A MODEL FOR DETECTING LACK OF INVARIANCE FOR ITEM RESPONSES AND RESPONSE TIMES

A Dissertation

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Emily Hailey

May 2014

© Copyright by
Emily Hailey
All Rights Reserved
May 2014

TABLE OF CONTENTS

LIST OF TABLES v
LIST OF FIGURES vi
LIST OF APPENDICES vii

CHAPTER

I. INTRODUCTION 1
II. REVIEW OF LITERATURE.....10
III. METHODOLOGY 43
IV. RESULTS.....50
V. DISCUSSION.....63

REFERENCES 68

APPENDICES 79

LIST OF TABLES

TABLE

1. Mean RMSE and Bias for Response Time Parameters for Null Conditions ...50
2. Mean RMSE and Bias for Time Discrimination Parameters for LOI Conditions ...51
3. Mean RMSE and Bias for Item Time Intensity Parameters for LOI Conditions ...52
4. Mean RMSE and BIAS for Item Response Parameters for Null Conditions ..53
5. Mean RMSE and Bias for Discrimination Parameters for LOI Conditions ...54
6. Mean RMSE and Bias for the Threshold Parameters for LOI Conditions 55
7. Empirical Type I Error Rates 56
8. Empirical Type I Error Rates 56
9. Empirical Power Rates.....58
10. The Largest 10 Variance Components for Type I Error.....59
11. The Largest 10 Variance Components for Power 60
12. The β/SE Ratios for Item Responses and Response Times 61
13. MIMIC-IRTRT LOI Parameter Estimates and P-values.....62

LIST OF FIGURES

FIGURE

1. Illustration of SEM-IRTRT Model for Five Items 4
2. A MIMIC Model for Testing Uniform LOI 7
3. Illustration of the MIMIC-IRT model 8
4. Hierarchical Item Response and Response Time Model 14
5. Illustration of the SEM-IRTRT Model 18
6. A MIMIC-interaction model for testing non-uniform and uniform LOI 30
7. MIMIC-IRTRT Model for Item Responses and Response Times 40

LIST OF APPENDICES

APPENDIX

1. R Code for Data Simulation.....79
2. Mplus Code for MIMIC-IRTRT Model.....83

Abstract

The objective of this study was to conduct exploratory research to determine the viability of a new model for detecting lack of invariance (LOI) for both item responses and response times. LOI occurs when the property of parameter invariance, which states that item parameters are invariant across examinee populations and person parameter are invariant across sets of items, is violated (Rupp & Zumbo, 2006). LOI can be present for item responses, which refer to whether an examinee answered an item correctly or incorrectly, as well as item response times, which refer to how much time an examinee spent answering an item. LOI can be problematic as it has consequences for validity and fairness of the test (Gierl, 2005).

Currently, much of the research on LOI is relegated to studies of item responses (see Demars, 2004a, 2004b; Lord, 1977; Rupp & Zumbo, 2006; Wells, Subkoviak, & Serlin, 2002). Little research has been done on LOI for response times (Demars & Wise, 2010; Klein Entink, 2009; van der Linden, Schnipke, & Scrams, 2007), and no research has looked at LOI for item responses and response times at the same time. As such, this study evaluated a model, multiple indicator multiple cause model for detecting LOI in item responses and response times (MIMIC-IRTRT) that can examine LOI for both item response and response time simultaneously.

This study is conducted in two parts consisting of both a simulation and extant data analysis. In these studies, only uniform LOI was examined. In the simulation study, number of items, correlation between person ability and speed, number of LOI items,

type of LOI, and magnitude of LOI were manipulated. In the extant data analysis, high-stakes, college-level, health profession exam data that was suspected to possess compromised test items was analyzed with the MIMIC-IRTRT model. The results from both the simulation and extant data study provide support for the use of the MIMIC-IRTRT model in detecting LOI.

CHAPTER 1

INTRODUCTION

The objective of this study is to consider a new method for detecting a lack of parameter invariance in response times as well as item responses. The key concepts from this objective will be briefly discussed. The property of parameter invariance is a cornerstone in item response theory (IRT) and states that item parameters are invariant across examinee populations and person parameters are invariant across sets of items. This is a property of the parameters and therefore must be tested for parameter estimates. When the property does not hold for parameter estimates, there is said to be a lack of invariance (LOI; Rupp & Zumbo, 2006). A LOI has consequences for validity and fairness (Gierl, 2005). LOI can be present for item responses, which refer to whether an examinee answered an item correctly or incorrectly, as well as item response times, which refer to how much time an examinee spent answering an item.

The literature in educational measurement currently contains a substantial number of investigations and studies related to a LOI for item responses (see Demars, 2004a, 2004b; Lord, 1977; Rupp & Zumbo, 2006; Wells, Subkoviak, & Serlin, 2002). The research for LOI for response times is nascent and the research that is available examines a LOI in response time separate from a LOI in item response (Demars & Wise, 2010; Klein Entink, 2009; van der Linden, Schnipke, & Scrams, 2007). Currently, there is no method in place that would allow for the testing of LOI for both item responses and response times at the same time. The inclusion of response times with item responses has

been shown to improve person parameter estimation for item response parameters (Fox, Klein Entink, & van der Linden, 2007; van der Linden, Klein Entink, & Fox, 2010). In addition, eliminating LOI for both item response and response times can improve validity and fairness of a test (Gierl, 2005; Haladyna & Downing, 2004). As such, I propose a model that can examine LOI for both item response and response time simultaneously. The following sections provide a more detailed description of the key concepts for this study as well as the proposed model.

Overview and Background

Item response time has been used in psychological measurement as a means of examining cognitive processing for several decades (Luce, 1986). However, it has been neglected in large-scale educational measurement because traditional paper-and-pencil tests do not lend themselves to capturing item response time. It has only been through the introduction of computer-based testing that researchers can easily capture examinees' item response times in addition to their item responses. Important information can be gleamed from response times for both items and examinees. The information provided from response times may be helpful in improving item calibration, test design, item selection for adaptive test, diagnosis of aberrant responses, and test accommodations (Schnipke & Scrams, 1999; van der Linden, 2006, 2007; van der Linden, et al., 2010). As such many models have tried to best capture examinees item response times. Some models focus *solely* on modeling response times (see Maris,1993; Scheilechner, 1979; Schnipke & Scrams, 1997; van der Linden, 2006). Some models estimate item responses and response time *separately* (see Bejar & Yocom, 1991; Embretson, 1998; Gorin, 2005; Mulholland, Pellegrino, & Glaser, 1980; Primi, 2001), while others have purported joint

models for estimating item response and response time *simultaneously* (see Roskam, 1997; Thissen, 1983; van Breukelen, 1989; Verhelst, Verstraalen, & Jansen, 1997). Van der Linden's (2007) hierarchical item response and response time model extends upon the models that estimate both item responses and response times simultaneously by allowing for relationships between the two model types. An item response model (IRM) and a log-normal response time model are put in a hierarchical framework, where the first level is the separate models and the second level represents the relationships between the parameters of the models. The item response model is the two-parameter normal-ogive model, where the probability of a correct response on item *i* for examinee *j* is given as

$$P(U_{ij} = 1; \theta_j, a_i, b_i) = \phi(a_i(\theta_j - b_i))$$

where ϕ is the normal distribution function, θ_j is the ability parameter for examinee *j*, a_i is the discrimination parameter for item *i*, and b_i is the difficulty parameter for item *i*.

The response time model is the log-normal model described by van der Linden (2006) and states that the observed response time, t_i, for an examinee on item *i* is a realization of a random variable *T* as given by

$$f(t_i; \tau, \alpha_i, \beta_i) = \frac{\alpha_i}{t_i\sqrt{2\pi}}\, exp\left\{-\frac{1}{2}\left[\alpha_i\left(ln\, t_i - (\beta_i - \tau_j)\right)\right]^2\right\}.$$

Wherein this model t_i is the observed response time for item *i*, τ_j is the speed of the test for examinee *j* (i.e. the person speed parameter), β_i is the time intensity parameter for item *i*, and α_i is the item time discrimination parameter. The larger τ_j, the less amount of time the examinee spends on all items and the faster she operates. The item time intensity parameter has a similar interpretation. The larger the item time intensity, the more time examinees spend on the item. Finally, the item time discrimination parameter

is the reciprocal of the standard deviation of the response-time distribution. The larger its value, the less variability in log-times on item *i* for all examinees.

Recently, Sen (2012) proposed a structural equation model (SEM-IRTRT) approach to van der Linden's hierarchical model. This approach combines a two-parameter logistic IRT model with a log-normal model for item response time (Finger & Chuah, 2009) into a single confirmatory factor analysis model (CFA). Finger and Chuah (2009) transformed van der Linden's response time model (2006) into a CFA framework. In this model, the log of response time, ln_{t_i}, follows a single-factor model with intercept:

$$ln_{t_i} = v_i + \lambda_i \xi + \varepsilon_i$$

where v_i represents the time intensity for item *i*, λ_i is the loading factor for item *i* on the single factor, ξ is the person speed parameter, and ε_i is the residual. The resulting structural equation model that incorporates both item response and response times is detailed in Figure 1.

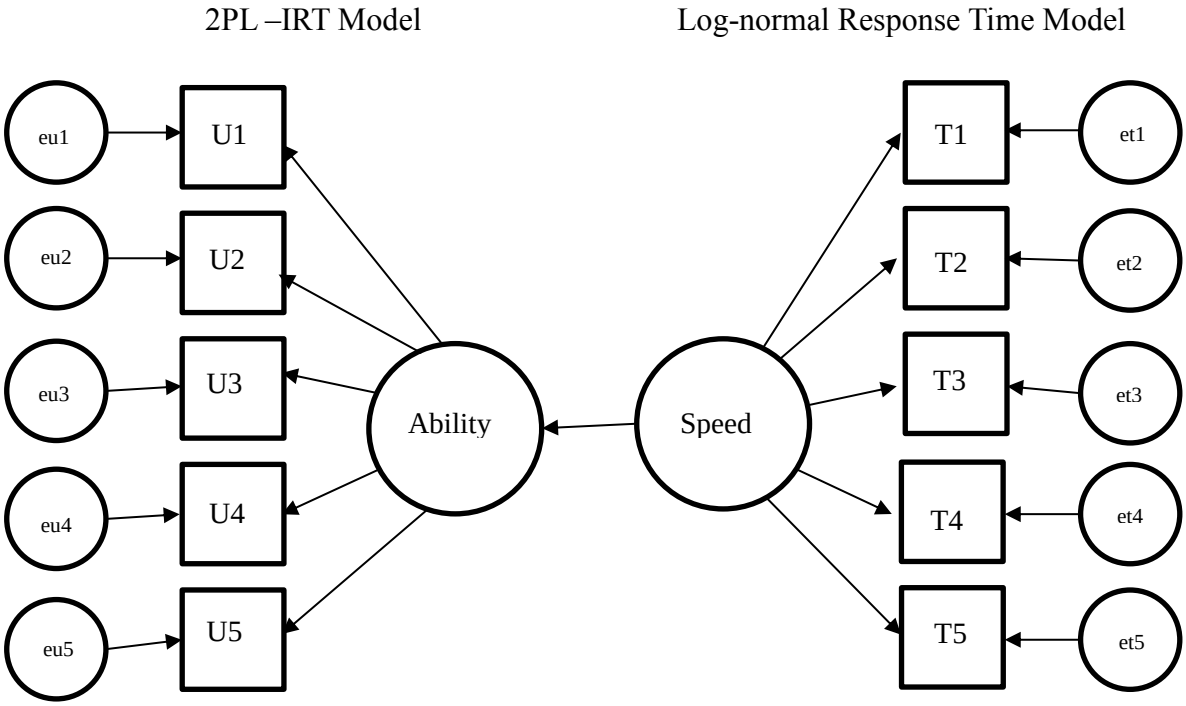

Figure 1. Illustration of the SEM-IRTRT model for five items.

Parameter Invariance

One important measurement property is that of parameter invariance where item parameters are invariant across examinee populations and person parameters are invariant across a set of items, referred to as item parameter invariance and person parameter invariance, respectively. Parameter invariance is a property of the parameters and may not always hold for the estimates. As such, the estimates need to be tested to determine whether parameter invariance holds. When the property does not hold, it is referred to as a lack of variance (LOI; Rupp & Zumbo, 2006). If the LOI occurs between two groups on the same test, it is referred to as differential item functioning (DIF; Dorans & Holland, 1993). If the LOI occurs between two groups taking the same test on different occasions, it is referred to as item parameter drift (drift; Goldstein, 1983).

A LOI can be problematic for a variety of reasons. When there is a LOI, then results cannot be generalized across examinee populations or measurement conditions (Rupp & Zumbo, 2006). Also, if parameter estimates are not invariant, the results from scaling and equating will not be accurate (Hu, Rogers, & Vukmirovic, 2008; Sukin, 2010; Wells et al., 2002). Finally, issues of validity and fairness arise in the presence of a LOI (Gierl, 2005).

LOI can manifest as uniform LOI, non-uniform LOI or both. Uniform LOI only impacts item difficulty estimates. Non-uniform LOI impacts item discrimination estimates. If both item difficulty and discrimination estimates are impacted, then both uniform and non-uniform LOI are present (Mellenbergh, 1982).

Most of the existing research on LOI has largely focused on IRT models. The concept of LOI is relatively new to response time models and as such there is not a

consensus on the terminology. *Differential speededness* refers to the situation in timed computer adaptive testing in which examinees are matched on ability but experience different items with varying time intensities and therefore may experience varying levels of time pressure (van der Linden, Breithaupt, Chuah & Zhang, 2007). *Differential rapid-guessing* refers to the situation where sub-groups of examinees (those who exhibit rapid-guessing behavior and those who exhibit solution-based behavior), who are matched on ability, require different amounts of time to respond to an item (DeMars & Wise, 2010). *Time differential item functioning* (time-DIF) refers to the situation where examinees, who are matched on speed, require different amounts of time to respond to an item (Klein Entink, 2009). The latter definition of LOI in response times is the one of interest in this study.

There are several methods for detecting LOI for item responses, such as Mantel-Haenszel, (Holland & Thayer, 1998), Lord's (1980) chi-square, SIBTEST (Shealy & Stout, 1993), Raju's (1990) signed and unsigned areas, Kim and Cohen's (1991) signed and unsigned closed-intervals, and the multiple indicators multiple causes (MIMIC) model (Muthén, 1985). The MIMIC model is particularly well-suited to examining LOI in item response and item response time. The MIMIC model utilizes a confirmatory factor analysis framework to examine LOI. The model for testing uniform LOI in item responses relies on the following equation:

$$y_i^* = \alpha_i\theta + \phi_i z + \varepsilon_i,$$

where y_i^* is the latent response variable for item i, θ is the latent trait, α_i is discrimination parameter for item i, z is a dummy variable indicating group membership, ϕ_i indicates the relationship between the grouping variable and the item response, and ε_i is the

random error. The presence of LOI is determined by examining the significance of the estimate, ϕ (Woods & Grimm, 2011).

This model can be illustrated using a confirmatory factor analysis approach as shown in Figure 2.

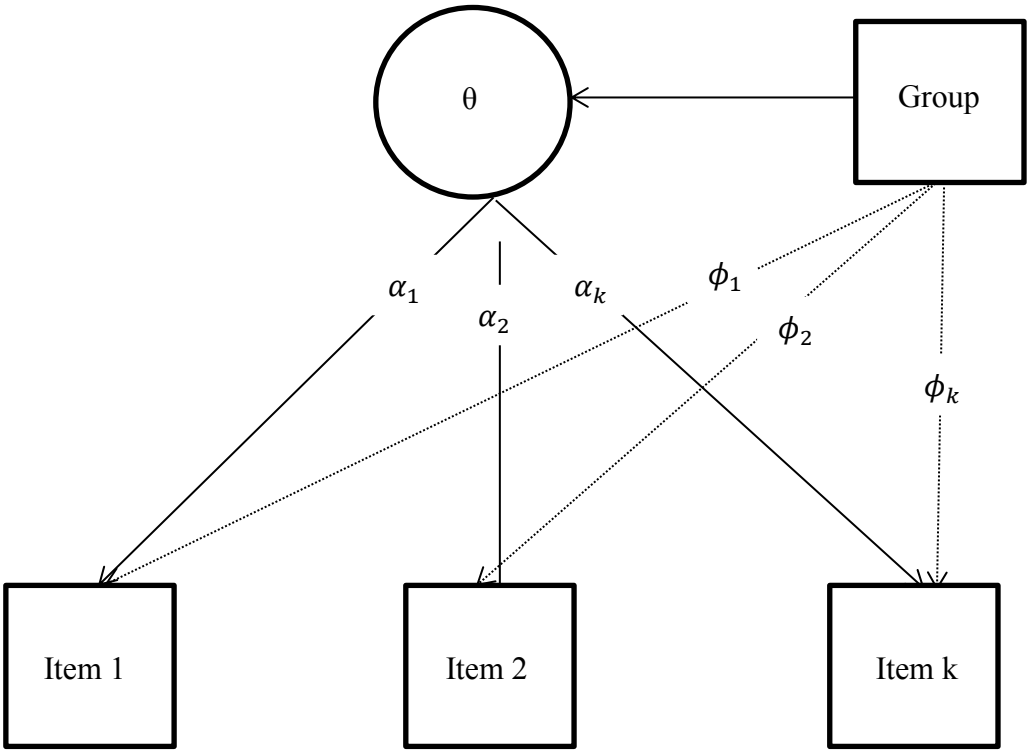

Figure 2. A MIMIC model for testing uniform LOI.

