regression are consistent with those for simple latent regression.

Generally speaking, the biases in R^2 and regression coefficients for the PLS approach appear to be slightly lower than that in IRT approach, but again are appreciably larger than the biases exhibited by the IRT-SEM approaches. As was known from theory (Chin, 1998; Barclay, et al., 1995; Schneeweiss, 1993; Wold, 1982), PLS parameter estimates do suffer from finite item bias. What is new in these preliminary results is that the bias may be no less than that exhibited by the simple two-step methods, though further investigation is needed before definitive conclusions can be drawn. Again, similar PLS simulation results were found for both simple and multiple latent regression.

The application of the study results to data from two separate research studies (entrepreneurship and work-family conflict) was examined. These applications illustrated and confirmed the bias arising from the two-step CTT regression approach and the PLS approach. It is clear that different approaches using the same data will generate contrasting estimates of the structural latent variable relationships. This raises the spectre of different researchers drawing different conclusions from the same data due solely to the type of the latent variable structural approach used to test the relationships. This variation in quantitative estimates between different methodologies was recently noted in a presentation to Statistics Canada's Advisory Committee on Statistical Methods by Cyr & Cartwright (2004), who commented that

“ ... comparability issues may arise across analyses that purport to use the same scale data but produce different results. For high-stakes policy analysis, this becomes a troublesome issue that casts doubt upon any analyses conducted using these data.”

11.2 Recommendations

Recommendations can be summarized in the following catchphrases:

First, *the conventional two-step approach should be used with caution*. It should be used when there is a long scale, many item categories, high measurement precision, and when the items are symmetric. Nevertheless, for simple regression models, the two-step approach might be useful in preliminary or exploratory analyses if we keep in mind that the estimates are attenuated (i.e., conservative). However, for multiple regression or multi-equation models, the two-step approach can lead both to attenuated and inflated parameter estimates.

Second, *do not choose the easy way out; consider IRT-SEM instead*. The potential for misleading results is great if measurement error is ignored. This is especially of concern in business studies, where on average, measurement error accounts for about 44% of the observed variance in a latent variable score (Cote & Buckley, 1987). Thus, the IRT-SEM approach should be considered as an alternative to the two-step approach given a large enough sample, due to its ability to take measurement error into account.

Third, *for a small sample size, use constrained IRT-SEM*. Although using IRT scores for subsequent analyses is problematic, IRT still has its merits. IRT has proven to be a solid technique for scale construction which provides a better evaluation of measurement precision. Thus, for *scale developers*, it is recommended that IRT be used for scale construction based on large normative samples and the results, in particular the item parameters, be made available for use in subsequent research. Then, *scale users*

can translate the IRT parameters into discrete SEM parameters and treat them as fixed in the IRT-SEM estimation. That is, using previously calibrated items will offset the effect of small sample size to a considerable extent.

It is recognized that the third recommendation might be difficult to implement for both scale developers and scale users since IRT is a model-based method that entails a mathematically complex theory. From a practical point of view, IRT requires diagnostic tests of model assumptions; however, the key assumption of unidimensionality in IRT is also shared by CTT, a fact which is sometimes ignored by researchers. Therefore, the final recommendation of this thesis is that greater attention should be paid to the teaching of both IRT and CTT measurement theories and practices in business schools.

11.3 Directions for Future Research

As noted, IRT-SEM modeling is designed for discrete item indicators. Despite the availability of the IRT-SEM approach, the standard SEM method for continuous item indicators is more commonly used in studies involving discrete item indicators, particularly when the items have five or more response categories. It will be important to extend this study to explore the difference in bias between using the discrete SEM approach and the continuous SEM approach for analyzing indicator variables which are discrete (binary or ordinal), so that we may assess the bias due to ignoring the categorization. The effects of categorization error have been studied to some extent, as described by Bollen (1989, p.p. 433-447), but this research is somewhat dated and needs to be integrated with the results of the current study.

In addition, research is needed to quantify the bias in PLS estimates for polytomous data under various conditions, such as number of categories, number of items, measurement precision, sample size, symmetry of indicator distribution, etc, and in particular to compare PLS to the IRT-SEM as a method for latent variable modeling when samples are small.

The study has shown that IRT (EAP) and CTT scores are highly comparable in general, and the advantages of IRT do not carry over when IRT scores are used in a two-step analysis. It is not known to what degree these results are specific to the type of scoring method used. A similar simulation study should be carried out by employing different IRT score prediction methods, such as the weighted likelihood estimator

(WLE) which is now used to produce scores in some large international educational surveys such as the OECD's Program for International student Assessment (PISA) (Adams & Wu, 2000).

The performance criterion used to compare the various latent variable approaches in the simulation study was the relative bias of structural regression coefficient estimates, or the corresponding R^2 . The mean square error of the parameter estimates can be further considered as an evaluation criterion, since it combines the variance and bias of the parameter estimates (mean squared error = variance + square of the bias). This would be particularly useful when the study is extended to investigate the interface between discrete and continuous indicators, as West, Finch, & Curran (1995) reports some results suggesting that categorization error also affects the standard errors of parameter estimates.

Finally, the current study should be extended to explore a greater range of simulation conditions, for example, sample sizes below 300, and also more variation in the measurement model parameters of the simulated latent variables. Further, conditions leading to both attenuation and inflation of parameter estimates should be studied. These developments would require further development of the meta-models, and might necessitate the inclusion of non-linear terms to capture trends due to small sample sizes and other variations in simulation settings.