Multiple Indicator Multiple Cause Model for Item Response and Response Time

Much of the research has focused on evaluating LOI for item responses only. Therefore, I propose a model that can evaluate LOI for both item responses and response times. The proposed model combines Sen's (2012) SEM formulation of the hierarchical model and the MIMIC model. This model will be referred to as the multiple indicator multiple cause model for detecting LOI in item responses and response times (MIMIC-IRTRT) model. An illustration depicting the MIMIC-IRTRT is provided in Figure 3.

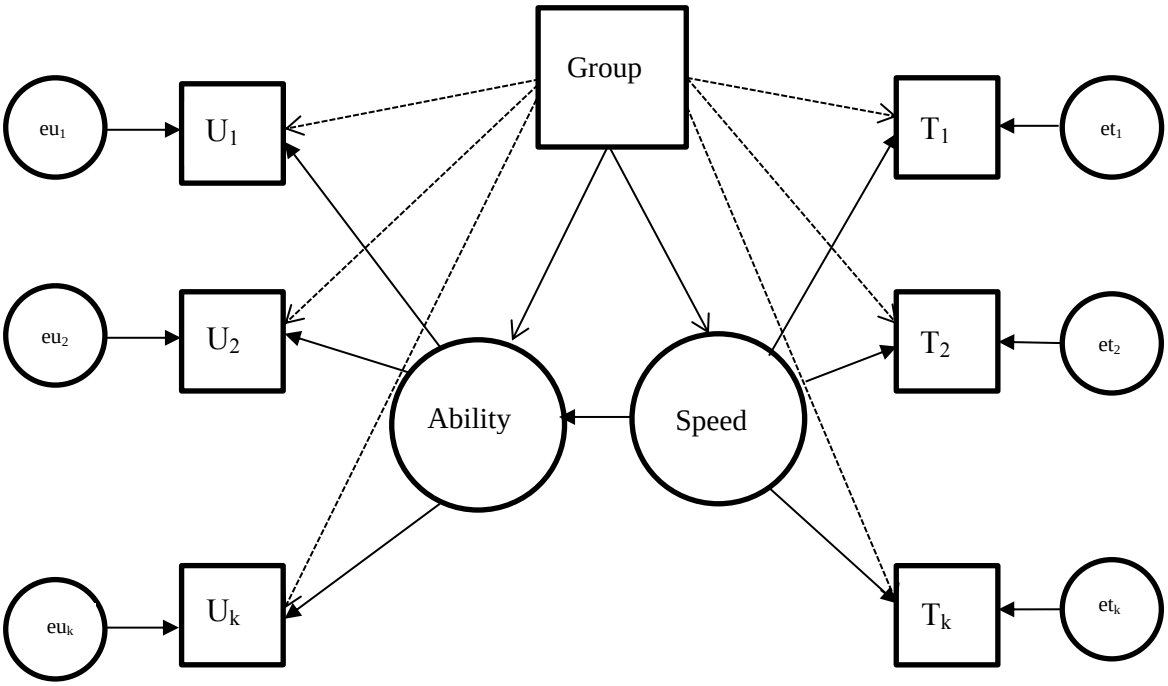

Figure 3. Illustration of the MIMIC-IRTRT model

The MIMIC-IRTRT model is run with LOI effect paths, paths from the grouping variable to the item response or response time, for both item response and response time at the same time. Items are flagged as exhibiting LOI in item responses when there are significant LOI effect paths for item responses. Items are flagged as exhibiting LOI in the response time when there are significant LOI effect paths for response times.

The purpose of this study is two-fold (1) describe a new model for detecting LOI in both item response and item response time and (2) investigate the viability of the proposed model in detecting LOI under different conditions of test length, type of LOI, number of LOI items, magnitude of LOI, and the correlation between person ability and speed.

In the following chapter, van der Linden's (2007) hierarchical model will be described in greater detail. In addition, relevant research on parameter invariance is

reviewed. Particularly, lack of invariance in item response and response time models will be reviewed. The review of the relevant research will (a) situate the proposed research questions in the context of preceding research and (b) illustrate the need for the proposed study in the field of measurement.

Chapter 3 details the research design and the procedures to be employed in the proposed study.

CHAPTER 2

REVIEW OF THE LITERATURE

This chapter provides a brief description of the history of response time models and an in-depth review of van der Linden's response time model, a popular model that incorporates both item responses and response times. The property of parameter invariance is defined as well as important concepts related to violations of the property. Empirical studies investigating violations of parameter invariance in item response models are summarized and methods for detecting a lack of parameter invariance (LOI) are described. The concept of LOI in response time models is introduced and research is summarized. Finally, a model for detecting a LOI in models that include both item responses and response times is introduced and described in detail.

Response Time Models

Response times for test items are an important source of information for items and examinees. The information provided from response times may be helpful in improving item calibration, test design, item selection for adaptive test, diagnosis of aberrant responses, and test accommodations. Many response time models have been developed in hopes of best capturing the response time distribution. These initial approaches model response times independently of responses for the items, and they, vary by the type of underlying response time distribution. Maris' (1993) model posits a gamma distribution. Scheiblechner's (1979) work uses an exponential distribution, and Schinpke and Scrams (1997) and van der Linden (2006) both utilize the log-normal distribution. However,

with these approaches, the item responses are not taken into account and examinee information is lost. There are approaches that model response time and item responses independently (Bejar & Yocom, 1991; Embretson, 1998; Gorin, 2005; Mulholland et al., 1980; Primi, 2001). Other approaches model item response and item response time by treating response time as a fixed facet in an item response model (Roskam, 1997; Thissen, 1983; van Breukelen, 1989; Verhelst et al., 1997). A limitation of the combined approach is that response time is assumed to be independent of the person ability and the speed at which the examinee takes the test. As such, a limitation of these approaches is that they do not account for the relationship between person ability and speed.

To address this limitation, van der Linden created a model for item response and response time that allows for a relationship between the two sources of information. This model is referred to as the Hierarchical Item Response and Response Time Model and will be described in detail in the following section.

Van der Linden's Hierarchical Item Response and Response Time Model

Van der Linden (2007) developed a model that includes both item responses and response times. The two types of models are placed in a hierarchical framework where the first level is simply the item response model and response time model for each examinee and item. The second level represents the relations between the parameters in the models.

First-Level Models. This level of modeling is illustrated by selecting two specific models for the responses and times on the items; however, other variations could be selected (van der Linden, 2007).

As the first-level model for responses for examinees $j = 1, \ldots, N$ on item $i = 1, \ldots, n$, the two- parameter normal-ogive model (2PNO) is used, which gives the probability of a correct response on item i for examinee j as

$$P(U_{ij} = 1; \theta_j, a_i, b_i) = \phi(a_i(\theta_j - b_i)) \quad (1)$$

where ϕ denotes the normal distribution function, θ_j is the ability parameter for examinee j, a_i is the discrimination parameter for item i, and b_i is the difficulty parameter for item i.

The response time model is the log-normal model described by van der Linden (2006) and states that the observed response time, t_i, for an examinee on item i is a realization of a random variable T as given by

$$f(t_i; \tau, \alpha_i, \beta_i) = \frac{\alpha_i}{t_i\sqrt{2\pi}} \, exp\left\{-\frac{1}{2}\left[\alpha_i\left(ln\, t_i - (\beta_i - \tau_j)\right)\right]^2\right\}. \quad (2)$$

Wherein t_i is the observed response time for item i, τ_j is the speed of the test for examinee j (i.e. the person speed parameter), β_i is the time intensity parameter for item i, and α_i is the item time discrimination parameter. The larger τ_j, the less amount of time the examinee spends on all items and the faster she operates. The item time intensity parameter has a similar interpretation. The larger the item time intensity, the more time examinees spend on the item. Finally, the item time discrimination parameter is the reciprocal of the standard deviation of the response-time distribution. The larger its value, the less variability in log-times on item i for all examinees.

Second-Level Models. This level allows for the incorporation of the speed-accuracy trade-off, sometimes referred to as the speed-ability trade-off for achievement tests. This trade-off is described by a monotonically decreasing relationship between speed and "effective" ability (as evidenced by the number of correct items). This trade-

off is modeled by allowing for a relationship between the person parameters at the population level. The population model describes the joint distribution of the person parameters, θ and τ, in a population, *P*, from which the examinees are assumed to be sampled. This distribution is assumed to be bivariate normal,

$$\xi_j \sim MVN(\mu_p, \Sigma_{\mathrm{p}}), \tag{3}$$

where

$$\mu_p = (\mu_\theta, \mu_\tau) \tag{4}$$

and covariance matrix

$$\Sigma_{\mathrm{p}} = \begin{pmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\tau^2 \end{pmatrix}. \tag{5}$$

The parameter ρ denotes the covariance between the person parameters.

The item-domain model describes the joint distribution of the item parameters a_i and b_i in the response model and α_i and β_i in the response time model. This distribution is assumed to have a multivariate normal distribution

$$\psi_i \sim MVN(\mu_{\mathrm{I}}, \Sigma_{\mathrm{I}}), \tag{6}$$

where

$$\mu_{\mathrm{I}} = (\mu_a, \mu_b, \mu_\alpha, \mu_\beta) \tag{7}$$

and covariance matrix

$$\Sigma_{\mathrm{I}} = \begin{pmatrix} \sigma_{\mathrm{a}}^2 & \sigma_{ab} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_{\mathrm{b}}^2 & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha}^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta\alpha} & \sigma_{\beta}^2 \end{pmatrix}. \tag{8}$$

In order to deal with the indeterminacy issue and identify the model, constraints need to be put in place. The mean speed, μ_τ, is set to zero. This constraint allows the

average item parameter, μ_β, to be, "equated to the average expected logtime over persons and items" (van der Linden, 2006, p. 185) and τ_j to be interpreted as the deviation from the average (van der Linden, 2006). In addition to the constraint that $\mu_\tau = 0$ on the response time side of the model, identifiability is obtained by also setting $\mu_\theta = 0$ and $\sigma_\theta^2 = 1$. These two constraints are typical in IRT parameter estimation.

This model can be illustrated as shown in Figure 4.

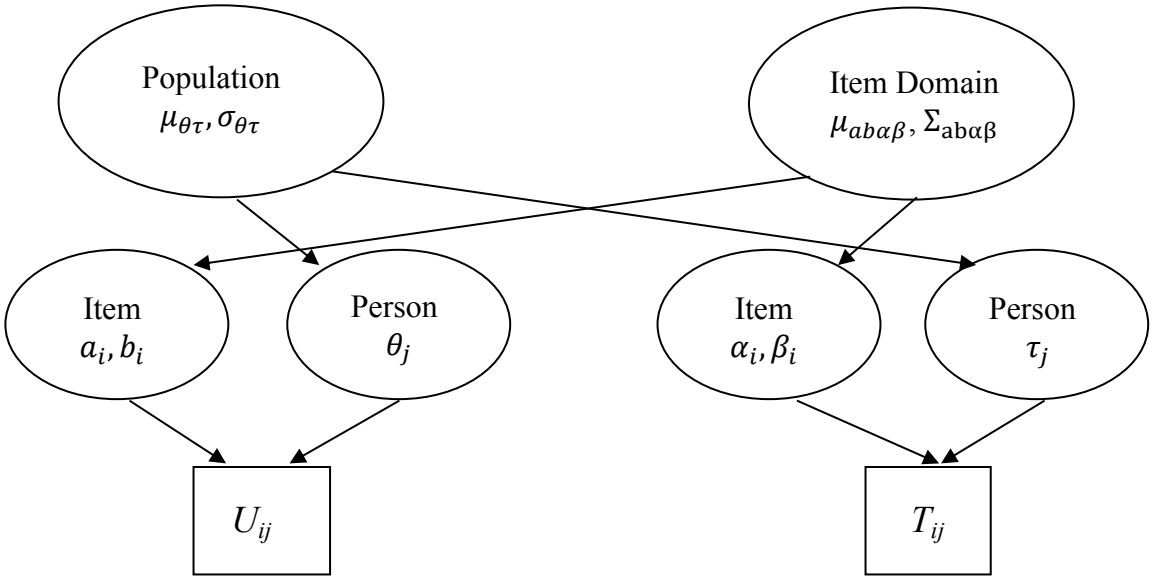

Figure 4. Hierarchical item response and response time model.

Research on the hierarchical model has highlighted some of its benefits such as the improvement in the estimation of examinee ability, detection of differential speededness and aberrant test behavior, and utility in adaptive test design (Fox et al., 2007; van der Linden, 2007; van der Linden et al., 2010).

Assumptions. There are several assumptions that need to be met for the model to hold. First, the stationarity assumption, which states an examinee is operating at a fixed speed. That is, examinees are assumed to settle on speed at the start of the test and stay at that speed for the remainder of the test. As a result of this assumption, this model

excludes changes in examinee speed as a result of learning, fatigue or test-taking strategy (Fox, et al., 2007). However, the results are not less useful for examinees that have minor fluctuations in speed because these fluctuations can be detected using residual analysis (van der Linden et al., 2007). Second, in addition to response time as a random variable, response is also considered to be a random variable. Third, separate item and person parameters are assumed for both the distributions of the responses and times. This assumption allows for the comparison of examinee speed across different items, which could be very useful in adaptive testing where one is trying to control the level of speededness of a test (van der Linden et al., 2007; van der Linden et al., 1999). Fourth, the model assumes conditional independence between the responses and response times for the level of ability and speed at which the examinee operates. While this assumption may seem counterintuitive, because responses and response time are nested within the combination of examinees and items, it follows the same line of reasoning as local independence for IRT models:

> For a fixed item, if a response model fits and the same holds for a response-time model, their person parameters capture all person effects on the response and response-time distributions. If these parameters are held constant, no potential sources of covariation are left and the response and the response time on an item become independent (van der Linden, 2007, p. 292).

Finally, the relationship between speed and ability for a population of examinees is modeled separately from the impact of these parameters on the responses and times of individual examinees. This is also done for the relationship between time and response parameters for the item.

Estimation. Van der Linden described a Markov Chain Monte Carlo (MCMC) method of estimation for the hierarchical model. The MCMC method is a Bayesian approach and as such priors are assumed for all parameters so that the parameters can be estimated using MCMC.

As priors for the population and item models, normal-inverse-Wishart prior distributions were selected. That is,

$$\Sigma_{\mathrm{p}} \sim Inv - Wishart(\Sigma_{p0}^{-1}, v_{p0})$$

$$\mu_p|\Sigma_p \sim MVN\left(\mu_{p0}, \frac{\Sigma_{\mathrm{p}}}{\kappa_{p0}}\right)$$

$$\Sigma_{\mathrm{I}} \sim Inv - Wishart(\Sigma_{I0}^{-1}, v_{I0})$$

$$\mu_I|\Sigma_{\mathrm{I}} \sim MVN\left(\mu_{I0}, \frac{\Sigma_{\mathrm{I}}}{\kappa_{I0}}\right),$$

where $v_{p0} \geq 2$ is a scalar degrees of freedom parameter, Σ_{p0} is a $2\ x\ 2$ scale matrix for the prior on Σ_{p}, and μ_{p0} and κ_{p0} are the vector with the means of the posterior distribution and the strength of prior information about the means. The parameters for the prior distribution Σ_{I} and μ_I are defined similarly.

The joint posterior distribution of the parameters is given by

$$f(\xi, \psi, \mu_p, \mu_I, \Sigma_{\mathrm{p}}, \Sigma_{\mathrm{I}}|u, t)$$

$$\propto \prod_{j=1}^{J}\prod_{i=1}^{I} f(u_{ij}; \theta_j, a_i, b_i) f(t_{ij}; \tau_j, \alpha_i, \beta_i)$$

$$\times f(\xi_j; \mu_p, \Sigma_{\mathrm{p}}) f(\psi_i;\ \mu_I, \Sigma_{\mathrm{I}}) f(\mu_p, \Sigma_{\mathrm{p}}) f(\mu_I, \Sigma_{\mathrm{I}})$$

The Gibbs sampler will be used to estimate the parameters. The Gibbs sampler iterates through draws from the full conditional distribution one block of parameters given all the remaining parameters. The conditional distribution of the blocks of

parameters can be derived from the joint posterior distribution equation above. Van der Linden (2007) noted that the parameter estimates tended to converge quickly using the hierarchical model.

Structural Equation Approach to Hierarchical Response Time Model

Finger and Chuah (2009) presented a confirmatory factor analysis (CFA) conceptualization of van der Linden's response time model that was estimated with maximum likelihood. In this model, the log of response time, ln_{t_i}, follows a single-factor model with intercept as follows:

$$ln\, t_i = v_i + \lambda_i \xi + \varepsilon_i$$

where v_i is the latent intercept for item *i*, λ_i is the loading factor for item *i* on the single factor, ξ is the level or score on the single factor, and ε_i is the residual.

This model assumes that the factor and residual have an expected value of zero and are independent of one another, E(ξ) = 0, E(ε_i) = 0, and E($\xi\varepsilon_i$) = 0. In addition, factor scores are assumed to have a variance of 1 and the factor loading for each of *n* items on the single factor is fixed to -1 (i.e., $\lambda_i = -1$, *i* = 1,…, *n*). Based on this parameterization, the mean and covariance structures are as follows

$$\mu' = (\mu_1, \mu_2, \ldots, \mu_n)'$$

$$= [E(\ln t_1), E(\ln t_2), \ldots, E(\ln t_n)]$$

$$= (v_1, v_2, \ldots, v_n)',$$

$$\Sigma = \lambda\lambda' + \psi^2 = (-1_n)(-1'_n) + \psi^2,$$

where $\lambda' = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$, $1'_n$ is a row vector of size *n* with all elements equal to -1, and ψ^2 is an *n* x *n* diagonal matrix of residual variances with the *i*-th diagonal element equal to ${\psi_i}^2$.

When the log-response time is normally distributed, the density function for the model is given by

$$f(\ln t_i | v_i,\ \lambda_i = -1, \xi, {\psi_i}^2) = \frac{1}{\sqrt{2\pi{\psi_i}^2}} exp\left\{\frac{[\ln t_i - (v_i - \xi)]^2}{-2{\psi_i}^2}\right\}.$$

This model is the same as van der Linden's response time model, such that ξ and v from the CFA model are equivalent to τ (the person speed parameter) and β (the item time intensity parameter) from van der Linden's parameterization. In addition, ${\psi_i}^2 = \alpha^{-2}$, where α is the discrimination parameter from van der Linden's response time model.

Sen (2012) combined a two-parameter normal ogive model for item response with Finger and Chuah's (2009) CFA model for item response time. Her model is referred to as the SEM-IRTRT model, and it is illustrated in Figure 5.

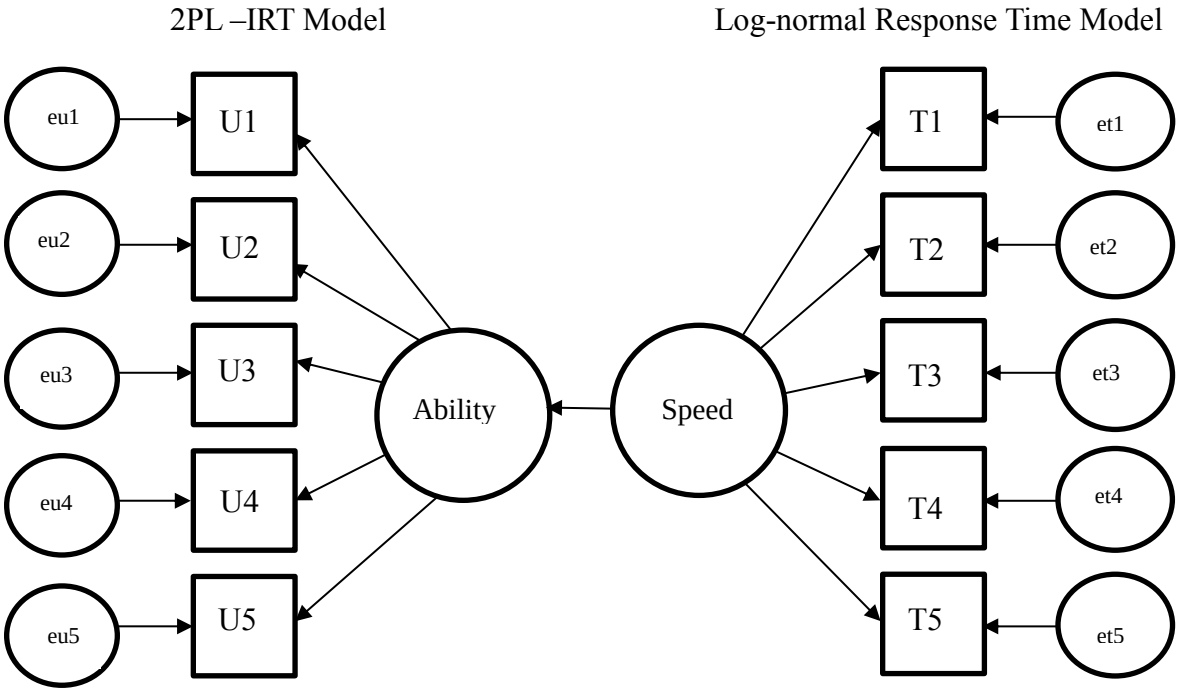

Figure 5. Illustration of the SEM-IRTRT model.

This model can be estimated using either Bayesian techniques or maximum likelihood estimation (Sen, 2012).

Sen (2012) compared parameter recoveries of the SEM-IRTRT model to van der Linden's hierarchical model (2007) as well as to a 2PL model utilizing marginal maximum likelihood estimation and a 2PL model utilizing MCMC estimation. She found both the SEM-IRTRT model and van der Linden's model improved ability estimation over the 2PL models when the correlation between person ability and speed was high, but not the item response parameters. In addition, she found that the SEM-IRTRT model performed comparably to van der Linden's model in estimation of item parameters. In some instances, such as estimation of time discrimination the SEM-IRTRT model performed better than van der Linden's model.

Parameter Invariance

One of the advantages of using IRT is that the item parameters are invariant across examinee populations and the person parameters are invariant across sets of items. Parameter invariance indicates that item parameters do not depend on the examinee population and person parameters do not depend on the distribution of items on the test (Lord, 1980). It is a property that only holds for the population parameters or when the model fits the data. Parameter estimates should be tested for a LOI because invariance is not guaranteed to hold for a sample (Rupp & Zumbo, 2006).

A LOI is a concern in educational and psychological testing for two main reasons. First, parameter invariance is necessary if one wants to generalize across examinee populations or measurement conditions. Rupp and Zumbo (2006) state that for inferences:

> to be equally valid for different populations of examinees or different measurement conditions, parameters in the psychometric models used for data analysis need to be invariant; if parameters are not invariant, the statistical foundation for inferences is not identical across the populations or measurement conditions, and hence the inferences are not generalizable across those to the same degree (p. 64).

When parameter invariance does not hold, we are only able to talk about that particular group of examinees tested on that particular set of items, which greatly limits the utility of the testing situation.

Second, a LOI may adversely influence scaling and equating procedures and result in biased scores (Hu et al., 2008; Sukin, 2010; Wells et al., 2002). That is, if parameter estimates that exhibit a LOI from one testing occasion are put on the same scale as parameter estimates from another testing occasion, the scores that are now thought to be comparable are actually not , which results in misleading conclusions about examinee performance.

As implied by the previous two reasons, LOI affects the inferences we make about test scores. Test scores are affected by something not accounted for by our model. The influence of this secondary construct on test scores is a source of construct irrelevant variance and bias because it "yields scores or promotes score interpretations that result in different meanings for members of different groups" (Gierl, 2005, p. 3). LOI raises question about test fairness and equity in the outcomes of testing. Therefore, detecting the presence of LOI is an important step in establishing valid score inferences and making fair and equitable decisions about examinees.

Research on Testing Invariance

A LOI among groups taking the same test such as different race or gender groups, is referred to as *differential item functioning* (DIF; Dorans & Holland, 1993), and a LOI among different test administration groups, such as groups taking the same test in different years, is called *item parameter drift* (drift; Goldstein, 1983). Most of the extant research has focused on DIF and drift in item response models, while very little attention has been given to LOI in response time models. The research summarized below focuses and DIF and drift because of its prevalence in the literature, but it also addresses the available research on LOI in response time models.

Differential item functioning. An item that requires different item parameters for distinct groups of examinees exhibits DIF. It is illustrated with a separate item characteristic curve (ICC) for each group and it indicates that examinees of the same ability have different probabilities of answering an item correctly (Dorans & Holland, 1993). DIF is a threat to test validity because it implies group specific inferences about test scores. In fact, the presence of DIF indicates that a measure taps into a nuisance dimension (Ackerman, 1992). Thus, the presence of DIF raises concerns relating to fairness and equity in testing (Samuelson, 2005).

When examining DIF, the performance of two groups is compared. The primary group of interest is referred to as the focal group, which usually consists of examinees in the minority such as Black, Hispanic, Asian, Females, or English language learners. The other group is referred to as the reference group and it typically consists of a majority group such as Whites, Males, or those who speak English as a first language (Holland & Wainer, 1993). A DIF analysis entails the comparison of the focal group performance to

reference group performance, but the comparison is not simply the group difference in mean item performance (Wainer, 1993; Zwick, 1990). This comparison is referred to as impact. DIF refers to the group difference in mean item performance conditional on examinee ability (Wainer, 1993; Zwick, 1990). Impact can be written as

$$P(X_i = 1|G = F) \neq P(X_i = 1|G = R)$$

where X, is the response to Item i, X=1 indicates a correct response and F represents the focal group and R represents the reference group. It contrasts with DIF which involves the added dimension of comparison of item performance, conditional on ability (Zwick, 1990),

$$P(X_i = 1|T = t, G = F) \neq P(X_i = 1|T = t, G = R) \text{ for all } t,$$

where T represents the ability on which item responses depend.

DIF can manifest as uniform DIF, non-uniform DIF or both. According to Mellenbergh (1982) uniform DIF occurs when one group is consistently advantaged across the entire ability scale and the ICCs are parallel. Uniform DIF exists in the difficulty parameter only. In contrast to uniform DIF, non-uniform DIF occurs such that one group is advantaged at one part of the scale whereas the other group is advantaged at another part of the scale. This would be seen by intersecting ICCs and indicates that there is DIF in the discrimination parameter (Swaminathan & Rogers, 1990). An example of non-uniform DIF would be that for examinees who score at or below the mean (e.g. $\theta \leq 0$), the Focal group is favored whereas for those scoring above the mean (e.g. $\theta > 0$) the Reference group is favored. Non-uniform DIF is often more difficult to detect as some methods do not account for the canceling out effect that occurs between positive and negative regions of area.

Research on DIF. DIF analysis can be conducted for any known groups of examinees such as race, gender, language, or disability groups. A large portion of DIF research has compared racial and ethnic groups (see Barnes & Wells, 2009; Bleistein & Wright, 1987; Hauser & Kingsbury, 2004; Kulick, 1984; Kulick & Dorans, 1983; Lord, 1977; Rogers, Dorans, & Schmitt, 1986; Schmitt, 1985, 1988; Schmitt & Bleistein, 1987; Schmitt & Dorans, 1988) and gender groups (see Barnes & Wells, 2009; Gierl, Khaliq, & Boughton, 1999; Hauser & Kingsbury, 2004; Lawrence, Curley, & McHale, 1988). DIF research also exists on the comparison of different language groups (see Alderman & Holland, 1981; Angoff & Sharon, 1974; Hulin, Drasgow & Parsons, 1983; Sinharay, Dorans, & Liang, 2009; Snetzler & Qualls, 2000), and students with and without disabilities (Abedi, Leon, & Kao, 2008) and students with and without test accommodations (see Bolt, 2004; Cohen, Gregg, & Deng, 2005; Koretz, 1997; Koretz & Hamilton, 1999). The preponderance of DIF research testifies to the importance of ensuring test fairness and eliminating sources of construct irrelevant variance.

DIF analyses are utilized in high stakes tests such as the SAT, TOEFL, GRE, and certification exams (see Alderman & Holland, 1981; Gu, Drake, & Wolfe, 2006; Lawrence, Curley, & McHale, 1988; Woo & Dragan, 2012), large scale state assessments (see Bolt, 2004; Koretz, 1997; Koretz & Hamilton, 1999), low stakes tests (Barnes & Wells, 2009), psychological assessments (see Abad, Colom, Rebollo, & Escorial, 2004; Mitchelson, Wicher, LeBreton, & Craig, 2009; Sheppard, Han, Colarelli, Dai, & King, 2006), and medical assessments (see Sacco, Casado, & Unick, 2011; Woodard, Auchus, Godsall, & Green, 1998). DIF detection is fairly straightforward. Understanding the cause of DIF is not, but there are a few common explanations that arise in DIF studies.

After researchers discover the presence of DIF, they often want an explanation as to why it occurs. One of the common explanations of DIF is an examinee's familiarity with the content of the item (Eells, Davis, Havighurst, Herrick & Tyler, 1951; Jensen, 1980; Reynolds & Kaiser, 1990; Striker & Emmerich, 1999). Another explanation is an examinee's interest in the content of the item (Eells et al., 1951). A third explanation is negative emotional reaction associated with the content (Wendler & Carlton, 1987). Another explanation is differential speededness. That is, some students tend to respond to items more slowly than others students which may account for the presence of DIF (Schmitt & Bleistein, 1987; Schmitt & Dorans, 1990).

Item parameterdrift. Another occurrence of LOI is item parameter drift referred to simply as drift. Drift occurs when there is a, "differential change of item parameters over subsequent testing occasions" (Wells et al., 2002, p. 77). Instead of treating the groups of interest as two groups within a single testing occasion as you do in DIF, the groups for drift analyses are examinees from two different testing occasions. The testing occasions can represent different forms given on a single occasion with a common set of items on or the same form given over multiple occasions.

Research on drift. Research on drift is much smaller than the body of work on DIF. Research has shown that drift over 1 year is not as problematic as drift over a longer period of time and more testing occasions. Wells et al. (2002) found that when item discrimination and item difficulty parameters were simulated to have drift over the course of one year, there was a small impact on ability estimates. Similarly, Rupp and Zumbo (2003a, 2003b) found that examinees' scores were only slightly impacted, unless

there was an unusually large amount of drift simulated; however, these studies only focused on the impact of drift across two occasions.

It has been shown that drift may be more problematic when studied longitudinally. DeMars (2004a, 2004b) examined patterns of drift over four years on one test of U.S. History and Political Science and a second test of information literacy. She showed that while the impact of drift on item parameters may be small for one year, over the course of four years the impact could be very large (DeMars, 2004a, 2004b). Wollack, Sung, and Kang (2005) examined drift over seven years on a college-level placement test. They found that drift effects did not seem to cumulate over time. A possible explanation of this was it was random drift (e.g., an item increasing in difficulty between occasion 1 and 2, and then decreasing in difficulty between occasion 2 and 3) and not systematic drift (e.g., an item becoming increasingly harder over multiple testing occasions) as well as having both positively and negatively drifting items that cancel each other out. In addition, the choice of linking method was found to play a role in the ability estimates, with the Haebara and Stocking Lord methods[1] performing better than the mean-mean and mean-sigma methods[2] (Wollack et al., 2005). Wollack, Sung, and Kang (2006) extended their research on the effects of compounding drift on examinee ability estimates. They found that for a test that included both random and systematic drift over several years could lead to substantial bias on ability estimates. From these studies it appears that systematic drift is more problematic than random drift for a testing program that spans more than one year.

---

[1] These are iterative methods that rely on characteristic curves produced during an IRT analysis to put different test versions on the same scale

[2] These methods rely simply on the item estimates produced during an IRT analysis to put different test versions on the same scale.

Item parameters may change due to the use of different test forms for multiple reasons such as: a shift in curriculum and instructional emphasis (Bock, Muraki, & Pfeiffenberger,1988; DeMars, 2004b; Han & Guo, 2011; Sykes & Fitzpatrick, 1992), disclosure of the items by previous test takers and practice (Guo, 2009; Han & Guo, 2011), changes in the construct over time (Babcock & Albano, 2012; Martineau, 2004; 2006), or changes in placement of item on the test (Kingston & Dorans, 1984). Regardless of the cause of drift, it poses a threat to assessment procedures that require a stable scale.

Methods for Detecting Lack of Invariance in Item Response Models

Both DIF and drift involve pairwise comparisons and as such any DIF procedure could be utilized to study drift (DeMars, 2004a). There are multiple procedures that can be utilized to detect DIF or drift. These procedures will be referred to as LOI detection procedures. Some methods involve observed data and others involve IRT parameter estimates.

Observed data procedures. Procedures to detect LOI that are based on observed data generally utilize the overall test score for conducting analyses. Two of the most common procedures are the Mantel-Haenszel procedure and the SIBTEST.