11.4 Limitations

The IRT-SEM model assumes that the underlying latent variable is unidimensional and normally distributed. However, scales used in the social sciences have not been adequately demonstrated to be uni-dimensional (Shevlin, et al., 1997). For example, the Cognitive Dissonance After Purchase scale developed by Sweeney, et al., (2000) and the Interactivity of Websites scale constructed by Liu (2003) are multidimensional. Of course, careful evaluation of scales using discrete exploratory and confirmatory factor analysis can identify such difficulties, and the results of this study can then be applied to appropriate uni-dimensional sub-scales. A more fundamental limitation is that some latent variables in business (and other behavioural disciplines) may not in fact have normal underlying distributions.

Shinn (1974) reviewed the application of psychophysical scaling to attitude measurement and suggested that perceptions are not normally but rather log-normally distributed (O'Brien, 1981). Abandoning the assumption of underlying normality has major theoretical and practical consequences and is an important area for future research in latent variable theory, which will require the concerted efforts of many researchers. It is interesting to note that such an effort may soon be underway. The Statistical and Applied Mathematical Sciences Institute (SAMSI), an American network of major research institutions, has identified Latent (Hidden) Variable Models in the Social Sciences as one of its theme areas for 2004-05, and the Statistical Society of Canada is currently negotiating an involvement in this project. Thus it may be possible for the author and her collaborators to play a small role in such future developments.

Appendix A

A Proof That $E(e) = 0$ and $\rho(t, e) = 0$

As noted, the process of measuring the characteristic of a group of subjects (i.e., respondents) can be modeled as follows: (a) the subjects are chosen at random from a population and (b) the observed score for each subject is a random value from this subject's own distribution. Thus, there are two sources of the variation in the observed score x : one is due to the differences among the subjects, and the other is due to the differences between all the possible observed scores, the x_p 's, specific to this particular subject (i.e., variation in random error pertaining to this subject). In other words, an observed score x for a subject is drawn from the joint distribution composed of the population distribution of subjects and the observed score distribution of this particular subject.

Then, there are two distinct expectations: one is $E_p(.)$ representing the expectation over a particular subject's random error distribution; the other is $E_s(.)$ denoting the expectation over the population of subjects. From statistical probability theory, $E(.)$,

namely expectation over the joint distribution is equivalent to: $E \equiv E_s E_p$. Then, the results $E(e) = 0$ and $\rho(t, e) = 0$ can be proved as follows.

$$(1) E(e) = 0$$

Proof: $E(e) = E_s(E_p(e_p)) = 0$, since $E_p(e_p) = 0$ by definition.

$$(2) \rho(t, e) = 0$$

Proof:

$$\begin{aligned} \text{cov}(t, e) &= E(te) - E(t)E(e) = E(te), \text{ since } E(e) = 0 \\ &= E_s[E_p[t_p(x_p - t_p)]] = E_s[t_p E_p(x_p) - t_p^2], \\ &\quad \text{since } E_p(t_p) = t_p \text{ and } E_p(x_p) = t_p \\ &= E_s[t_p^2 - t_p^2] = 0 \end{aligned}$$

$$\text{Thus, } \rho(t, e) = \frac{\text{cov}(t, e)}{\sqrt{\text{var}(t) \text{var}(e)}} = 0.$$

Appendix B

The Derivation of the Formula for Accuracy

This appendix will spell out the derivation of the accuracy index for the binary case with arbitrary probability of correct response. The accuracy index was measured as the shared variance of the CTT standardized total score (Z) with the latent variable as defined in this thesis, i.e., $\rho^2(Z, \eta)$. The standardized total score is defined as

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)}}, \text{ where } X \text{ is the observed total score for each individual respondent.}$$

As introduced in Section 2.2, the one-factor model for a binary item, i , can be expressed as:

$$x_i^* = \lambda\eta + \varepsilon_i, \quad i = 1, \dots, n$$

and the observable discrete indicator variable is then modeled as

$$\begin{aligned} x_p &= 1 \text{ if } x^* \geq \tau, \\ x_p &= 0 \text{ otherwise,} \end{aligned}$$

where η is the latent variable, ε_i is an uncorrelated respondent-specific error with variance θ_i^2 , and x_i^* is a continuous indicator of η . We assume the variances of η and x_i^* are equal to one, i.e., $\text{var}(\eta) = \text{var}(x_i^*) = 1$, and λ 's are all the same for all items.

This represents a simplified case, but is sufficient for expository purpose. The final results can be generalized. The subscripts for each subject (respondent) are ignored for simplicity, and the latent variable η for each subject is assumed to be independent from each other. Therefore, the probability of a positive answer is as follows:

$$P_i = P(x_i = 1) = P(x_i^* \geq \tau_i) = \int_{\tau_i}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^*)^2}{2}} dx_i^* = 1 - \Phi(\tau_i). \quad (\text{B.1})$$

where Φ is the cumulative standard normal density.

Now, we are ready to derive the formula for the accuracy index as follows.

$$\rho^2(Z, \eta) = \frac{[\text{cov}(Z, \eta)]^2}{\text{Var}(Z)\text{Var}(\eta)} = \frac{[\text{cov}(Z, \eta)]^2}{\text{Var}(Z)}, \text{ since } \text{var}(\eta) = 1. \quad (\text{B.2})$$

The square root of the numerator of Equation (B.2) can be worked out as:

$$\begin{aligned} \text{cov}(Z, \eta) &= \text{cov}\left(\frac{X - \bar{X}}{\sqrt{\text{var}(X)}}, \eta\right) = \frac{1}{\sqrt{\text{var}(X)}} \text{cov}\left(\sum_{i=1}^n x_i, \eta\right), \text{ where } X = \sum_{i=1}^n x_i. \\ &= \frac{1}{\sqrt{\text{var}(x)}} \sum_{i=1}^n \text{cov}(x_i, \eta) = \frac{n}{\sqrt{\text{var}(x)}} \text{cov}(x_i, \eta), \end{aligned}$$

assuming n items with identical characteristics.