Mantel-Haenszel. The Mantel-Haenszel procedure (Holland & Thayer, 1988) is one of the most widely utilized and well known procedures for conducting LOI analysis (Clauser & Mazor, 1998). It can only be used with categorical data as it utilizes a three-way contingency table that tabulates item responses for two different groups at multiple test score levels. The Mantel-Haenszel approach tests conditional independence of two categorical variables – group membership (focal or reference) and item response – at

each stratum. The Mantel-Haenszel statistic is distributed approximately as a chi-square with one degree of freedom and is used as a measure of statistical significance. The common-odds ratio is used to as a measure of effect size for binary items, and the standardized mean difference is often used as a measure of effect size for polytomous items (Zwick, 1993). These effect size measures indicate the magnitude of LOI and reflect the degree of practical significance.

SIBTEST. The simultaneous item bias test (SIBTEST) is a nonparametric significance test for detecting LOI. Instead of the observed test score that is used to match examinees in the Mantel-Haenszel procedure, SIBTEST involves a regression correction method to match examinees on their latent ability level. This correction controls the type I error rates. A benefit of SIBTEST is that it generalizes to multiple dimensions (Shealy & Stout, 1993), unlike other LOI detection methods.

SIBTEST requires two distinct non-overlapping subsets of items in a test. One subset, referred to as the valid subtest, which contains the items that are assumed to only measure the construct that the test is designed to measure. The other subset, referred to as the suspect subtest, which contains the items being tested for a LOI. The scores from the valid subtest are used to match examinees that have the same ability level across groups in order to test items from the suspect subtest for LOI. After examinees are matched, the item means are adjusted to correct for differences in the ability distributions for the focal and reference groups using a regression correction. These corrected estimates are the basis for the test statistic for examining LOI.

IRT-based methods. There are many procedures for detecting LOI that utilize the item and person parameter estimates from IRT models. Some methods directly

compare parameters for two different groups and other methods compute the area between ICCs.

Lord's χ^2. Lord's chi-square (Lord, 1980) tests the hypothesis that the item parameters in the reference group are equal to those in the focal group. Lord's chi-square simultaneously tests the null hypothesis, Ho: $a_{iR} = a_{iF}$ and $b_{iR} = b_{iF}$ for item i, where a_{iR} and b_{iR} are the discrimination and difficulty parameters for item i estimated for the reference group and a_{iF} and b_{iF} are the discrimination and difficulty parameters for item i estimated for the focal group. This method for detecting LOI utilizes matrix information and can be conducted with categorical data.

Ragju's signed and unsigned area between two ICCs. Raju (1988) presented another method for detecting items that LOI. This method utilizes the area between two ICCs, where large areas indicate LOI, and it can be used only with the Rasch, 2PL and 3PL with fixed c parameter (Raju, 1990). This method is not appropriate for the 3PL model where the c parameter is allowed to vary. He also points out that asymptotic formulas tend to work best when sample sizes are large and cautions using the procedures with small sample sizes.

Kim and Cohen's signed and unsigned closed-interval measure. Like Raju's signed and unsigned area between IRFs, Kim and Cohen's signed and unsigned closed-interval measures utilize the area between ICCs. However, they differ in that the closed-interval imposes limits on the interval of interest on the θ scale (Kim & Cohen, 1991). The closed interval area measures are used to determine if the area between the ICCs is larger than 0.

Robust z. This method for detecting LOI is based on robust statistical procedures (see Hogg, 1979; Huber, 1964; Huynh, 1982). It relies on the robustification of the traditional z statistic. Let *D* be the difference between a person's score on two variables and $\overline{D}$ be the mean of the difference, and *SD* be the standard deviation of the differences. The traditional *z* statistic is defined as $z = (D - \overline{D})/SD$. Given that the mean and standard deviation are influenced by outlying observations the z statistic is also influence by outliers. Therefore, a robust z is obtained by replacing the mean with the median and the standard deviation by the IQR (Huynh & Meyer, 2009). For testing LOI, *D* represents the difference in item parameters for a single item on two different test occasions and $\overline{D}$ is the mean of the item parameter differences. LOI is identified by comparing an item's robust *z* statistics to a critical value form the standard normal distribution.

An advantage that the robust *z* method has over other methods for detecting LOI is it can test for LOI in each item parameter (e.g. difficulty, discrimination) separately. Other IRT-based procedures test for LOI in item parameters simultaneously.

Multiple indicator multiple cause model. The multiple indicator multiple cause model (MIMIC), popularized by Muthén (1985), utilizes a confirmatory factor analysis (CFA) framework to examine LOI. The model for testing uniform LOI relies on the following equation:

$$y_i^* = \alpha_i\theta + \phi_i z + \varepsilon_i,$$

where y_i^* is the latent response variable for item *I* (where $y_i^* > \tau_i$, an observed variable,y_i=1; τ_i is referred to as the threshold parameter and is related to item difficulty), θ is the latent trait, α_i is discrimination parameter for item *i*, z is a dummy variable indicating group membership, ϕ_i indicates the relationship between the grouping

variable and the item response (i.e., the group difference in the threshold), and ε_i is the random error. Uniform LOI is evaluated by examining the significance of ϕ_i, which tells us the group differences in the threshold parameter.

The model for testing non-uniform LOI relies on the interaction between group membership and the latent trait and can be represented by the following equation

$$y_i^* = \alpha_i\theta + \phi_i z + \omega_i\theta z + \varepsilon_i$$

where ω_i indicates the non-uniform LOI effects (Woods & Grimm, 2011). This model is illustrated using a confirmatory factor analysis approach in Figure 6.

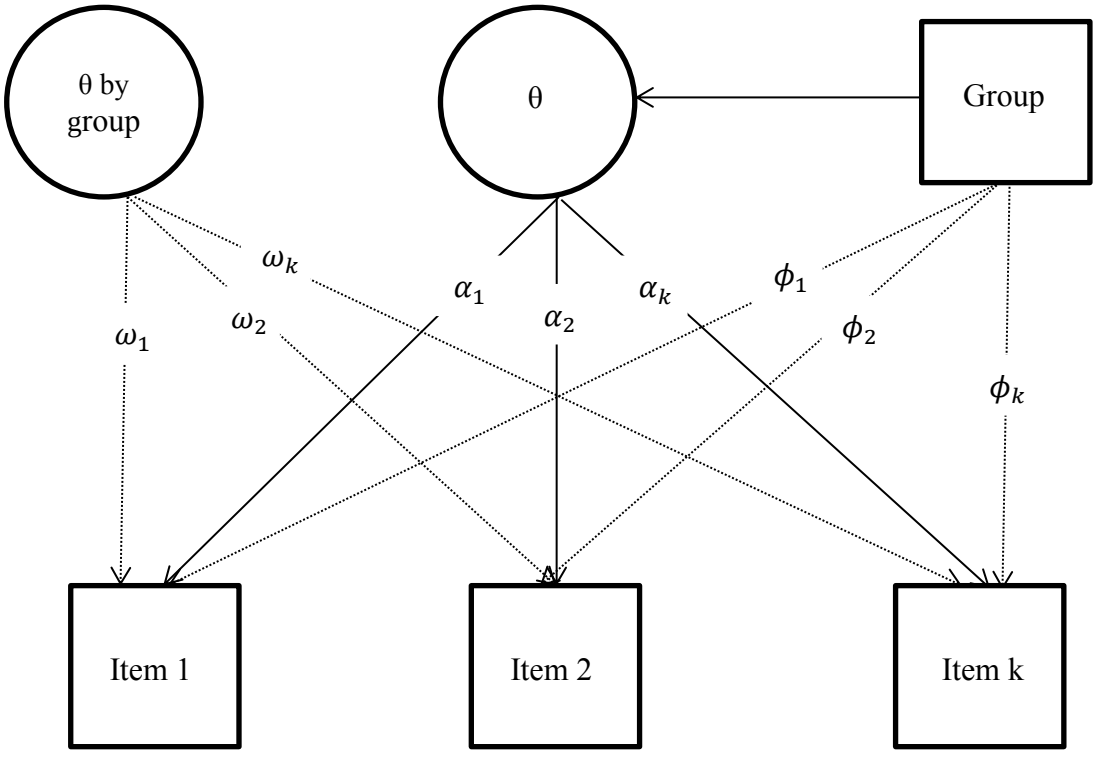

Figure 6. A MIMIC-interaction model for testing non-uniform and uniform LOI.

There are multiple approaches for testing for LOI with the MIMIC model, but no consensus on which approach is best. One approach starts with the baseline model that has no paths from group variable to item and uses large modification indices to identify

which path may exhibit LOI (Muthén, 1988 & 1989; Oort, 1992 & 1998). The main limitation with this approach is that there is not a definition of what makes a large index. In addition, statistical significance of modification indices is impacted by sample size (Woods, 2009b). Therefore, using modification indices to identify LOI may not be a preferable approach.

Several alternative approaches utilize a subset of anchor items that are invariant and used to define the matching criterion. Anchors are defined in the model by constraining the path from the group variable to the item to zero. In one variation of the anchor item approach, all but the studied item are used as anchorswhen evaluating LOI. This method performs well when there is no LOI in any of the anchor items (Cohen, Kim & Wollack, 1996; Kim & Cohen, 1998). However, if LOI is present among the anchor items, the the result is misspecification of the null distribution, inaccurate estimate of parameters, and inflation of the type I error rate (Stark, Chernyshenko, & Drasgow, 2006; Wang, 2004; Wang & Yeh, 2003; Woods, 2009a).

A second variation of the anchor item approach, and perhaps the most popular one, utilizes designated anchors (Christensen, Jorm, Mackinnon, Korten, Jacomb, Henderson, et al., 1999; Fleishman, Spector, & Altman, 2002; Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000; Woods, 2009a). This approach first identifies LOI-free items and utilizes them as anchor items. Once anchor items are established, then the studied items are tested simultaneously for LOI. Woods (2009a) showed that setting LOI-free anchor items to between 10% and 20% of the total number of items lead to higher power rates than utilizing a single LOI-free item and fewer type I errors than utilizing all other items as anchors.

Research comparing methods. Studies of LOI detection methods typically evaluate power and type I error rates under a variety of conditions. Some popular conditions to examine include: test length, sample size, number of LOI items, magnitude of LOI, and difference in ability distributions. These conditions are prevalent in studies of LOI detection as they have been shown to impact power and type I error rates (see Finch, 2005; Meyer & Huynh, 2010, Woods, 2009b). To compare the various methods for detecting LOI, research regarding power and type I error will be presented.

Power. Many studies have examined the power of various LOI detection methods. Lord's chi square performed better than Raju's interval methods for a 2PL model especially when the sample size was small ($N = 100$), the test was short ($k = 20$), and the percentage of LOI items was high (i.e. 20%; Cohen & Kim, 1993). Donoghue and Isham (1998) also found that Lord's chi square had the highest power when compared to Raju's interval methods, Kim and Cohen's closed-interval methods, and the Mantel-Haenszel procedure for a 3PL model, but this finding only applied when the pseudo-guessing parameter was fixed to a constant value. It had higher levels of power when detecting LOI in the difficulty parameter and LOI in both the difficulty and discrimination parameters (Donoghue & Isham, 1998; Green, Smith, & Habing, 2010), but had a harder time detecting LOI in only the discrimination parameter (Donoghue & Isham, 1998). .

The Mantel-Haenszel procedure had higher power than Lord's chi square, Raju's interval methods, Kim and Cohen's interval methods, MIMIC model, and robust z ($N = 1,000$) for detecting LOI in the difficulty parameter (Donoghue & Isham, 1998; Finch, 2005; Green et al., 2010; Meyer & Huynh, 2010) and LOI in the difficulty and

discrimination parameters (Donoghue & Isham, 1998; Green et al., 2010). Research has shown that when sample size increases, the Mantel-Haenszel's power increases (Clauser, Mazor, & Hambleton, 1994; Swaminathan & Rogers, 1990), but so too does the power for other methods. Indeed, when the sample size was large (N = 4,000), the Mantel-Haenszel and the robust z had comparable levels of power (Meyer & Huynh, 2010). .

SIBTEST had comparable power rates to the Mantel-Haenszel method (Shealy & Stout, 1993), but higher than the Mantel-Haenszel when the items favored the focal group (Gotzmann, Wright, & Rodden (2006). That is, when the LOI was in the positive direction, SIBTEST performed better.

Raju and Kim and Cohen's signed interval methods, had good power rates for LOI in the difficulty only and both difficulty and discrimination parameters; however, it showed extremely low levels of power for detecting LOI in the discrimination parameters. The power rates were almost as small as the nominal alpha levels for detecting discrimination only. The unsigned interval methods had lower power rates for detecting difficulty LOI only and higher power rates for detecting discrimination LOI than its signed counterpart (Donoghue & Isham, 1998).

The MIMIC model was shown to have high power rates when the test was long (k = 50) or the model contained no pseudo-guessing parameter (Finch, 2005). When there was no pseudo-guessing parameter present, the MIMIC model performed better than the SIBTEST and Mantel-Haenszel procedures. In addition, the difference in mean ability distribution did not have a large impact on power rates (Finch, 2005).

For Lord's chi square, Mantel-Haenszel, and robust z, Green et al. (2010) found that when there were fewer LOI items or larger magnitude of LOI, power rates were

better. Their finding suggests that a purification method may work best. This procedure involves a multistage process in which a preliminary analysis is run to flag items that are exhibiting LOI. The second stage involves re-running the analysis without the flagged items (Marco, 1977).

Type I error. Lord's chi square has been shown to have error rates 7 to 10 times larger than the nominal level, when the guessing parameter is freely estimated (Donoghue & Isham, 1998; Meyer & Huynh, 2010). When the pseudo-guessing parameter is freely estimated, error rates increase as the sample size and number of common items increase (Meyer & Huynh, 2010). When the pseudo-guessing parameter is fixed, error rates are at or below the nominal level (Green et al., 2010; Kim, Cohen, & Kim, 1994). For 2PL models, Cohen and Kim (1993) found that when groups differed in ability distributions, higher error rates were observed.

The method of estimating item response model parameters can have an impact on Lord's chi-square test. McLaughlin and Drasgow (1987) found that joint maximum likelihood estimation (JMLE) lead to highly inflated type I error rates. However, marginal maximum likelihood estimation (MMLE) and marginal Bayesian estimation (MBE) for the 2PL parameters resulted in type I error rates that were around the nominal alpha level (Cohen & Kim, 1993; Lim & Drasgow, 1990).

The Mantel-Haenszel procedure has been shown to have error rates at or below nominal levels for a variety of conditions (Donoghue & Isham, 1998; Finch, 2005; Green et al., 2010; Meyer & Huynh, 2010; Roussos & Stout, 1996). Error rates decreased as sample size increased (N = 1,000 & 4,000) (Meyer & Huynh, 2010) when using large sample sizes; however, for small samples sizes (N = 100, 200, 500, & 1,000), error rates

increased slightly (Rossous & Stout, 1996) as the sample size increased. Error rates for the Mantel-Haenszel procedure also increased as the difference between the group ability distributions increased (Rossous & Stout, 1996).

SIBTEST has been shown to have error rates comparable to the Mantel-Haenszel procedure (Shealy & Stout, 1993). Like the Mantel-Haenszel procedure, error rates increased as the sample size and the difference between groups ability distributions increased for small sample sizes (Rossous & Stout, 1996).

Research has shown robust *z* to be a worthwhile LOI detection approach in Rasch equating as it has type I error rates close to nominal levels (Arce-Ferrer, 2008; Huynh & Rawls, 2009). Green et al. (2010) examined the robust *z*'s utility for the three-parameter logistic model (3PL) model. They found that the robust *z* statistic had highly inflated type I error rates. Error rates increased slightly as the sample size increased and the percentage of LOI items decreased. Error rates decreased as the number of common items increased. Meyer and Huynh (2010) examined the robust *z* procedure for the 3PL and generalized partial credit model (GPCM) and showed that sample size and number of common items were related to the performance of the robust *z* statistic. Robust *z* performs at optimal levels when sample sizes are at least 3,000 examinees, common items make up about 40% of the total test. Power for the robust *z* can be increased by increasing sample size and the number of common items. Arce and Lau (2011) showed that the nominal level also played a role in the robust *z*'s performance. More conservative nominal levels such as .10 also lead to better performance of the robust *z*.

Donoghue and Isham (1998) found that the test for significance for Raju's signed area performed as expected when there was no LOI present; however, the significance

test for the unsigned area had type I error rates that were nearly twice the nominal rate. For 2PL models, Cohen and Kim (1991) found that when groups differed in ability distributions, higher error rates were observed. In contrast, Cohen and Kim (1991) showed that signed closed-interval procedure for the 3PL model produced error rates well below the nominal alpha level. The error rates for the signed closed-interval procedure were lower when the pseudo-guessing parameter was freely estimated, while the unsigned closed-interval procedure for the 3PL had slightly inflated error rates. In a study conducted by Donoghue and Isham (1998), both the signed and unsigned closed-interval procedures performed well with error rates at and under the nominal alpha levels.