Now, $\text{cov}(x_i, \eta) = E(x_i \eta) - E(x_i)E(\eta) = E(x_i \eta)$, since it is assumed that $E(\eta) = 0$ (a standard assumption). Thus,

$$\begin{aligned} \text{cov}(x_i, \eta) &= \int \eta x_i(x_i^*) dF(\eta, x_i^*), \text{ with } x_i^* \geq \tau_i, \\ &= \int \eta dF(\eta, \varepsilon), \text{ over the region where } x_i^* \geq \tau_i \text{ or } \eta \geq \frac{\tau_i - \varepsilon_i}{\lambda}. \end{aligned}$$

This leads to

$$\text{cov}(x_i, \eta) = \frac{1}{2\pi\sqrt{1-\lambda^2}} \int_{-\infty}^{+\infty} \int_{\frac{\tau_i - \varepsilon_i}{\lambda}}^{+\infty} \eta e^{-\frac{\eta^2}{2}} e^{-\frac{\varepsilon^2}{2(1-\lambda^2)}} d\eta d\varepsilon$$

which can, after some manipulation, be reduced to

$$\text{cov}(x_i, \eta) = \lambda \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau_i^2}{2}} = \lambda \phi(\tau_i), \text{ where } \phi(\tau_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau_i^2}{2}}.$$

Thus, the numerator of Equation (B.2) becomes

$$[\text{cov}(Z, \eta)]^2 = \frac{n^2}{\text{var}(X)} \lambda^2 \phi^2(\tau_i), \text{ where } \tau_i = \Phi^{-1}(1 - P_i). \quad (\text{B.3})$$

The square root of the denominator of Equation (B.2) can be worked out as:

$$\begin{aligned} \text{var}(Z) &= \text{var}\left(\frac{X - \bar{X}}{\sqrt{\text{var}(X)}}\right) = \frac{1}{\text{var}(X)} \text{var}\left(\sum_{i=1}^n x_i\right) \\ &= \frac{1}{\text{var}(X)} \left[n \text{var}(x_i) + \frac{n(n-1)}{2} \rho \text{var}(x_i) \right], \text{ assuming } \rho(x_i, x_j) = \rho \\ &= \frac{1}{\text{var}(X)} \left[n P(1-P) + \frac{n(n-1)}{2} \rho P(1-P) \right], \text{ assuming all } P_i = P. \\ &= \frac{n}{\text{var}(X)} [P(1-P)(1 + (n-1)\rho)] \end{aligned} \quad (\text{B.4})$$

Employing Equations (B.3) and (B.4) for Equation (B.2), then the accuracy index can be written as:

$$\rho^2(Z, \eta) = \frac{[\text{cov}(Z, \eta)]^2}{\text{Var}(Z)\text{Var}(\eta)} = \frac{n\lambda^2\phi^2(\tau)}{P(1-P)(1 + (n-1)\rho)}, \quad (\text{B.5})$$

where $\tau_i = \Phi^{-1}(1 - P_i)$ and $\phi(\tau_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau_i^2}{2}}$.

Based on the formula of Equation (B.5), we can then calculate the accuracy index for the CTT standardized total score by hand. Table B.1 shows the accuracy measures for CTT based on both simulation results and the above formula for symmetric and asymmetric binary cases. It can be seen that both quantities are very close; this indicates that our simulation estimates of accuracy in Table 9.4 are correct. It is also worth noting that in Table B.1, the values of Cronbach's alpha are always greater than the accuracy measures, especially for the case of 5 asymmetric items, where Cronbach's alpha is

twice as large as the accuracy. Thus, Cronbach's alpha is not a lower bound for the accuracy; on the contrary, it is an optimistic estimate of the accuracy.

Table B.1: Cronbach's alpha vs. accuracy for binary indicators

Assumptions: (1) $P(x_i = 1) = P$; (2) $\tau_i = \tau$; (3) $\lambda_i = \lambda = \sqrt{0.6}$, i.e., $CD=0.6$ for all items; (4) $\rho(x_i, x_j) = \rho = 0.2683$				
Number of items: (n)	Marginal frequency distribution: $(P, 1-P)$	Cronbach's alpha $\rho^2(T(\eta), \eta)$	Calculated Accuracy: $\rho^2(Z, \eta)$	Simulated Accuracy: $\rho^2(Z, \eta)$
5	0.5 , 0.5	0.774	0.726	0.725
10	0.5 , 0.5	0.873	0.818	0.818
5	0.95 , 0.05	0.637	0.324	0.320
10	0.95 , 0.05	0.780	0.393	0.389

Appendix C

A Mathematical Explanation of the Relationship between the Correlation Coefficient of Exploratory Latent Variables and the Magnitude of the Parameter Bias

The purpose of this appendix is to prove a mathematical explanation of the relationship between the correlation coefficient of the two independent latent variables and the magnitude of the parameter bias shown in Table 9.19 and Plot 9.7 for the multiple latent regression model.