Finch (2005) found that error rates for the MIMIC model were inflated when the test was short (e.g. 20 items) and the model underlying the data was 3PL, but at or below nominal levels when test length was increased or there was no pseudo-guessing parameter. In addition, he found that the difference in mean ability distribution did not have a large impact on power rates (Finch, 2005). When p-value corrections were implemented on data from a short test error rates were all found to be at or below nominal levels (Woods, 2009). Woods and Grimm (2011) found that for detecting non-uniform LOI, the MIMIC-interaction model had an inflated type I error rate and suggested using a Benjamini-Hochberg (1995) p-value correction to account for this inflation.

Research on Lack of Invariance in Response Time Models

LOI has been studied extensively for item response models, but substantially fewer studies evaluate LOI in response time models. LOI is similar to other concepts that researchers have explored in response time models but these concepts vary in definition

and may not strictly be considered LOI. The different definitions vary in their matching criterion as well as the parameters of interest.

Differential speededness is refers to the situation where examinees, who are matched on ability, require different amounts of time to respond to an item (van der Linden et al., 2007). This concept is studied in the context of computer adaptive testing, where the items an examinee receives depends on his/her response to items of different difficulty. The concern is that examinees that get more difficult items may require more time to answer the items, and therefore not have as much time to answer items at the end of test. Bridgeman and Cline (2004) found that when examinees who took the GRE analytic and quantitative sections were matched on ability level near the end of the test those who had a less time-consuming test (i.e., the items required less time) had higher scores by about 25 points than those who had more time-consuming tests. This implies that examinees with a more time-consuming test tend to respond incorrectly to end of test items. A possible explanation is that the examinees run out of time and have to rush.

DeMars and Wise (2010) also matched examinee on ability to compare examinees on response time performance. They utilized the term *differential rapid-guessing* to refer to different sub-groups, who are matched on ability, exhibiting differences in response time for a given item. Their definition implies a particular type of examinee behavior (i.e. rapid-guessing) and not just the effect of a time limit. They examined lack of differential rapid-guessing in the context of a computer-based low-stakes test. They wanted to determine if differential rapid-guessing can lead to LOI in the item responses and be detectable using standard identification methods such as the Mantel-Haenzsel procedure. They found that differential rapid-guessing can lead to LOI favoring the

group with longer response times. Their results suggest an additional explanation for LOI; specifically, it can be caused by differences in examinee behavior (e.g. rapid-guessing).

Klein Entink (2009) used a different criterion for identifying comparable examinees. He matched examinee on response speed, not latent ability, to define the concept of *time-DIF*. This phenomenon refers to the situation in which groups of test takers, who are matched on response speed, differ in their response time for a given item. In addition to coining the term time-DIF, he proposed a method for testing for it using a dependent samples t-test. In particular, he suggested that the null hypothesis for assessing time-DIF between group 1 and group 2 is $\beta_i^1 - \beta_i^2 = 0$ and the alternative hypothesis is that $\beta_i^1 - \beta_i^2 \neq 0$ where β_i equals the item time intensity parameter. Little research has been conducted with this method. For the purposes of this study, I will be utilizing Klein Entink's concept of time-DIF when I discuss a LOI in the response times and I will employ the term time-LOI.

Construct irrelevant variance is a threat to validity (Haladyna & Downing, 2004). LOI among item responses suggests that something other than the target construct is affecting scores. In a similar fashion, time-LOI suggests that something other than the intended construct is influencing an examinees performance. It could be that time limits cause examinees to switch their test taking strategies and rapidly respond to items near the end of the test. It could also suggest that the cognitive demand for a test item differs for groups of examinees. For example, word problems on a math test may take exceptionally more time to complete for the students with low reading speed. The extra demand of reading may change the time some examinee need to complete the item and

also possibly the quality of the item response. Thus, time-LOI also suggest that something other than the intended construct is affecting test performance resulting on construct irrelevant variance.

Detection of time-LOI is important throughout the testing process. If an item is flagged as exhibiting time-LOI in a piloting stage of an exam, test developers will have an opportunity to revise the item prior to operationalization. For items exhibiting time-LOI on an operational computer-adaptive test, this could indicate the need to refresh the item pool. If items exhibiting time-LOI go undetected in a criterion-referenced testing scenario, this could have consequences for the decisions made regarding examinee proficiency.

A variety of procedures are available for detecting LOI and researchers have proposed method for evaluating time-LOI. However, to avoid the adverse influence of LOI and time-LOI on test scores, a method is needed to simultaneously detect both types is needed.

A Multiple Indicator Multiple Cause Model for detecting LOI for Item Responses and Response Times

The proposed model combines Sen's (2012) SEM formulation of the hierarchical model and the model. This model is referred to as the multiple indicator multiple cause model for detecting LOI for item response and response times (MIMIC-IRTRT) model. An illustration of the MIMIC-IRTRT is provided in Figure 7.

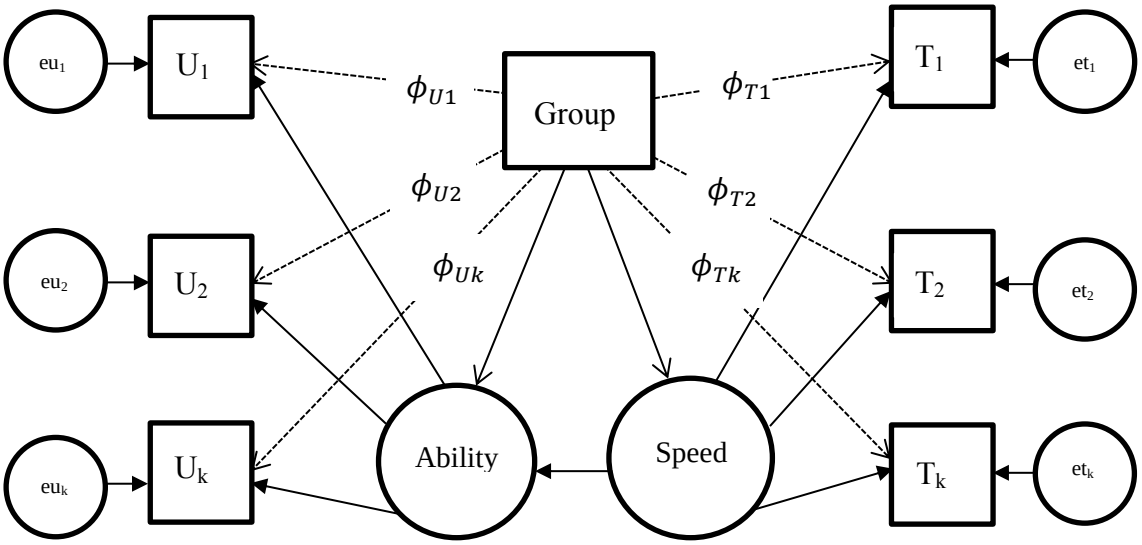

Figure 7. MIMIC-IRTRT model for item responses and response times.

In the MIMIC-IRTRT model, uniform LOI is determined by examining the LOI effects for item responses, ϕ_{Uk}, and response times, ϕ_{Tk}. There are model constraints put in place so that the model is identifiable and the issue of indeterminacy is resolved. The mean and standard deviation of person speed are constrained to 0 and 1, respectively. The factor loadings, paths from speed to response time, for speed are fixed to -1 and the factor loading for the first item response, path from ability to response, is fixed to 1.

In order to examine the LOI effects, a multiple step process must be under taken. First, anchor items need to be established. These help to identify the model and also serve to define the factor, ability and speed, on which the groups are matched (Woods, 2009b). Anchor items are selected from a LOI-free subset of items. In simulation studies it is common to select a percentage of items for anchor items, generally 10 to 20 percent, from the subset of items that were not manipulated to have LOI. For real data, the anchor items need to be tested empirically. There are different procedures to choose from to test the items, one of the most promising procedures is that described by Woods (2009). This

procedure involves running a preliminary MIMIC-IRTRT model with a single item being tested for LOI using all other items as anchors. The ratios of LOI effect to standard error are obtained for each item and then are ranked ordered by absolute value. The items with the smallest ratios are selected as the anchor items.

Once anchor items are established, testing for LOI on the studied item (i.e., non-anchor items) can begin. The underlying framework for LOI detection using the MIMIC-IRTRT model involves the estimation of direct and indirect effects for the grouping variable. The indirect effect is captured by regressing the latent variable onto the group variable and tells us whether there is a mean difference on the latent variable across the groups. This indirect effect accounts for the group differences on the latent variable. The direct effect, also referred to as the LOI effect, is captured by regressing the item response or response time for the studied items onto the group variable and tells us whether there is a difference in the response probabilities or response times across the groups. The direct effect indicates whether there is LOI, after controlling for mean differences in the latent variable for the groups. In order to get these indirect and direct effects, the MIMIC-IRTRT model is run with LOI effect paths in place for all the studied items for both item responses and response times and no LOI effect path for the anchor items. An item is designated as exhibiting LOI if the LOI effect is statistically significant.

This study will focus on examining uniform LOI for both item response (referred to as response-LOI) and response time (referred to as time-LOI).

Summary

Response time models are promising as they provide more information about items and examinees. One of the most popular response time models is van der Linden's (2007) hierarchical model, but research with this model in particular and response time models in general have been largely focused on improvements to item and parameter estimation. Research relating to a LOI in response time models is in its early stages. Moreover, research on the detecting a LOI for item response models are not appropriate for response time models because response time data is continuous and not categorical like item response data. Only one method, the dependent t-test, has been suggested as an option for detecting time-LOI (Klein Entink, 2009); however, this test only applies to response times and the viability of this method has not been assessed. A method that would allow for simultaneous evaluation of LOI for both item responses and response times has not been developed.

The current study will address the gaps in the literature in three ways. It will (1) propose a MIMIC model for detecting LOI for response times, (2) present a model for examining LOI for both item responses and response times, and (3) evaluate the viability of the proposed model. As such, this study will provide the most comprehensive model for evaluating LOI to date.

The next chapter provides a description of the methodology used to evaluate the utility of the MIMIC model for detecting LOI for both item responses and response times.

CHAPTER 3

METHODOLOGY

The evidence presented in the previous chapters illustrates the lack of tools for evaluating LOI in response time models and LOI in joint models of item response and response time. The proposed study will fills this gap in the literature by proposing a model for evaluating LOI in item response and response time and also by evaluating the viability of the proposed model through a simulation study. The model will then be applied to extant data to demonstrate its use in practical testing situations. In particular, this study will answer the following research questions:

1. What are the empirical power and type I error rates of the MIMIC-IRTRT model for of identifying uniform LOI?
2. How do sample size, type of LOI, magnitude of LOI and other factors affect the type I error rates and empirical power of the model?

Study 1: Simulation

Data Simulation. The simulation includes five fully-crossed conditions and null conditions. The fully-crossed conditions include the: (a) test length (20 or 40 items) (b) type of LOI (response-LOI, time-LOI, or both) (c) number of LOI items (3 or 6), (d) correlation of examinee ability and person speed (0, .4, or .8), and (e) magnitude of LOI (small or large). The null conditions represent cases with no LOI, and provide the basis for calculating type I error rates. There are a total of six null conditions, as the no LOI cases still vary by test length, which has two levels, and correlation between person speed

and ability, which has three levels. The sample size was fixed at 1,000 examinees per test form. There are 2 x 3 x 2 x 3 x 2 = 72 fully-crossed conditions plus six null conditions, for a total of 78 conditions. The MIMIC-IRTRT model was applied to each condition. Each condition was replicated 100 times.

Pairs of data were generated to represent two test forms. Generating item parameters were based on analysis of extant data from the computer-based certification exam[3] and were generated according to van der Linden's hierarchical model [see Equations 1-9]. For the item response model, the 2PL was utilized as M*plus*[4] (Muthén & Muthén, 1998-2011) cannot currently estimate 3PL models. The item parameters were sampled from a multivariate normal distribution with means as follows

$$\mu_P = (\mu_\theta, \mu_\tau) = (0, 0).$$

The correlation matrix was configured as follows

$$\Sigma_{\mathrm{P}} = \begin{pmatrix} 1 & \rho_{\theta\tau} \\ \rho_{\tau\theta} & 1 \end{pmatrix}$$

where $\rho_{\theta\tau}$ represents the correlation between person ability and speed. The level of correlation is manipulated in the simulation.

The item parameters were sampled from a multivariate normal distribution with means and as follows

$$\mu_I = \left(\mu_a, \mu_b, \mu_\alpha, \mu_\beta\right) = (0.616, -0.689, 1.016, 4.133)$$

The covariance matrix was configured as follows

[3] This data does not contain a grouping variable. Therefore, it cannot be analyzed for LOI.
[4] M*plus* will be utilized for the data analysis

$$\Sigma_I = \begin{pmatrix} a & b & \alpha & \beta \\ 0.06 & 0.01 & 0.04 & 0.02 \\ 0.01 & 0.55 & 0.02 & 0.17 \\ 0.04 & 0.02 & 0.05 & -0.01 \\ 0.02 & 0.17 & -0.01 & 0.78 \end{pmatrix}$$

Since the model that is utilized to simulate the data is not exactly the same as the model utilized for estimation, a check of parameter recovery will be conducted.

Both Form X and Form Y used the same generating parameters in the null conditions. In other conditions, Form Y parameters for selected items were shifted by a constant to simulate time-LOI. The particular value of the constant was a condition of the simulation. As time-LOI has not been examined before, there are no established cutoffs for what is considered small or large time-LOI. Therefore, constants of 0.4 and 0.8 were selected based on previous LOI studies to represent small and large amount of LOI, respectively. These constants were selected based on LOI constants for discrimination parameters as time intensity has a similar scale. Small response-LOI was represented by a constant of 0.6 and large response LOI was represented by 0.8.

Parameter recovery. Parameter recovery was evaluated to check the quality of the simulation and the ability of the model to accurately estimate parameters. Parameter recovery of the item parameters was evaluated through bias and root mean squared error (RMSE). Bias is given by

$$bias = \frac{1}{R}\sum_{r=1}^{R}(\hat{\delta}_{ri} - \delta_i),$$

and RMSE is computed as,

$$RMSE = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\delta}_{ri} - \delta_i)^2}{R}},$$

where $\hat{\delta}_{ri}$ is the parameter estimate in replication r, and δ_i, is the true parameter for item i, R represents that total number of replications. Bias and RMSE were computed for the item parameters – time discrimination, time intensity, discrimination[5], and threshold[6].

Details of testing procedure. For all conditions, the full MIMIC-IRTRT model was applied to simultaneously test for response and time-LOI. M*plus* 6.0 (Muthén & Muthén, 1998-2011) was utilized for model parameter estimation.

Anchor items were selected from the LOI-free item subset. For simulation studies, it is common practice, in order to reduce computational burden, to assume that items that were not simulated to have LOI are LOI-free (see Cheng, Shao, Lathrop, 2013; Woods, 2009b; Woods & Grimm, 2011). If the assumption is violated, Finch (2005) showed that the MIMIC model is fairly robust to the presence of LOI in some of the anchor items. Therefore, I will select anchor items from the subset of items that were not simulated to have LOI. The number of anchor items was set to 20% of the total items. For the 20 item test, the last four items on the test were selected as anchor items for both item responses and response time. For the 40 item test, the last eight on the test were selected as anchor items for both item responses and response time. Once anchor items were established, the studied items were tested for LOI. To do this, the MIMIC-IRTRT was run with the LOI effect paths present for all the studied items for both item responses and response times and no LOI effect paths for the anchor items. The model was fit to the data using the robust maximum likelihood estimator[7], which uses the expectation maximization (EM) algorithm. The convergence criterion is set to 0.0001. An item was

[5] The discrimination values were rescaled as the model fixes the loading of the first item to 1.
[6] The threshold was selected for parameter recovery, rather than the difficulty parameter, as it is the parameter being examined when testing for LOI
[7] Denoted MLR in M*plus*

flagged as exhibiting LOI when the p-value for the LOI effect path in the MIMIC-IRTRT model is less than the nominal level of 0.05.

Evaluation of MIMIC-IRTRT model. After the parameter estimates were obtained, the utility of the MIMIC-IRTRT model was evaluated on the basis of type I error and statistical power. A variance components analysis was also conducted to examine the influence of each study condition on power and type I errors.

***Type I error*.** Empirical type I error rates were computed by counting the number of times non-LOI items were flagged as exhibiting LOI and dividing this value by the total number of times all items could be flagged as exhibiting LOI (e.g. the product of the number of items and the number of replications).

***Power*.** Empirical power rates were computed in a similar fashion by introducing known amounts of LOI to population item parameters for a fixed number of items and then finding the proportion of times the detection method flagged the items known to have LOI.

***Variance components analysis*.** The simulation design factors include test length, correlation between person ability and speed, type of LOI, number of LOI items, and magnitude of LOI. A variance component analysis was performed in order to examine the relationship between the design factors and either empirical power rates or type I error rates. The variance component analysis helps identify the conditions that have the most influence on power and Type I error.