We follow the discussion of the consequences of measurement errors for multiple regression by Bollen (1989, pp.159-161). Consider the true multiple latent regression model given by

$$\eta = \Gamma\xi + \zeta, \quad (\text{C.1})$$

where η is a single latent response variable, ξ is a $m \times 1$ vector of latent explanatory variables, Γ is a $1 \times m$ matrix of regression coefficients and ζ represents a disturbance term independent of ξ with $E(\zeta) = 0$. The estimated model with observed variables is

$$y = \Gamma^* \mathbf{x} + \zeta^* . \quad (\text{C.2})$$

The relations of $\mathbf{x}$ to ξ , and y to η are

$$\mathbf{x} = \xi + \delta \quad \text{and} \quad y = \eta + \varepsilon , \quad (\text{C.3})$$

where $\mathbf{x}$, ξ , and δ are $m \times 1$ matrices, and y , η , and ε are scalars. The measurement errors (δ, ε) are uncorrelated with ξ , η , ζ , and with each other. Only one observed variable x is assumed for each latent variable ξ and only one observed variable y is used for η . For simplicity, the observed variables are assumed to have the same scale as the latent variables (i.e., $\Lambda_x = \mathbf{I}$ and $\Lambda_y = \mathbf{I}$). Bollen (1989) showed the relationship of Γ to Γ^* is

$$\Gamma^{*'} = \Sigma_{xx}^{-1} \Sigma_{xy'} = \Sigma_{xx}^{-1} \Sigma_{x\xi'} \Gamma' \quad (\text{C.4})$$

Equation (C.4) suggests if we ignore the error in y and $\mathbf{x}$, the Γ^{*} we end up with will differ from the true Γ' by the factor $\Sigma_{xx}^{-1} \Sigma_{x\xi'}$.

In our simulation, the multiple regression model for Equation (C.1) was designed as

$$\eta = (\gamma_1 \ \gamma_2) \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \zeta , \quad (\text{C.5})$$

with the covariance matrix of ξ , Φ , i.e.,

$$\Phi = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \quad (\text{C.6})$$

The covariance matrix of $\mathbf{x}$ is

$$\Sigma_{\mathbf{x}\mathbf{x}'} = \Phi + \text{cov}(\delta) = \Phi + \theta_{\delta}. \quad (\text{C.7})$$

The covariance matrix of $\mathbf{x}$ and ξ then is

$$\Sigma_{\mathbf{x}\xi'} = E(\xi\xi' + \delta\xi') = \Phi. \quad (\text{C.8})$$

Employing Equations (C.7) and (C.8) for the formula of Γ^{*} , the Equation (C.4) then becomes

$$\Gamma^{*'} = (\Phi + \theta_{\delta})^{-1} \Phi \Gamma'. \quad (\text{C.9})$$

In our simulation, the structural regression coefficients are assumed to be same, i.e.,

$\gamma_1 = \gamma_2 = \gamma$. Then, Equation (C.9) can be simplified as

$$\gamma^* = \frac{1 + \theta - \rho^2 + \rho\theta}{(1 + \theta)^2 - \rho^2} \gamma = A\gamma, \quad (\text{C.10})$$

where γ^* is one of the two identical elements in Γ^* . θ_1 and θ_2 are then the diagonal elements of θ_{δ} and it is assumed $\theta_1 = \theta_2 = \theta$ in our simulation. It is easily proved that $A \leq 1$. It can be seen from Equation (C.10) that A is smaller if ρ is negative compared to that with a positive ρ . In other words, the shrinkage of γ^* is greater when $-1 \leq \rho \leq 0$. Therefore, the parameter bias is greater when the correlation coefficient between the two exploratory variables is negative. The above explanation assumes continuously measured items but nevertheless serves as an indication of the expected trend when only discrete variables are used.

It is also possible to show, via Equation (C.9) that the bias due to ignoring measurement error can lead to inflation of parameter estimates in certain cases. To prove this, consider the matrix operator norms defined by Noble & Daniel (1977, p.p. 164-166). In particular, consider the norm $\|\mathbf{A}\|_1$, for a matrix $\mathbf{A}$, defined by

$$\|\mathbf{A}\|_1 = \max_{\mathbf{x} \neq \mathbf{0}} \left\{ \frac{\|\mathbf{Ax}\|_1}{\|\mathbf{x}\|_1} \right\}$$

where the vector norm $\|\mathbf{v}\|_1 = \sum_i |v_i|$, for any vector $\mathbf{v}$ with elements v_i . From Theorem 5.7 of Noble and Daniel (1977),

$$\|\mathbf{A}\|_1 = \max_j \sum_{i=1}^m |a_{ij}|,$$

namely, the maximum absolute column sum. Let $\mathbf{A} = (\boldsymbol{\Phi} + \boldsymbol{\theta}_\delta)^{-1} \boldsymbol{\Phi}$ from Equation (C.9). Then for the case of two exogenous latent variables, it can be shown that

$$\|(\boldsymbol{\Phi} + \boldsymbol{\theta}_\delta)^{-1} \boldsymbol{\Phi}\|_1 \leq 1 \text{ when } \rho \leq \theta_1 + \frac{\theta_1}{\theta_2}, \text{ where } \theta_1 \text{ and } \theta_2 \text{ are the diagonal elements of}$$

$\boldsymbol{\theta}_\delta$, and $\theta_2 \geq \theta_1$. The norm is thus less than one whenever $\rho \leq 0$. It is also less than one, when $\theta_1 = \theta_2 = \theta$, since $\rho \leq 1 + \theta$. However, if $\theta_2 \gg \theta_1$ and θ_1 is small, then for larger values of ρ , $\|(\boldsymbol{\Phi} + \boldsymbol{\theta}_\delta)^{-1} \boldsymbol{\Phi}\|_1 > 1$. In this case, the effect of the linear transformation (C.9) on $\boldsymbol{\Gamma}$ will be to increase its norm, so that some of the elements of $\boldsymbol{\Gamma}$ will be increased. Provided the norm is less than one, as in the cases described above, the effect of bias will be to attenuate the regression coefficients.

In summary, the bias due to ignoring measurement error can lead to attenuation of parameter estimates, but it can also cause inflation. For two explanatory variables, this

can happen, when there is a large difference in the magnitude of the measurement errors of the explanatory variables. In this thesis, only cases of attenuation have been explored, as these comprise the most typical and realistic cases, particularly for the case of two exogenous latent variables.