Study 2: Extant data

Study 2 utilized extant data in order to demonstrate the utility of the MIMIC-IRTRT in real world settings and compare results to the MIMIC model for item

responses. The extant data included item responses and response times for a college-level, high stakes exam in a health profession field. A total of 13,662 examinees responded to 150 test items. Examinees came from 416 different schools, 424 programs, and included five different degrees types. The test vendor wanted the data to be confidential. Therefore, limited information was provided about the actual items and other variables in the data file. As a result, generic identifiers will be utilized for the groups that are tested for LOI. For the LOI analysis, examinees who identified as degree A (N = 6,761), dummy coded as zero, were compared to examinees who identified as degree B (N = 6,586), dummy coded as one. Computing resources did not permit the inclusion of all items in the analysis. Therefore, a subset of 30 items was utilized for analysis in this study.

Anchor item selection. Prior to LOI testing, designated anchor items were selected empirically based on the procedure described by Woods (2009b). The anchor selection procedure will be briefly described below. The first step in the procedure is to test each item individually in a separate MIMIC-IRTRT model. The second step is to compute the ϕ / standard error (SE) ratio. This is the ratio of the LOI effect divided by the standard error. The third step is to rank the absolute values of the ratios in order from smallest to largest. The fourth and final step of the procedure is to select the items with the smallest ratio, as these items are thought to be LOI-free, as the designated anchor items.

Data analysis. Once anchor items were established, all the studied items were analyzed for LOI using the both the MIMIC model for item responses and the MIMIC-IRTRT model so that comparisons can be made about the items identified as exhibiting

LOI. As in the simulation study, an item was flagged as exhibiting LOI when the p-value for the LOI effect path was less than the nominal level of 0.05.

CHAPTER 4

RESULTS

Chapter 4 provides results from the simulation study that attempt to answer the research questions concerning the viability of the MIMIC-IRTRT model and factors the impacts the results from the MIMIC-IRTRT model, as well as, an extant data analysis to demonstrate the MIMIC-IRTRT model in a practical setting.

Study 1

Parameter recovery. Parameter recovery of the item parameters was completed to check the quality of the simulation. Both bias and RMSE values were utilized in the evaluation of the recovery of item parameters.

P***arameter recovery for response times*.** Table 1 shows the average RMSEs and bias for response time parameters in the null conditions. Overall, time discrimination has lower RMSEs and bias than time intensity; however, both parameters showed very little bias and small RMSEs.

Table 1
Mean RMSE and bias for response time parameters for null conditions

Number of Items	Ability and Speed Correlation	Time Discrimination		Time Intensity	
		Bias	RMSE	Bias	RMSE
20	0	0.00	0.02	0.01	0.03
	0.4	0.00	0.02	-0.01	0.04
	0.8	0.00	0.02	0.00	0.03
40	0	0.00	0.02	0.02	0.04
	0.4	0.00	0.02	-0.01	0.03
	0.8	0.00	0.02	0.01	0.03

Table 2 shows the average RMSEs and bias for time discrimination in the LOI conditions. For all conditions, the bias and RMSE values are very small. The bias values are all positive, indicating a consistent, but slight, overestimation of the time discrimination parameter.

Table 2
Mean RMSE and bias for time discrimination parameters for LOI conditions

Number of Items	Items with LOI	Magnitude of LOI	Ability and Speed Correlation	Type of LOI					
				Time		Response		Time & Response	
				Bias	RMSE	Bias	RMSE	Bias	RMSE
20	3	Small	0	0.00	0.02	0.00	0.02	0.00	0.02
			0.4	0.00	0.02	0.00	0.02	0.00	0.02
			0.8	0.00	0.02	0.00	0.02	0.00	0.02
		Large	0	0.00	0.02	0.00	0.02	0.00	0.02
			0.4	0.00	0.02	0.00	0.03	0.00	0.03
			0.8	0.00	0.02	0.00	0.02	0.00	0.02
	6	Small	0	0.00	0.02	0.00	0.02	0.00	0.02
			0.4	0.00	0.02	0.00	0.02	0.00	0.02
			0.8	0.00	0.02	0.00	0.02	0.00	0.02
		Large	0	0.00	0.02	0.00	0.02	0.00	0.02
			0.4	0.00	0.02	0.00	0.02	0.00	0.02
			0.8	0.00	0.02	0.00	0.02	0.00	0.02
40	3	Small	0	0.00	0.02	0.00	0.02	0.00	0.02
			0.4	0.00	0.02	0.00	0.02	0.00	0.02
			0.8	0.00	0.02	0.00	0.02	0.00	0.02
		Large	0	0.00	0.02	0.00	0.02	0.00	0.02
			0.4	0.00	0.02	0.00	0.02	0.00	0.02
			0.8	0.00	0.02	0.00	0.02	0.00	0.02
	6	Small	0	0.00	0.02	0.00	0.02	0.00	0.02
			0.4	0.00	0.02	0.00	0.02	0.00	0.02
			0.8	0.00	0.02	0.00	0.02	0.00	0.02
		Large	0	0.00	0.02	0.00	0.02	0.00	0.02
			0.4	0.00	0.02	0.00	0.02	0.00	0.02
			0.8	0.00	0.02	0.00	0.02	0.00	0.02

Table 3 shows the average RMSEs and bias for the time intensity parameters for the LOI conditions. The bias and RMSE values are reasonably small and consistent across conditions. This bias values are both positive and negative, indicating over- and under- estimation of the time intensity parameters.

Table 3
Mean RMSE and bias for item time intensity parameters for LOI conditions

Number of Items	Items with LOI	Magnitude of LOI	Ability and Speed Correlation	Type of LOI					
				Time		Response		Time & Response	
				Bias	RMSE	Bias	RMSE	Bias	RMSE
20	3	Small	0	0.01	0.03	0.01	0.03	-0.00	0.03
			0.4	-0.01	0.03	-0.00	0.03	0.00	0.03
			0.8	0.01	0.03	-0.01	0.03	0.00	0.03
		Large	0	0.00	0.03	-0.01	0.03	0.00	0.03
			0.4	-0.01	0.03	0.00	0.03	-0.00	0.03
			0.8	-0.01	0.04	0.00	0.03	-0.01	0.03
	6	Small	0	-0.01	0.03	-0.01	0.03	0.00	0.03
			0.4	0.00	0.03	0.00	0.03	0.00	0.03
			0.8	-0.00	0.03	0.00	0.03	0.00	0.03
		Large	0	0.00	0.03	0.00	0.03	-0.01	0.03
			0.4	0.01	0.03	-0.00	0.03	0.00	0.03
			0.8	0.02	0.04	0.01	0.03	0.01	0.03
40	3	Small	0	-0.01	0.03	0.00	0.03	0.01	0.03
			0.4	0.00	0.03	0.00	0.03	0.00	0.03
			0.8	0.01	0.03	-0.00	0.03	0.01	0.03
		Large	0	-0.00	0.03	0.01	0.03	0.01	0.03
			0.4	-0.00	0.03	0.00	0.03	0.00	0.03
			0.8	-0.01	0.03	0.00	0.03	0.01	0.03
	6	Small	0	0.02	0.04	-0.00	0.03	0.00	0.03
			0.4	0.00	0.03	0.01	0.03	0.01	0.03
			0.8	0.02	0.03	-0.01	0.03	0.01	0.03
		Large	0	0.01	0.03	0.00	0.03	-0.00	0.03
			0.4	0.00	0.03	-0.00	0.03	-0.01	0.03
			0.8	0.00	0.03	0.01	0.03	0.01	0.03

P***arameter recovery for item response parameters***. Table 4 shows the average RMSEs and bias for the item response parameters for the null conditions. The bias and RMSE values are slightly larger for the item response parameters than the response time parameters. Both the threshold and discrimination parameters produce relatively small bias and RMSE values across conditions. The RMSE values are slightly larger for the discrimination parameters than for the threshold parameters.

Table 4
Mean RMSE and bias for item response parameters for null conditions

Number of Items	Ability and Speed Correlation	Discrimination		Threshold	
		Bias	RMSE	Bias	RMSE
20	0	0.01	0.08	-0.01	0.07
	0.4	-0.03	0.15	0.01	0.07
	0.8	0.01	0.13	0.00	0.07
40	0	-0.02	0.12	0.03	0.07
	0.4	0.01	0.08	-0.01	0.07
	0.8	0.02	0.11	0.01	0.06

Table 5 shows the average RMSEs and bias for the discrimination parameters for the LOI conditions. As in the null conditions, bias and RMSE values for the discrimination parameters are reasonably small. The majority of the bias values are positive, which indicates an overestimation of the discrimination parameters.

Table 5
Mean RMSE and bias for discrimination parameters for LOI conditions

Number of Items	Items with LOI	Magnitude of LOI	Ability and Speed Correlation	Type of LOI					
				Time		Response		Time & Response	
				Bias	RMSE	Bias	RMSE	Bias	RMSE
20	3	Small	0	0.00	0.10	0.02	0.10	0.00	0.10
			0.4	0.02	0.09	0.01	0.11	0.01	0.10
			0.8	0.00	0.09	0.00	0.09	0.02	0.09
		Large	0	0.02	0.13	0.02	0.09	0.01	0.12
			0.4	0.02	0.11	0.01	0.09	0.01	0.10
			0.8	0.03	0.13	0.68	0.69	0.01	0.09
	6	Small	0	0.00	0.10	0.02	0.12	0.04	0.14
			0.4	0.00	0.10	0.03	0.10	0.01	0.14
			0.8	0.00	0.09	0.01	0.08	0.03	0.12
		Large	0	-0.04	0.18	0.02	0.10	0.04	0.15
			0.4	0.02	0.11	0.01	0.12	0.03	0.14
			0.8	0.02	0.12	0.00	0.08	0.01	0.10
40	3	Small	0	0.01	0.09	0.01	0.08	0.02	0.10
			0.4	0.01	0.08	0.01	0.09	0.02	0.10
			0.8	0.01	0.10	0.01	0.09	0.01	0.08
		Large	0	0.02	0.08	0.02	0.09	0.01	0.12
			0.4	0.01	0.08	0.00	0.08	0.03	0.12
			0.8	0.01	0.10	0.02	0.02	0.01	0.10
	6	Small	0	0.03	0.11	0.04	0.15	0.01	0.09
			0.4	0.00	0.08	0.01	0.09	0.00	0.07
			0.8	0.01	0.07	0.01	0.08	0.03	0.12
		Large	0	0.00	0.03	0.12	0.06	0.02	0.08
			0.4	0.02	0.11	0.02	0.12	0.02	0.09
			0.8	0.00	0.08	0.01	0.08	0.01	0.10

Table 6 shows the average bias and RMSE values for the threshold parameters in the LOI conditions. The bias and RMSE values do not appear to show an obvious patterns and are relatively small for all LOI conditions. The bias values are both positive and negative, which indicates both over- and under-estimation of the threshold parameters.

Table 6
Mean RMSE and bias for threshold parameters for LOI conditions.

Number of Items	Items with LOI	Magnitude of LOI	Ability and Speed Correlation	Type of LOI					
				Time		Response		Time & Response	
				Bias	RMSE	Bias	RMSE	Bias	RMSE
20	3	Small	0	0.01	0.07	-0.03	0.07	0.02	0.07
			0.4	-0.00	0.07	-0.03	0.08	-0.01	0.07
			0.8	0.01	0.07	-0.02	0.07	0.02	0.07
		Large	0	-0.00	0.07	-0.01	0.07	0.02	0.07
			0.4	-0.03	0.07	-0.04	0.08	-0.01	0.07
			0.8	-0.01	0.07	-0.00	0.07	-0.02	0.07
	6	Small	0	-0.02	0.07	0.01	0.07	-0.01	0.07
			0.4	0.01	0.07	0.00	0.07	-0.02	0.07
			0.8	-0.02	0.07	0.01	0.07	0.02	0.07
		Large	0	-0.03	0.07	-0.01	0.07	0.01	0.07
			0.4	-0.01	0.07	0.01	0.07	0.00	0.07
			0.8	0.02	0.07	0.00	0.07	0.03	0.07
40	3	Small	0	-0.02	0.07	-0.00	0.07	-0.00	0.07
			0.4	-0.01	0.07	0.00	0.07	-0.01	0.07
			0.8	0.01	0.07	-0.01	0.07	0.01	0.07
		Large	0	0.02	0.07	-0.01	0.07	-0.01	0.07
			0.4	0.00	0.07	0.01	0.07	0.01	0.07
			0.8	-0.04	0.08	-0.01	0.07	0.01	0.07
	6	Small	0	-0.02	0.07	0.02	0.07	0.04	0.08
			0.4	-0.01	0.07	0.01	0.07	0.01	0.07
			0.8	0.03	0.07	-0.02	0.07	0.01	0.07
		Large	0	0.01	0.07	0.02	0.07	-0.02	0.07
			0.4	0.00	0.07	-0.02	0.07	0.01	0.07
			0.8	-0.00	0.07	0.04	0.08	0.01	0.07

Type I error rates. Empirical type I error rates for the null cases for the MIMIC-IRTRT model are between 0.04 and 0.05 (see Table 7). These values are close to the nominal level of 0.05. For the fully crossed conditions, the empirical power rates ranged from 0.04 to 0.06 (see Table 8). It appears that the type I error rate for the MIMIC-IRTRT model is not heavily affected by the conditions tested in this study, as all conditions produce error rates right around the nominal level.

Table 7
Empirical Type I Error Rate

Number of Items	Ability and Speed Correlation	Type I Error
20	0	0.04
	0.4	0.05
	0.8	0.05
40	0	0.05
	0.4	0.05
	0.8	0.05

Table 8
Empirical Type I Error Rates

Number of Items	Items with LOI	Magnitude of LOI	Ability and Speed Correlation	Type of LOI		
				Time	Response	Time & Response
20	3	Small	0	0.05	0.05	0.05
			0.4	0.05	0.05	0.05
			0.8	0.05	0.05	0.05
		Large	0	0.05	0.05	0.05
			0.4	0.05	0.05	0.05
			0.8	0.04	0.05	0.05
	6	Small	0	0.04	0.05	0.05
			0.4	0.05	0.04	0.05
			0.8	0.06	0.05	0.05
		Large	0	0.05	0.05	0.05
			0.4	0.05	0.05	0.05
			0.8	0.05	0.05	0.05
40	3	Small	0	0.05	0.05	0.05
			0.4	0.05	0.05	0.05
			0.8	0.05	0.05	0.05
		Large	0	0.05	0.05	0.05
			0.4	0.05	0.05	0.05
			0.8	0.05	0.04	0.05
	6	Small	0	0.05	0.05	0.05
			0.4	0.04	0.06	0.05
			0.8	0.04	0.05	0.05
		Large	0	0.06	0.05	0.05
			0.4	0.05	0.05	0.05
			0.8	0.05	0.05	0.05

Power rates. The MIMIC-IRTRT model exhibits more empirical power to detect uniform LOI in the response time parameters than it does in the item response parameters or when LOI in present in both parameter types (see Table 9). Averaged over all conditions, empirical power to detect uniform LOI for response times is 1.00, but it is 0.94 for item response and 0.95 for both response time and item response.

The level of power for the MIMIC-IRTRT model also seems to be influenced by the magnitude of LOI. When the magnitude is small, the empirical power is 0.94 on average; however, when the magnitude is large, the empirical power is 0.99 on average. This result has more of an impact for detecting uniform LOI in item responses and both response times and item responses, as the power for response times is 1.00 regardless of condition.

Table 9
Empirical Power Rates

Number of Items	Items with LOI	Magnitude of LOI	Ability and Speed Correlation	Type of LOI		
				Time	Response	Time & Response
20	3	Small	0	1.00	0.96	0.88
			0.4	1.00	0.87	0.83
			0.8	1.00	0.90	0.99
		Large	0	1.00	0.99	0.99
			0.4	1.00	1.00	1.00
			0.8	1.00	1.00	0.99
	6	Small	0	1.00	0.88	0.91
			0.4	1.00	0.96	0.91
			0.8	1.00	0.86	0.80
		Large	0	1.00	0.99	0.96
			0.4	1.00	0.99	0.98
			0.8	1.00	0.94	1.00
40	3	Small	0	1.00	0.84	0.95
			0.4	1.00	0.95	0.97
			0.8	1.00	0.85	1.00
		Large	0	1.00	1.00	1.00
			0.4	1.00	0.98	1.00
			0.8	1.00	1.00	0.96
	6	Small	0	1.00	0.92	0.96
			0.4	1.00	0.96	0.92
			0.8	1.00	0.87	0.87
		Large	0	1.00	1.00	1.00
			0.4	1.00	0.98	1.00
			0.8	1.00	0.97	0.98
		Average		1.00	0.94	0.95

Variance components analysis. The simulation design factors include the number of items, the number of LOI items, type of LOI, magnitude of LOI, and the correlation between person ability and speed. The relationship between the design factors and the detection of LOI items were examined by utilizing a variance components analysis.

Table 10 provides the variance components estimates for type I error rate. The largest variance component for type I error rate was .13, which makes up 38% of the total variation, for the interaction between number of items, number of LOI items, type of LOI, and magnitude of LOI. All the other variance components had moderate to negligible impact on the type I error rates. This result is in line with the type I error output, as the error rates were almost identical for all conditions

Table 10
The Largest 10 Variance Components for Type I Error

Rank	Factor	Variance Estimate	Percent Total Variance
1	No. of items×No. of LOI items×Type of LOI×Magnitude of LOI	.13	38
2	No. of items×No. of LOI items×Type of LOI	.05	14
3	No. of items×Ability and Speed Correlation×No. of LOI items×Type of LOI	.03	7
4	Ability and Speed Correlation×No. of LOI items×Type of LOI×Magnitude of LOI	.02	7
5	No. of items×Abiliy and Speed Correlation	.02	6
6	No. of items×Ability and Speed Correlation×Magnitude of LOI	.02	6
7	Type of LOI	.01	4
8	No. LOI items	.01	3
9	No. of LOI items×Magnitude of LOI	.01	2
9	Ability and Speed Correlation×Magnitude of LOI	.01	2

Note: The variances are calculated using the dependent variable of type I error times 100

For power, the variance components analysis indicated that there was much more variance due to the study conditions than seen for type I error (see Table 11). The most influential component was magnitude of LOI with a variance of 10.29 (21%), followed by the interaction between type of LOI and magnitude of LOI with a variance of 8. 35 (17%). The interaction of all the factors has the third largest influence on power, followed by the interaction of all the factors except the number of items. Type of LOI has the fifth largest influence with a variance estimate of 4.67. All other factors had a moderate to slight impact on power.

Table 11
The Largest 10 Variance Components for Power

Rank	Factor	Variance Estimate	Percent Total Variance
1	Magnitude of LOI	10.29	21
2	Type of LOI ×Magnitude of LOI	8.53	17
3	Abilty and Speed Correlation×No. of LOI items×Type of LOI ×Magnitude of LOI	7.94	16
4	No. of items×Ability and Speed Correlation×No. of LOI items×Type of LOI ×Magnitude of LOI	7.42	15
5	Type of LOI	4.67	10
6	No. of items×Type of LOI ×Magnitude of LOI	2.85	6
7	No. of items×Ability and Speed Correlation×No. of LOI items×Magnitude of LOI	2.57	5
8	Ability and Speed Correlation×Type of LOI	1.64	3
9	Ability and Speed Correlation×No. of LOI items	1.42	3
10	No. of items×Ability and Speed Correlation×No. of LOI items	0.68	1

Note: The variances are calculated using the dependent variable of power rate times 100

The variance components analysis results are consistent with what was seen for the type I error and power outputs. The individual design factors have very little impact on the type I error results, but the magnitude of LOI and the type of LOI have large impacts on the power results.

Study 2: Analysis of Extant Data

Anchor item selection. The anchor item selection procedure described previously was applied to the data. Table 12 provides the ratios for both item responses and response times. Three items from each response type were chosen as anchor items. These absolute values of the ratios were rank ordered and the three smallest ratios for each response type were selected for anchors. The anchors for response time were 3, 7, and 27. The anchors for item responses were 14, 27, and 28 for both the MIMIC and MIMIC-IRTRT models.

Table 12
The ϕ/SE ratios for item responses and response times

Item	Ratio		
	MIMIC-IRTRT		MIMIC
	Response Times	Item Responses	Item Responses
1	-2.079	7.705	7.709
2	1.541	-4.510	-4.545
3	**0.203**	-8.746	-8.761
4	-4.629	2.339	2.320
5	1.202	2.723	2.720
6	-2.334	3.478	3.469
7	**0.168**	3.796	3.808
8	5.094	-3.300	-3.326
9	-4.107	2.154	2.149
10	-3.490	3.466	3.464
11	-5.984	1.751	1.739
12	-5.637	9.208	9.198
13	-0.961	-1.022	-1.021
14	-1.666	**0.563**	**0.558**
15	-2.538	6.025	6.033
16	0.544	-2.281	-2.301
17	-1.358	-3.804	-3.813
18	7.657	-3.613	-3.630
19	4.230	6.494	6.500
20	5.383	-5.180	-5.202
21	-4.696	13.598	13.584
22	3.641	-3.981	-4.018
23	9.265	-4.173	-4.189
24	6.797	-11.668	-11.686
25	-5.678	3.824	3.816
26	2.796	-3.696	-3.714
27	**0.107**	**-0.159**	**-0.189**
28	0.689	**-0.394**	**-0.393**
29	1.453	-0.761	-0.777
30	-1.397	-1.165	-1.165

Note: Bolded values indicate that the item was selected as an anchor

Data analysis. The MIMIC-IRTRT was run for the rest of the items using the designated anchor items. Table 13 shows the parameter estimates for LOI for both item responses and response times. From MIMIC-IRTRT model, items 6, 8, 9, 10, 12, 15, 18, 19, 20, 21, 22, 23, 24, 25, and 26 were identified as exhibiting both time and response-LOI. Items 4 and 11 were identified as exhibiting time-LOI only and items 1, 2, 3, 5, 7,

and 17 were identified as exhibiting response-LOI only. For the MIMIC model items 1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, and 26 were identified as exhibiting response LOI.

Table 13
MIMIC-IRTRT and MIMIC LOI parameter estimates and p-values

Item	LOI Effects		
	MIMIC-IRTRT		MIMIC
	Response Times	Item Responses	Item Responses
1	-0.008 (0.053)	**0.273 (0.000)**	**0.275 (0.000)**
2	0.005 (0.220)	**-0.192 (0.000)**	**-0.190 (0.000)**
3	--	**-0.314 (0.000)**	**-0.313 (0.000)**
4	**-0.016 (0.000)**	0.100 (0.067)	0.101 (0.061**)**
5	0.004 (0.335)	**0.097 (0.012)**	**0.098 (0.011)**
6	**-0.008 (0.033)**	**0.184 (0.014)**	**0.188 (0.013)**
7	--	**0.134 (0.001)**	**0.136 (0.001)**
8	**0.015 (0.000)**	**-0.172 (0.005)**	**-0.170 (0.005)**
9	**-0.015 (0.000)**	**0.165 (0.039)**	**0.167 (0.037)**
10	**-0.011 (0.002)**	**0.122 (0.001)**	**0.123 (0.001)**
11	**-0.025 (0.000)**	0.060 (0.165)	0.062 (0.154)
12	**-0.019 (0.000)**	**0.353 (0.000)**	**0.354 (0.000)**
13	-0.004 (0.329)	-0.036 (0.319)	-0.036 (0.333)
14	-0.006 (0.120)	--	--
15	**-0.010 (0.018)**	**0.232 (0.000)**	**0.234 (0.000)**
16	0.001 (0.739)	-0.094 (0.063)	-0.093 (0.067)
17	-0.004 (0.213)	**-0.158 (0.002)**	**-0.157 (0.001)**
18	**0.035 (0.000)**	**-0.155 (0.002)**	**-0.153 (0.003)**
19	**0.013 (0.001)**	**0.239 (0.000)**	**0.243 (0.000)**
20	**0.021 (0.000)**	**-0.223 (0.000)**	**-0.220 (0.000)**
21	**-0.018 (0.000)**	**0.509 (0.000)**	**0.509 (0.000)**
22	**0.014 (0.002)**	**-0.166 (0.032)**	**-0.161 (0.035)**
23	**0.035 (0.000)**	**-0.199 (0.001)**	**-0.197 (0.001)**
24	**0.025 (0.000)**	**-0.430 (0.000)**	**-0.429 (0.000)**
25	**-0.018 (0.000)**	**0.176 (0.005)**	**0.179 (0.004)**
26	**0.009 (0.026)**	**-0.155 (0.001)**	**-0.153 (0.001)**
27	--	--	--
28	0.002 (0.666)	--	--
29	0.004 (0.302)	-0.030 (0.504)	-0.028 (0.524)
30	-0.006 (0.178)	-0.044 (0.298)	-0.042 (0.323)

Note: Bolded values indicate that the item was flagged as indicating LOI

If the LOI effect is positive, that indicates that the thresholds and time intensities are larger for degree B examinees. A negative LOI effect indicates that the thresholds and time intensities are smaller for degree B examinees.

CHAPTER 5

DISCUSSION

The purpose of this study was to examine the viability of the newly proposed MIMIC-IRTRT model for detecting LOI in both item responses and response times. Both simulation and extant data analyses were conducted toward this end and a discussion of the major findings from each of these analyses are provided in the following sections.

Study 1: Simulation

The simulation study included five factors (i.e., number of items, correlation between person ability and speed, number of LOI items, type of LOI items, and magnitude of LOI) resulting in a total of 78 conditions. Overall, the results were favorable for the use of the MIMIC-IRTRT model in detecting LOI for both item responses and response times. Parameter recovery indicated that the parameters were well recovered. Overall, bias and RMSE values were smaller for response time parameters than item response parameters. Type I error was at the nominal level for all conditions. Power rates were high and impacted by the magnitude of LOI and type of LOI. The difference in the bias and RMSE values for the response time and item response parameters may help to explain the differences in the power rates due to type of LOI. As the purpose of this study was to examine the MIMIC-IRTRT model, other item response methods for detecting LOI were not utilized. As such, a direct comparison cannot be made between the results from the MIMIC-IRTRT model and existing methods for detecting LOI in the item response parameters; however, general comparisons can be

discussed to help situate the results from this study. If the results of this study are compared to studies that utilized the traditional methods for detecting LOI, including the Mantel-Haenszel, SIBTEST, Lord's chi-square, and the MIMIC model, in the 2PL model for the item responses, then both the type I error rates and the power appear to be similar where the type I errors are around the nominal level and a power is relatively high (see Finch, 2005; Kim & Cohen, 1995; Kim, Cohen, & Kim, 1994; Lim & Drasgow, 1990; Woods, 2009). The performance of the MIMIC-IRTRT model when LOI in both item responses and response times was present was comparable to the performance of the model when there was a LOI in just the item responses. As for the performance of the MIMIC-IRTRT model in detecting LOI in response times, there are no other methods to compare; however, given that the model detected the LOI 100% of the time, with an error rate right at the nominal level, confidence in the performance for detecting LOI in response times is high and one can say that the MIMIC-IRTRT model performs very well when identifying LOI for response time.

Study 2: Extant Data

The extant data analysis consisted of applying the MIMIC-IRTRT model to a college-level, high stakes exam, in a health profession field. Examinees with degree A and degree B were compared. Results from the MIMIC-IRTRT model were compared to the MIMIC model for item responses. The MIMIC-IRTRT model revealed that 23 out of 30 (~77%) items exhibited some type of LOI. Fifteen items exhibited LOI in both response time and item responses. Six items exhibited LOI for responses only and two items exhibited LOI for times only. The traditional MIMIC model for item responses identified a total 21 items exhibiting LOI. These items were the same items identified as

having both response LOI and time LOI or just response LOI in the MIMIC-IRTRT model. Through comparison of the two types of models it does not seem that the MIMIC-IRTRT model improves the detection of response-LOI for this particular dataset; however, it does seem clear that the MIMIC-IRTRT model does help to flag potentially problematic items that may go unnoticed with traditional LOI detection methods (i.e., items that exhibit time-LOI).

The exceptionally high number of items that were identified as exhibiting some type of LOI, brings up the importance of effect size measures for distinguishing between statistical significance and practical significance. To date, there is not an established measure for determining practical significance for the MIMIC model.

Limitations/Future Directions

As this was an exploratory study, examining a newly proposed model there are limitations and many areas that could be addressed in future studies. One limitation has to do with the item response model used in the study. Because M*plus* is not currently capable of fitting a 3PL model, this study was only able to examine the utility of the MIMIC-IRTRT model for the 2PL model; however, it has been stated that M*plus* may introduce the 3PL model in the future and at that time, this study could be extended to examine the utility of the proposed model when a 3PL model is fitted to the response data (Sen, 2012). Future studies may also want to incorporate polytomous response models, where there are more than two response options, in order to determine the utility of the MIMIC-IRT for more complex data types.

In addition, decisions had to be made about which conditions to include and how many levels of each condition. In the future, other conditions may be examined to see the

impact on the results, such as including different sample sizes, more magnitude levels, different numbers of anchor items and differences in ability levels for the two groups. Based on the extant data results, one may also want to examine a higher percentage of LOI items. The simulation study only included datasets with up to 30% of items exhibiting LOI. By incorporating more conditions, one would get a better understanding of the situations in which the model works best and determining if factors like different ability level distributions and high percentage of LOI items impacts the power and type I error results. In addition, by incorporating more magnitude levels for the response times, one may be able to identify the threshold for LOI detection on the time intensity parameter.

Based on the results of the extant data analysis, one potential avenue for future exploration could include simulation studies that compare the performance of the MIMIC-IRTRT model for detecting LOI in item responses to established methods for detecting item responses, such as the MIMIC model, under a variety of conditions. This would help determine how the MIMIC-IRTRT model compares to methods currently being implemented to detect LOI in item responses. By looking at these comparisons, one might get a better understanding of whether the MIMIC-IRTRT model improves the detection of LOI in the item responses.

Another important direction for future research is sparked by the lack of an effect size measure for the MIMIC-IRTRT model. For traditional LOI detection methods, like the Mantel-Haenszel and SIBTEST, the significance test is just one part of the detection process. Measures of effect size, like the common odds ratio and standardized mean difference, allow the researcher to determine the magnitude of LOI, so that items can be

categorized as exhibiting negligible, moderate, or large amounts of LOI (Zwick, 1993). As such, establishing a credible effect size measure for the MIMIC-IRTRT model is a worthwhile avenue for future research.

Conclusion

The ability to detect LOI for both item responses and response times is of interest to test developers and administrators as this can have impacts on validity and fairness. With the introduction of computer-based testing, response times are easily captured and the information captured from response times can help in the detection of LOI. There has been little research that examines LOI in response times and no research on the combination of LOI in both item responses and response times. As such, I introduced the MIMIC-IRTRT model, which can detect LOI for both item response and response times. This study revealed that the model is exceptionally good at detecting LOI in response times and performs comparably to current methods for detecting LOI in item responses. In sum, the MIMIC-IRTRT model may be a viable option for improving detection of LOI and therefore requires further investigation in the future.

REFERENCES

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the raven's advanced progressive matrices: evidence for bias. *Personality and Individual Differences, 36*, 1459-1470.

Abedi, J., Leon, S., & Kao, J. (2008). *Examining differential item functioning in reading assessments for students with disabilities*. (CRESST Tech. Rep. No. 744). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ackerman, T. A. (1992). An explanation of differential item functioning from a multidimensional perspective. *Journal of Educational Measurement, 24*, 67-91.

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language*. (Research Rep. No. 81-16). Princeton, NJ: Educational Testing Service.

Angoff, W. H. & Sharon, A. T. (1974). The evaluation of differences in test performances of two or more groups. *Educational and Psychological Measurement, 34*, 807-816.

Arce-Ferrer, A. (2008). Comparing screening approaches to investigate stability of common items in Rasch equating. *Journal of Applied Measurement*, *9*, 1-11.

Arce, A. J. & Lau, C. A. (2011). *Statistical properties of 3PL robust z: An investigation with real and simulated data sets.* Presented at the annual meeting of the National Council on Measurement in Education. New Orleans, Louisiana.

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement, 36,* 565-580.

Baker, F. B. (1992). *Item response theory.* New York: Springer-Verlag.

Barnes, B. J., & Wells, C. S. (2009). Differential item functional analysis by gender and race of the national doctoral program survey. *International Journal of Doctoral Studies, 4*, 77-96.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15*, 129–137.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57,* 289-300.

Bleistein, C. A., & Wright, D. (1987). Assessment of unexpected differential item difficulty for Asian-American candidates on the Scholastic Aptitude Test. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (RM-87-1). Princeton, NJ: Educational Testing Service.

Bock, R.D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Bolt, S. E. (2004, April). *Using DIF analyses to examine several commonly-held beliefs about testing accommodations for students with disabilities.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego.

Cheng, Y., Shao, C., & Lathrop, Q. (2013). The mediated MIMIC model for explaining the underlying mechanism of differential item functioning. Paper presented at National Council of Measurement in Educational annual meeting, San Francisco, CA.

Christensen, H., Jorm, A. F., MacKinnon, A. J., Korten, A. E., Jacomb, P. A., Henderson, A. S., et al. (1999). Age differences in depression and anxiety symptoms: A structural equation modeling analysis of data from a general population sample. *Psychological Medicine, 29*, 325–339.

Clauser, B. & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *ITEMS*, 281-294.

Clauser, B., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*, 67-78.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice, 20*, 225–233.

Cohen, A. S., & Kim, S. (1993). A comparison of Lord's χ^2 and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*, 39-52.

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*, 15–26.

DeMars, C. E. (2004a). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, 17, 265-300.

DeMars, C.E. (2004b). Item *parameter drift: The impact of the curricular area*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

DeMars, C.E. & Wise, S.L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? International Journal of Testing, 10, 207-229.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*, 33-51.