Bibliography

- Aaker, D. A., Kumar, V., & Day., G. S. (2004). *Marketing research* (8th ed.). NY, New York: John Wiley and Sons.
- Adams, R. & Wu, M. (2000). *PISA 2000 Technical Report*. OECD: Organisation for Economic Co-operation and Development.
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: an approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22(1), 47-76.
- Alkhadher, O. & Clarke, D. D. (1998). Equivalence and predictive validity of paper-and-pencil and computerized adaptive formats of the differential aptitude tests. *Journal of Occupational & Organizational Psychology*, 71(3): 205-217.
- Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. V). Berkeley: University of California.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Arbuckle, J. L., & Wothke, W. (1999). *AMOS 4.0 User's Guide*, SPSS.
- Arnett, D. B., Laverie, D. A., & Meiers, A. (2003). Developing parsimonious retailer equity indexes using partial least squares analysis: a method and applications. *Journal of Retailing & Consumer Services*, 79(3): 161-180.
- Auh, S., Salisbury, L. C., & Johnson, M. D. (2003). Order Effects in Customer Satisfaction Modelling. *Journal of Marketing Management*, 19(3/4): 379-400.
- Bandalos, D. L. & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269-296). Mahwah, NJ, US: Lawrence Erlbaum Associates.

- Bagozzi, R. P. & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling*, 1: 35-67.
- Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. New York: M. Dekker.
- Baker, F. B. (2001). *The Basics of Item Response Theory* (2nd ed.): ERIC Clearinghouse on Assessment and Evaluation.
- Balasubramanian, S. K. & Kamakura, W. A. (1989). Measuring Consumer Attitudes Toward the Marketplace With Tailored Interviews. *Journal of Marketing Research*, 26(3): 311-326.
- Barclay, D. W., Higgins, C., & Thompson, R. (1995). The partial least squares (PLS) approach to causal modeling: Personal computer adaptation and use as an illustration. *Technology Studies*, 2(2): 285-309.
- Bartholomew, D. J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology*, 34, 93-99.
- Baumgartner, H. & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, 13(2): 139-161.
- Bawa, K. & Srinivasan, S. S. (1997). Coupon attractiveness and coupon proneness: A framework for modeling coupon redemption. *Journal of Marketing Research*, 34(4): 517-525.
- Bearden, W. O., & Netemeyer, R. G. (1999). *Handbook of marketing scales: multi-item measures for marketing and consumer behavior research* (2nd ed.). Thousand Oaks, Calif.: Sage Publications.
- Bentler, P. M. (1995). *EQS Structural Equations Program Manual*, Encino, Calif.: Multivariate Software.
- Birkinshaw, J., Morrison, A., & Hulland, J. (1995). Structural and competitive determinants of global integration strategy. *Strategic management journal*, 16(8): 637-655.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 396-479). Reading, MA: Addison-Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.

- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm", *Psychometrika*, 46(4), 443-459.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35(2), 179-197.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin*, 110(2), 305-314.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, N.J.: L. Erlbaum.
- Brooker, G. (1975). An instrument to measure consumer self-actualization. *Advances in Consumer Research*, 2, 563-575.
- Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean-and covariance-structure models. In G. Arminger & C. C. Clogg & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185-250). New York: Plenum Press.
- Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern Methods for business research* (pp. 295-336). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Chin, W. W. & Todd, P. A. (1995). On the use, usefulness and ease of use of structural equation modeling in MIS research: a note of caution. *MIS Quarterly*, 19(2): 237-246.
- Christofferson, A. (1975). Factor Analysis of Dichotomized Variables. *Psychometrika*, 40(1), 5-32.
- Churchill, G. A. J. and Iacobucci, D. (2002). *Marketing Research: Methodological Foundations*, South-Western.
- Cooper, D. R., & Schindler, P. S. (1998). *Business Research Methods* (6th ed.). Irwin/McGraw-Hill.
- Cote, J. A. & Buckley, M. R. 1988. Measurement error and theory testing in consumer research: An illustration of the importance of construct validation. *Journal of Consumer Research*, 14(4): 579-583.
- Craig, C. S. & Douglas, S. P. (2000). Survey research: instrument design, *International Marketing Research*: 199-227. New York, NY: John Wiley & Sons.

- Craig, C. S. & Douglas, S. P. (2000). Multi-country scale development, *International Marketing Research*: 256-288. New York, NY: John Wiley & Sons.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Croon, M. (2002). Using Predicted Latent Scores in General Latent Structure Models. In G. A. Marcoulides & I. Moustaki (Eds.) *Latent Variable and Latent Structure Models* (pp. 195-223), Mahwah, NJ: Lawrence Erlbaum Associates.
- Cyr, A. & Cartwright, F. (2004). Measurement methods and techniques, a paper presented to *Statistics Canada's Advisory Committee on Statistical Methods*, Meeting No. 38.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1-38.
- DeVellis, R. F. (1991). *Scale Development: Theory and Applications* (Vol. 26). Newbury Park, California: SAGE.
- Dunbar, G. (1998). *Data Analysis for Psychology*. Naomi Meredith.
- Duxbury, L. & Higgins, C. (1991). Gender differences in work-family conflict. *Journal of Applied Psychology*, 76: 60-74.
- Duxbury, L. & Higgins, C. (2001). Work-life balance in the new millennium: Where are we? Where do we need to go? *CPRN Discussion Paper*, W/12. CPRN: Ottawa.
- Duxbury, L. & Higgins, C. (2002). *The 2001 National Work-Life Conflict Study: Report One*. Ottawa: Health Canada.
- Duxbury, L. & Higgins, C. (2003). *Work-Life Conflict in Canada in the New Millennium: A Status Report*, Health Canada.
- Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, 20(3), 201-212.
- Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 1-15). Mahwah, N.J.: L. Erlbaum Associates.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: L. Erlbaum Associates.

- Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357-381.
- Fay, R. E. (1992). *When are inferences from multiple imputation valid?* Paper presented at the Proceedings of the Survey Research Section of the American Statistical Association, Washington, DC.
- Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent development, and applications* (pp. 15-38). New York: Springer-Verlag.
- Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, 45(1), 30-45.
- Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78(2), 350-365.
- Fuller, W. A. (1987). *Measurement Error Models*, New York: Wiley.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: Macmillan.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.
- Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Newbury Park, Calif.: SAGE.
- Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *Counseling Psychologist*, 27(3), 353-383.
- Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, 71(1), 105-131.
- Hattie, J. (1984). An empirical study of various indices for determining unidimensionality, *Multivariate Behavioral Research*, 19, 49-78.

- Hattie, J., Krakowski, K., Rogers, H. J. and Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality, *Applied Psychological Measurement*, 20:1, 1-14.
- Hoijtink, H., & Boomsma, A. (1996). Statistical inference based on latent ability estimates. *Psychometrika*, 61(2), 313-330.
- Holbert, R. L. & Stephenson, M. T. (2002). Structural Equation Modeling in the Communication Sciences. *Human Communication Research*, 28(4): 531-551.
- Hunt, S. D., Chonko, L. B., & Wood. V. R. (1985). Organizational commitment and marketing. *Journal of Marketing*, 49, 112-126.
- Johansson, J. K. & Yip, G. S. (1994). Exploiting globalization potential: U.S. and Japanese strategies. *Strategic management journal*, 15(8): 579-601.
- Johnson, R. A., & Wichern, D. W. (1998). Factor analysis and inference for structured covariance matrices, *Applied Multivariate Statistical Analysis* (4th ed., pp. 514-586). Upper Saddle River, NJ: Prentice-Hall.
- Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*, Chicago: Scientific Software International.
- Jöreskog, K. G. (2000). *Latent variables scores and their uses*. Available: <http://www.ssicentral.com/lisrel/column6.htm> [2002, Nov. 12].
- Jöreskog, K. G., & Sörbom, D. (2001). *LISREL: User's Reference Guide* (2nd ed.). Lincolnwood, IL: Scientific Software International.
- Junker, B. (1988). *Statistical aspects of a new latent trait theory*. Unpublished doctoral dissertation, University of Illinois at Urbana – Champaign.
- Kahai, S. S., Sosik, J. J., & Avolio, B. J. (2004). Effects of Participative and Directive Leadership in Electronic Groups. *Group & Organization Management*, 69(1): 67-105.
- Kamakura, W. & Srivastava, R. K. (1982). Latent trait theory and attitude scaling: The use of information functions for item selection. *Advances in Consumer Research*, 9(1): 251-256.
- Kaplan, R.M. & Saccuzzo, D.P. (1993). *Psychological Testing: Principles, Applications, and Issues*. 3rd ed., Brooks/Cole.
- Kline, P. (1993). *The Handbook of Psychological Testing*. Paul Kline.

- Kim, S. & Hagtvet, K. A. (2003). The impact of misspecified item parceling on representing latent variables in covariance structure modeling: A simulation study. *Structural Equation Modeling*, 10(1): 101-127.
- Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, 58(4), 587-599.
- Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement*, 7(1), 15-32.
- Lagace, R. R., Ingram, T. N., & Boorum, M. L. (1994). An exploratory study of buyers' perceptions of salesperson unethical behavior: scale development and validation. In R. Achrol & A. Mitchell (Eds.), *Enhancing Knowledge Development in Marketing*, Vol. 5: 118-123. Chicago: American Marketing Association.
- Laroche, M., Kim, C., & Tomiuk, M. A. (1998). Translation fidelity: An IRT analysis of Likert-type scale items from a culture change. *Advances in Consumer Research*, 25(1): 240-245.
- Laroche, M., Kim, C., & Tomiuk, M. A. (1999). IRT-based item level analysis: an additional diagnostic tool for scale purification. *Advances in Consumer Research*, 26: 141-149.
- Lawson, S. (1991). One parameter latent trait measurement: Do this results justify the effort? In B. Thompson (Ed.), *Advanced in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAI.
- Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete data likelihood, *American Statistician*, 37, 218-220.
- Liu, Y. (2003). Developing a scale to measure the interactivity of websites. *Journal of Advertising Research*, 45(2): 207-216.
- Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph*, 7.
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13, 517-548.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, 48(2), 233-245.

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79(386), 393-398.
- Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2003). Embedding IRT in structural equation models: a comparison with regression based on IRT scores. Accepted by *Structural Equation Modeling*.
- Lu, I. R. R., Thomas, D. R., & Orser, B. J. (2004). Latent variable modeling in business research: A comparison of two-step approach with structural equation modeling. *Proceedings of Administrative Sciences Association of Canada*, 32nd Annual ASAC Conference, Québec city, Québec, June 5-8, 2004.
- Lundstrom, W. J. & Lamont, L. M. (1976). The development of a scale to measure consumer discontent. *Journal of Marketing Research*, 13(November): 373-381.
- Mackenzie, S. B. (2001). Opportunities for improving consumer research through latent variable structural equation modeling. *Journal of Consumer Research*, 28: 159-166.
- MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62(6), 921-943.
- Maheswaran, D. & Shavitt, S. (2000). Issues and New Directions in Global Consumer Psychology. *Journal of Consumer Psychology*, 9(2): 59-66.
- Marcoulides, G. A., Ed. (1998). *Modern Methods for Business Research*. Mahwah, New Jersey, Lawrence Erlbaum Associates.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Marcoulides, G. A. (1998). Applied generalizability theory models. In G. A. Marcoulides (Ed.), *Modern Methods for Business Research* (pp.1-22). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Marcoulides, G. A. (1999). Generalizability theory: picking up where the Rasch IRT model leaves off? In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement: What Every Psychologist and Educator Should Know* (pp. 129-152). Mahwah, N.J.: L. Erlbaum Associates,.