Dorans, N.J., & Holland, P.W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). *Intelligence and cultural differences.* Chicago: University of Chicago Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380–396.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Hanszel, SIBTEST, and the IRT likelihood ration. *Applied Psychological Measurement, 29*, 278-295.

Finger, M.S., & Chuah, C.S. (2010). *Response time model estimation via confirmatory factor analysis*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences, 57B*, 275–284.

Fox, J.P., Klein Entink, R., van der Linden, W. (2007). Modeling of Responses and Response Times with the Package cirt. *Journal of Statistical Software, 20*, 1-14.

Gierl, M. J. (2005). Using dimensionalisty-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*, 3-14.

Gierl, M. J., Khaliq, S. N., & Boughton, K. A. (1999). *Gender differential item functioning in mathematics and science: prevalence and policy implications.* Paper presented at the Annual Meeting of the Canadian Society for the Study of Education.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*, 369-377.

Gotzmann, A., Wright, K., Rodden, L. (2006). *A comparison of power rates for items favoring the reference and focal group for the mantel-haenszel and SIBTEST procedures*. Paper presented at the American Educational Research Association (AERA) in San Francisco, California.

Green, J., Smith, J., & Habing, B. (2010*). A comparison of the robust z, mantel-haenszel, and lord's χ^2 methods for item drift detection*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Gu, L., Drake, S., & Wolfe, E. W. (2006). Differential item functioning for GRE mathematics items across computerized and paper-and-pencil testing media. *Journal of Technology, Learning, and Assessment, 5*.

Guo, F. (2009, February). *Quantifying impact of compromised items in CAT.* Paper presented at the 2009 National Council on Measurement in Education Meeting, San Diego, CA.

Haladyna, T. M., and Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17-27.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Item response theory.* Vol. 2. Hillsdale, NJ: Lawrence Erlbaum.

Han, K. T., & Guo, F. (2011). Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing. Graduate Management Admission Council, Research Reports, RR-11-02.

Hauser, C., & Kingsbury, G. (2004). *Differential item functioning and differential test functioning in the Idaho Standards Achievement Tests for spring 2003.* Lake Oswego, OR: Northwest Evaluation Association.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W. & Wainer, H. (1993). Differential Item Functioning. Routledge, London, UK.

Hogg, R. V. (1979). Statistical robustness: One view on its use in applications today. *The American Statistician, 33*, 108-115.

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32,* 311-333.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73-101.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Huynh, H. (1982). A comparison of four approaches to robust regression. *Psychological Bulletin*, *92*, 505-512.

Huynh, H., & Meyer, J. P. (2009). Use of robust z in detecting unstable items in item response theory. *Practical Assessment, Research and Evaluation, 15*, 1-8.

Huynh, H., & Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In Everett V. Smith Jr. & Greg E. Stone (Eds.) *Applications of Rasch Measurement in Criterion-Referenced Testing: Practice Analysis to Score Reporting*. Maple Grove, MN: JAM Press.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Kim, S. & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15*, 269-278.

Kim, S., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345–355.

Kim, S., Cohen, A. S., & Kim, H. (1994). An investigation of lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*, 217-228.

Kingston, M. K., & Dorans, N. J. (1984). Item location effect and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

Klein Entink, R. H. (2009). Statistical models for responses and response times. (Unpublished doctoral dissertation). University of Twente, Enschede.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles, CA: University of California, Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., & Hamilton, L. (1999). *Assessing students with disabilities in Kentucky: The effects of accommodations, format, and subject*. (CRESST Tech. Rep. No. 498). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kulick, E. (1984). *Assessing unexpected differential item performance of Black candidates on SAT form CSA6 and TSWE form E33* (SR-84-80). Princeton, NJ: Educational Testing Service.

Kulick, E., & Dorans, N. J. (1983). *Assessing unexpected differential item performance of Oriental candidates on SAT form CSA6 and TSWE form E33* (SR-83-106). Princeton, NJ: Educational Testing Service.

Lawrence, I. A., Curley, W. E., & McHale, F. J. (1988). *Differential item functioning for males and females on SAT-verbal reading subscore items*. (College Board Rep. No. 88-4). New York, New York: College Entrance Examination Board.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology*, 75, 164-174.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Portinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets and Zeitlinger.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Luce, D. R. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139-160.

Maris, E. (1993). Additive and multiplicative models for gamma distributed variables, and their application as psychometric models for response times. *Psychometrika*, 58, 445–469.

Martineau, J. A. (2004*). The effects of construct shift on growth and accountability models* (Unpublished doctoral dissertation). Michigan State University, East Lansing.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based value-added accountability. *Journal of Educational and Behavioral Statistics, 31*, 35-62.

McLaughlin, M. E. & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*, 161-173.

Mellenberg, G, J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7* , 105-108.

Meyer, P., & Huynh, H. (2010). *Evaluation of the Robust z procedure for detecting item parameter drift in 3PLM and GPCM mixed format items*. Paper presented at the annual meeting of National Council on Measurement in Education, Denver, CO.

Mitchelson, J. K., Wicher, E. W., LeBreton, J. M., & Craig, S. B. (2009). Gender and ethnicity differences on the abridged big five circumplex (AB5C) of personality traits: A differential item functioning analysis. *Educational and Psychological Measurement, 69*, 613-635.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology,12*, 252–284.

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10*, 121-132.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Erlbaum.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*, 150–166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*, 107–124.

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*, 41–70.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Reynolds, C.R., & Kaiser, S.M. (1990). Test bias in psychological assessment. In T. B. Gutkin & C.R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed., pp. 487-525). New York: Wiley.

Rogers, H. J., Dorans, N. J., & Schmitt, A. P. (1986). Assessing unexpected differential item performance of black candidates on SAT Form 3GSA08 and TSWE Form E43: November 1984 administration. Unpublished Statistical Report SR-86. Princeton, NJ: Educational Testing Service.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 187–208). New York: Springer.

Roussos, L.A., & Stout, W.F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33,* 215-230.

Rupp, A. A., & Zumbo, B. D. (2003a, April). *Bias coefficients for lack of invariance in unidimensional IRT models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Rupp, A. A., & Zumbo, B. D. (2003b). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *The Alberta Journal of Educational Research, XLIX,* 264-276.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*, 63-84.

Sacco, P., Casado, B. L., Unick, G. J. (2011). Differential item functioning across race in aging research: An example using a social support measure. *Clinical Gerontologist, 34*, 57-70.

Samuelson, K. (2005). Examining differential item functioning from a latent class perspective. (Unpublished doctoral dissertation). University of Maryland, College Park.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology, 19*, 18–38.

Schmitt, A. P. (1985). *Assessing unexpected differential item performance of Hispanic candidates on SAT form 3FSA08 and TSWE form E47* (SR-85-169). Princeton, NJ: Educational Testing Service.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement, 25*, 1-13.

Schmitt, A.P., & Bleistein, C.A. (1987). F*actors affecting differential item functioning for black examinees on scholastic aptitude test analogy items* (Research Report No. 87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A.P., & Dorans, N.J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, 27, 67-81.

Sen, R. (2012). Structural equation model approach to the use of response times for improving estimation in item response models. (Unpublished doctoral dissertation). University of Massachusetts, Amherst.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,*159-194.

Sheppard, R., Han, K., Colarelli, S. M., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment, 13*, 442-453.

Sinharay, S., Dorans,N. J., & Liang, L. (2009). First language of examinees and its relationship to differential item functioning. ETS RR-09-11. Princeton, New Jersey.

Snetzler, S., & Qualls, A. L. (2000). Examination of differential item functioning on a standardized achievement battery with limited English proficient students. *Educational and Psychological Measurement, 60*, 564–577.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1291-1306.

Sukin, T. M. (2010). Item parameter drift as an indication of differential opportunity to learn: An exploration of item flagging methods and accurate classification of examinees. (Unpublished Dissertation). University of Massachusetts, Amherst.

Swaminathan, H., & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT b values. *Journal of Educational Measurement, 29*, 201-211.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

van Breukelen, G. J. P. (1989). *Concentration, speed, and precision in mental tests*. (Unpublished Dissertation). University of Nijmegen, The Netherlands.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement, 44*, 117-130.

van der Linden, W. J., Klein Entink, R., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*, 327-347.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*, 195–210.

Verhelst, N. D., Verstraalen, H. H. F. M., & Jansen, M. G. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds*.), Handbook of modern item response theory* (pp. 169–185). New York: Springer.

Wainer, H. (1993). Measuring differential impact of items. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.3-23). Hillsdale, NJ: Erlbaum.

Wang,W. (2004). Effects of anchor item methods on detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education, 72*, 221–261.

Wang, W., & Yeh, Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479–498.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). Effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*, 77-87.

Wendler, C.L.W., & Carlton, S. R. (1987, April). *An examination of SAT-verbal items for differential performance by women and men: An exploratory study.* Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Wollack, J. A., Sung, H. J., & Kang, T. (2005, April). *Longitudinal effects of item parameter drift*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Wollack, J. A., Sung, H. J., & Kang, T. (2006, April). *The impact of compounding item parameter drift on ability estimation.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Woo, A. & Dragan, M. (2012). Ensuring validity of NCLEX with differential item functioning analysis. *Journal of Nursing Regulation, 2*, 29-31.

Woodard, J. L., Auchus, A. P., Godsall, R. E., & Green, R. C. (1998). An analysis of test bias and differential item functioning due to race on the mattis dementia rating scale. *The Journals of Gerontology, 53B*, 370-374.

Woods, C. M. (2009a). Empirical selection for anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*, 42-57.

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*, 1-27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause model. *Applied Psychological Measurement, 35*, 339-361.

Zwick, R. (1990). When do item response function and mantel-haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.

Zwick, R. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

APPENDIX A
R Code for Data Simulation

```
#=============================================================================
        # SIMULATE ITEM RESPONSE AND RESPONSE TIME DATA
        # Adapted from generate() function in cirt package by Entink
        # --Item responses stored in matrix called resp
        # --Item resposne time stored in matrix called time
#=============================================================================

library(cirt)
library(MBESS)
#=============================================================================
# persons parameter summary statistics
#=============================================================================
N<-1000
mut<-0  #mean person ability
muz<-0  #mean person speed

#=============================================================================
# generate person parameters
#=============================================================================
PersonGen<-function(N,mut,muz,rho, file){

        cor.mat<-matrix(c(1,rho,rho,1),nrow=2)
        SD<-c(1,.2236068)
        cov.mat<-cor2cov(cor.mat,SD)

        PX<-mvrnorm(N, mu=c(mut, muz), Sigma=matrix(c(1,cov.mat[1,2],cov.mat[1,2],.05), 2))
        thetaX<-PX[,1]
        zetaX<-PX[,2]

        PY<-mvrnorm(N, mu=c(mut, muz), Sigma=matrix(c(1,cov.mat[1,2],cov.mat[1,2],.05), 2))
        thetaY<-PY[,1]
        zetaY<-PY[,2]

        PersonParams<-cbind(thetaX, zetaX, thetaY,zetaY)
        names(PersonParams)<-c("thetaX", "zetaX", "thetaY", "zetaY")
        write.csv(x=PersonParams, file=file, quote=FALSE, row.names=FALSE)
        result<-list(thetaX=thetaX, thetaY=thetaY, zetaX=zetaX, zetaY=zetaY)
        result
}

#=============================================================================
# item parameter summary statistics
#=============================================================================
mua<-.6156964   #mean item discrimination
mub<--.6893613  #mean item difficulty
mup<-1.015708   #mean item time discrimination
mul<-4.132949   #mean item time intensity

cap<-.04        #cov between item discrimination and time discrimination
cbl<-.17        #cov between item difficulty and time intensity
cab<-.01        #cov between item discrimination and item difficulty
cal<-.02        #cov between item discrim and time intensity
cbp<-.02        #cov between item difficulty and time discrim
```

```
cpl<--.01          #cov between time discrim and time intensity

va<-.06            #variance item discrimination
vb<-.55            #variance item difficulty
vp<-.03            #variance time discrimination
vl<-.78            #variance time intensity

#============================================================================
# generate item parameters
#============================================================================I
ItemGen<-function(K,b_dif, l_dif, items_to_dif,file){

        X<-mvrnorm(K, mu = c(mua, mub, mup, mul), Sigma = matrix(c(va, cab, cap, cal, cab, vb, cbp,
        cbl, cap, cbp, vp, cpl, cal, cbl, cpl, vl), 4))

        alpha<-X[,1]            #item discrimination
        beta<-X[,2]             #item difficulty
        phi<-X[,3]              #time discrimination
        lambda<-X[,4]           #time intensity


        # add dif to difficulty parameters

        beta_dif<-beta

        for (kk in 1:K) {

                if(kk <= items_to_dif) {
                        beta_dif[kk] <- beta_dif[kk]+ b_dif
                } else {
                        beta_dif[kk] <- beta_dif[kk]
                }
        }

        # add dif to time intensity parameters

        lambda_dif<-lambda

        for (kk in 1:K) {

                if(kk <= items_to_dif) {
                        lambda_dif[kk] <- lambda_dif[kk]+ l_dif
                } else {
                        lambda_dif[kk] <- lambda_dif[kk]
                }
        }

        itemParams<-cbind(alpha, beta, beta_dif, phi, lambda, lambda_dif)
        itemParams<-as.data.frame(itemParams)
        names(itemParams)<-c("alpha", "beta", "beta_dif", "phi", "lambda", "lambda_dif")
        write.csv(x=itemParams, file = file,quote=FALSE, row.names=FALSE)

        params<-list(alpha=alpha, beta=beta, beta_dif=beta_dif,lambda=lambda, lambda_dif=lambda_dif,
        phi=phi)

}
```

```
#=============================================================================
# Create file names for item and person parameters for two different test forms.
#This function allows the creation of absolute file names for different conditions.
#-----------------------------------------------------------------------------------------
# condition = condition number
# path = path to output files
#=============================================================================f
fileNames<-function(condition, path){
        itemFile<-paste(path, condition, "/item-param-c", condition, ".txt", sep="")
        personFile<-paste(path,  condition, "/person-param-c", condition, ".txt", sep="")
        result<-list(item=itemFile, person=personFile)
        result
}

#=============================================================================
#function to generate item responses for each item
#=============================================================================
respGen<-function(kk, alpha, beta, theta, N) {
        draw<-runif(N, 0, 1)
        prob<-1/(1+exp(-alpha[kk]*(theta[1:N]-beta[kk])))
        y<-ifelse(draw < prob, 1, 0)
        y
}

#=============================================================================
#function to generate response times for each item
#=============================================================================
timeGen<-function(zeta, lambda, phi, K){

        time <- matrix(0, nrow = N, ncol = K)        #initialized the log-normal response times to zero
        time[1:N, ] <- time[1:N, ] + zeta[1:N]

        for (ii in 1:K) {
                time[1:N, ii] <- lambda[ii] - time[1:N, ii] + rnorm(N , mean=0, sd= 1/phi[ii])
                }
        time[1:N,]
}

#=============================================================================
#generate and store item responses and response times for all persons to all items
#=============================================================================
simdata<-function(condition, K, itemparams, personparams, reps){

        for (repnum in 1:reps){

        respX <- matrix(unlist(lapply(1:K, respGen, theta = personparams$thetaX, alpha =
        itemparams$alpha, beta = itemparams$beta, N = N)), ncol = K, nrow = N)

        respY <- matrix(unlist(lapply(1:K, respGen, theta = personparams$thetaY, alpha =
        itemparams$alpha, beta = itemparams$beta_dif,  N = N)), ncol = K, nrow = N)

        timeX<-timeGen(K=K, lambda=itemparams$lambda, phi=itemparams$phi,
        zeta=personparams$zetaX)
```

```
        timeY<-timeGen(K=K, lambda=itemparams$lambda_dif, phi=itemparams$phi,
        zeta=personparams$zetaY)

        time_resp<-merge(respX, timeX, by = "row.names")          #merge data for noDIF group
        time_resp["group"]<-NA
        time_resp["group"]<-0
        time_resp_dif<-merge(respY,timeY, by ="row.names")        #merge data for DIF group
        time_resp_dif["group"]<-NA
        time_resp_dif["group"]<-1
        time_data<-rbind(time_resp,time_resp_dif)

        write.table(x=time_data,file=paste(basePath, "/c-", condition, "/time_data-", repnum, ".csv",
        sep=""), row.names=FALSE, col.names=FALSE, sep =",")
}
```

APPENDIX B
Mplus Code for MIMIC-IRTRT Model

```
VARIABLE:  NAMES ARE examinee R1-R20 T1-T20 group;
           CATEGORICAL ARE R1-R20;
           USEVARIABLES ARE R1-R20 T1-T20 group;

ANALYSIS:  ESTIMATOR=mlr;

MODEL:  theta BY R1@1 R2-R20*;
        speed BY T1-T20@-1;
        speed@1;
        [speed@0];
        theta ON speed;
         theta ON group;
        speed ON group;

        !uniform LOI – items being tested for LOI
         T1-T16 ON group;
         R1-R16 ON group;
        ! R17-R20 and T17-T20 are used as anchors
```