- Martin, I. & Eroglu, S. (1993). Measuring a multi-dimensional construct: country image. *Journal of Business Research*, 28, 191-210.
- Martin, G. & Staines, H. (1994). Managerial competences in small firms. *Journal of Management Development*, 13(7): 23-34.
- McCollam, K. M. S. (1998). Latent trait and latent class models. In G. A. Marcoulides (Ed.), *Modern Methods for Business Research* (pp. 23-46), Mahwah, New Jersey: Lawrence Erlbaum Associates.
- McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monograph*, 15.
- McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research*, 31(2), 239-270.
- McDonald, R. P. (1999). *Test theory: a unified treatment*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Meade, A., Hecht, J., Lautenschlager, G., Barroso, C., & Stokes, G. (2000). Using item response theory to identify beta change over time in Likert-scale job satisfaction data, *Academy of Management Proceedings*.
- Meng, X. L. (1994). Multiple imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9, 538-573.
- Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, 11(1), 81-91.
- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics* 17(2), 131-154.
- Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22, 43-65.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115-132.
- Muthén, B. (1988). Some uses of structural equation modeling in validity studies: extending IRT to external variables. In H. Wainer and H. I. Braun (Eds.), *Test Validity* (pp. 213-238). Princeton, NJ: Educational Testing Service.
- Muthén, B., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, 28(1), 1-22.

- Muthén, L. K., & Muthén, B. O. (2001). *Mplus User's Guide*, Los Angeles, CA: Muthén & Muthén.
- Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality, *Journal of Educational Measurement*, 28:2, 99-117.
- Nandakumar, R. (1993). Assessing essential unidimensionality of real data, *Applied Psychological Measurement*, 17:1, 29-38.
- Ndalichako, J. L., & Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement*, 57(4), 580-589.
- Newton, K. (2001). *Management Skills for Small Business*. A report submitted to Small Business Policy Branch, Industry Canada.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). N.Y.: McGraw-Hill.
- Noble, B. & Daniel, J. W. (1977). *Applied Linear Algebra*. New Jersey: Prentice-Hall.
- Novick, M. R. (1966). The axioms and principal results of classical test theory, *Journal of Mathematical Psychology*, 3: 1-18.
- Novick, M. R. and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements, *Psychometrika*, 32(1): 1-13.
- O'Brien, R. M. (1981). Using rank category variables to represent continuous variables: Defects of common practice. *Social Forces*, 59: 1149-1162.
- Orser, B. J. & Riding, A. L. (2003). *Management Competencies and SME Performance Criteria: A Pilot Study*. A report to Small Business Policy Branch, Industry Canada.
- Orser, B., Riding, A., and Madill, J. (2002) The Need for Gender-based Small Business Programs: What Does the Evidence Say? *Frontiers of Entrepreneurship Research 2002, Summary Proceedings of the 22nd Annual Entrepreneurship Research Conference*, Boulder Colorado, June 6-9, 2002.
- Oshima, T. C. and Miller, M. D. (1990). Multidimensionality and IRT-based item invariance indexes: the effect of between-group variation in trait correlation, *Journal of Educational Measurement*, 27:3, 273-283.
- Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology*, 18(3), 267-296.

- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), 146-178.
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8, 287-312.
- Peter, J. P. (1979). "Reliability: a review of psychometric basics and recent marketing practices." *Journal of Marketing Research*. XVI (February), 6-17.
- Power, J. L. 2004. *Caregiving, Gender, and Role Interference*. Ph.D. Dissertation, Carleton University, Ottawa, Canada.
- Probst, T. M. (2003). Development and validation of the Job Security Index and the Job Security Satisfaction scale: A classical test theory and IRT approach. *Journal of Occupational & Organizational Psychology*, 76(4): 451-467.
- Ramsay, J. O. (1995). *TestGraf: A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data*. Montreal, Canada: McGill University.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.
- Ravichandran, T. & Rai, A. (2000). Quality management in systems development: An organizational system perspective. *MIS Quarterly*, 24(3): 381-415.
- Reckase, M. D. (1994) What is the 'correct' dimensionality for a set of item response data? in D. Laveault et al. (Eds.), *Modern Theories of Measurement: Problems and Issues* (pp. 87-92). University of Ottawa: Ottawa.
- Reckase, M. D. (1997). A linear logistic multidimensional item response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.
- Reise, S. P. 1999. Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Hersberger (Eds.), *The New Rules of Measurement* (pp. 219-242). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Rigdon, E. E. (1998). Structural equation modeling. In G. A. Marcoulides (Ed.), *Modern Methods for Business Research* (pp. 251-294). Mahwah, New Jersey: Lawrence Erlbaum Associates.

- Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance*, 14(2): 187-207.
- Robins, J. M., & Wang, N. S. (2000). Inference for imputation estimators. *Biometrika*, 87(1), 113-124.
- Rook, D. and Fisher, R. J. (1995). Normative influences on impulsive buying behavior. *Journal of Consumer Research*, 22, 305-313.
- Rost, J. (1990). Rasch model in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14: 271-282.
- Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *British Journal of Mathematical & Statistical Psychology*, 44: 75-92.
- Rubin, D. B. (1987). *Multiple Imputation For Non-Response in Surveys*, New York: Wiley.
- Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No.17.
- Schneeweiss, H. (1993). Consistency at large in models with latent variables. *Statistical modelling and latent variables*. D. J. B. a. M. D. K. Haagen. North Holland, Amsterdam, New York and Oxford, Elsevier Science Publishers B.V.: 299-320.
- Schoenberg, R., & Arminger, G. (1990). *LINCS2.0*. Kent. WA: RJS Software.
- Shevlin, M., Miles, J. N. V., & Bunting, B. P. (1997). Summated rating scales: a Monte Carlo investigation of the effects of reliability and collinearity in regression models. *Personality and Individual Differences*, 23(4), 665-676.
- Shinn, A. M. J. (1974). Relations between scales. In H. M. J. Blalock (Ed.), *Measurement in the Social Sciences: Theories and Strategies*. Chicago: Aldine.
- Shook, C. L., Ketchen Jr., D. J., Hult, G. T. M., & Kacmar, K. M. (2004). An assessment of the use of structural equation modeling in strategic management research. *Strategic Management Journal*, 24(5): 397-304.
- Singh, J. 2004. Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research*, 57(2): 184-208.
- Singh, J. & Howell, R. D. (1990). Adaptive designs for likert-type data: An approach for implementing marketing surveys, *Journal of Marketing Research*, 27(3): 304-321.

- Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, 25(4), 391-415.
- Skron dal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom." *Multivariate Behavioral Research*, 35(2): 137-167.
- Skron dal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, 66(4), 563-576.
- Slama, M. E. & Tashchian, A. (1985). Selected socio-economic and demographic characteristics associated with purchasing involvement. *Journal of Marketing*, 49, 72-82.
- Smith, J. B. & Bristor, J. M. (1994). Uncertainty orientation explaining differences in purchase involvement and external search. *Psychology & Marketing*, 11(6): 587-607.
- Steenkamp, J.-B. E. & Baumgartner, H. (2000). On of the use of structural equation models for marketing modeling. *International Journal of Research in Marketing*, 17: 195-202.
- Stephenson, M. T., & Holbert, R. L. (2003). A Monte Carlo simulation of observable versus latent variable structural equation modeling techniques. *Communication Research*, 30(3), 332-354.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52:4, 589-617.
- Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293-325.
- Stroud, A. H., & Secrest, D. (1966). *Gaussian Quadrature Formulas*, Englewood Cliffs, N.J.: Prentice-Hall.
- Suen, H. K. (1990). *Principles of Test Theories*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Sujan, H., Weitz, B., & Kumar, N. 1994. Learning orientation, working smart, and effective selling. *Journal of Marketing*, 58: 39-52.
- Sweeney, J. C., Hausknecht, D., & Soutar, G. N. (2000). Cognitive dissonance after purchase: A multidimensional scale. *Psychology & Marketing*, 17(5): 369-385.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393-408.

- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Thissen, D., & Steinberg, L. (1987). Taxonomy of item response models. *Psychometrika*, 51, 567-577.
- Thomas, D. R. (2001). *Item response theory and its application to the national longitudinal survey of children and youth*. Ottawa: Statistics Canada, Social Surveys Methods Division.
- Thomas, D.R., Lu, I.R.R., & Zumbo, B.D. (2002). Using IRT and Factor Scores in Regression and Other Analyses. Paper presented at the XIX International Methodology Symposium and Workshops: *Proceedings of the Modeling Survey Data for Social and Economic Research*, Statistics Canada, Ottawa.
- Thomas, D. R., Singh, A. C., & Roberts, G. R. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistics Review*, 3(64), 295-311.
- Thomas, D. R., & Cyr, A. (2002). *Applying item response theory methods to complex survey data*. Paper presented at the SSC Annual Meeting: Proceedings of the Survey Methods Section, Ottawa.
- Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45: 253-275.
- Thomas, D.R., & Zumbo, B.D. (2002). Item response theory and related methods with applications to complex sample surveys. Workshop lecture notes at the XIX International Methodology Symposium and Workshops. Statistics Canada, Ottawa.
- Traub, R. E. (1994). *Reliability for the Social Sciences: Theory and Applications* (Vol. 3). California: Sage.
- Tucker, L. (1971). Relations of factor score estimates to their use. *Psychometrika*, 36(4), 427-436.
- Tull, D.S., & Hawkins, D.I. (1993). *Marketing Research: Measurement and Method: A Text with Cases*. 6th ed., Macmillan.
- Van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, 25(3), 273-282.

- Wainer, H., & Thissen, D. (2001). True score theory: the traditional method. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 23-72). Mahwah, New Jersey: Lawrence Erlbaum.
- Walsh, W. B., & Betz, N. E. (2000). *Tests and assessment* (4th ed.). Upper Saddle River, N.J.: Prentice Hall.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427-450.
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables. In R. H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications* (pp. 56-75). California: SAGE.
- Wixom, B. H. & Watson, H. J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, 25(1): 17-38.
- Wold, H. (1982). Soft modeling: the basic design and some extensions. In K. G. Jöreskog & H. Wold (Eds.), *Systems Under Indirect Observation: Causality, Structure, Prediction* (Vol. 2, pp. 1-54). Amsterdam: North-Holland.
- Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science*, 7(4), 104-109.
- Zickar, M. J. & Highhouse, S. (1998). Looking closer at the effects of framing on risky choice: An item response theory analysis. *Organizational Behavior & Human Decision Processes*, 75(1): 75-91.
- Zickar, M. J. (2002). Modeling data with polytomous item response theory. In F.S. Drasgow, Neal (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 123-155). San Francisco, CA, US: Jossey-Bass/Pfeiffer.
- Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, 7(2): 168-190.
- Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, 56(4), 589-600.