This is to certify that the
dissertation entitled

EXTENDING THE PARTIAL CREDIT AND RATING SCALE MODELS USING THE HIERARCHICAL MULTIVARIATE GENERALIZED LINEAR MODEL

presented by

JONATHAN R. MANALO

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Measurement and Quatitative Methods

Kimberly A. Maier
Major Professor's Signature

14 December 2004
Date

MSU is an Affirmative Action/Equal Opportunity Institution

EXTENDING THE PARTIAL CREDIT AND RATING SCALE MODELS USING THE HIERARCHICAL MULTIVARIATE GENERALIZED LINEAR MODEL

By

Jonathan R. Manalo

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational
Psychology, and Special Education

2004

ABSTRACT

EXTENDING THE PARTIAL CREDIT AND RATING SCALE MODELS USING THE HIERARCHICAL MULTIVARIATE GENERALIZED LINEAR MODEL

By

Jonathan R. Manalo

In this dissertation, the Rating Scale and Partial Credit Models of Item Response Theory (IRT) are extended using a hierarchical multivariate generalized linear model (HMGLM). Specifically, previous extensions of IRT using hierarchical linear modeling (HLM) are discussed by highlighting their weaknesses and how by applying the HMGLM their weaknesses may be avoided. The HMGLM is also defined, in particular, as an extension of the Rating Scale and Partial Credit Models. A small simulation study is described to illustrate the accuracy of the parameter recovery for these models. Additionally, modeling extensions of the Rating Scale and Partial Credit Models are made by applying the HMGLM. Computational examples are provided to illustrate the application of these models.

Dedicated to my parents Alicia and Jesse for their constant support and love.

ACKNOWLEDGMENTS

First of all, I would like to thank my dissertation committee, for the freedom they allowed me, their support, constructive criticism, and insightful comments that made this dissertation a much better project. But most importantly, thank you Dr. Maier for accepting to be the chair of the committee and taking me on as your first student. Your direction, assistance, and support helped guide me through this dissertation.

Thank you Dr. Floden. Your deep thoughts pushed me to understand my topic in more meaningful ways. Without this my dissertation would have simply been over 100 pages of formulas without any real meaning.

Thank you Dr. Reckase for not only pushing me to think deeper about the psychometrics contained in this paper, but also thank you for the 5 years of guidance and wisdom you offered while I was at Michigan State University. Without this, I would have been just another MQM student.

Thank you Dr. Wolfe. Your friendship, guidance, and support for the past several years—from the University of Florida to Michigan State University (and to wherever life leads me)—have not only made me a better student, a better professional, and a better leader, but you have also made me a better person as well. For this I am extremely grateful, and for this you will always be the 'Master' and I will always be the 'Grasshopper.'

Second, I would like to thank my friends and family who helped motivate and support me. Especially, I would like to thank my parents Alicia and Jesse, my brothers

Jeff and Jesse, my little sister Jessica, my ole buddies Wayne and Way, and my New Mom Joyce for always being there. There is nothing like friends and family.

Lastly, I would like to thank my wife Margaret, my dogs Symbi and Isa, my dog in heaven Cream, and my baby boy on the way Eian. Although they cannot read or understand most of what I say, I am forever indebted to my dogs Symbi and Isa for they always provided me with a smile and love, unconditionally, when I needed it the most. To Cream: although you were not able to see me finish my dissertation and school, you were always there to distract me and pursue the finer things in life. Thank you. To my boy Eian: You are the main reason I finished my dissertation in one year and not five. Daddy is looking forward to the new chapter in his life (and Daddy has to pay for those toys). To Margaret: throughout my graduate career, especially when I was down on myself, you provided me with the support I needed; you provided me with the friendship I needed; you provided me with the love I needed, always. Thank you.

I did it.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1. Introduction

1-1. Motivation of the study

In recent years, educational researchers have combined the theory and methods of Hierarchical Linear Modeling (HLM; Goldstein, 2003; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) and Item Response Theory (IRT; Lord, 1980). For example, Kamata (1998, 2001), Maier (2000, 2001), Fox and Glas (1998), and Adams and Wilson (1996) used the HLM framework to define IRT models for dichotomously scored items. As they illustrate, one advantage of unifying HLM and IRT methods is that postulating IRT models becomes increasingly flexible. For example, traditional IRT models (e.g., 1-parameter model; Lord 1980) may be formulated to include covariates (Cheong & Raudenbush, 2000; Fox, In press, a; Kamata, 1998, 2001).

Another advantage of unifying IRT and HLM is that the IRT parameters and their standard errors may be estimated more precisely (Maier, 2000, 2001, 2002; Mislevy, 1987). That is, by applying the HLM framework, a Level-1 model is defined in which the item parameters in an IRT model are fixed and nested within a Level-2 model. The Level-2 model defines the person parameters as being randomly varying. By considering the nested relationship—an item level nested within a person level, the variation of the responses within persons and between persons is taken into consideration, and estimation methods may obtain better precision.

Unfortunately, with these advantages, a few disadvantages follow. For instance, although the aforementioned IRT models were suitable for items that were scored dichotomously, they were not suitable for items that were scored using partial credit (i.e., polytomous items). To compensate for this limitation, Adams and colleagues (Adams &

Wilson, 1996; Adams et al., 1997), Maier (2000, 2002), Patz and colleagues (Patz, 1996 as cited by Patz, Junker, and Johnson, 1999; Patz, Junker, and Johnson, 1999; Patz, Junker, Johnson, & Mariano, 2002), Donoghue and Hombo (2003), Rijmen, Tuerlinckx, De Boek, and Kuppens (2003), and Tuerlinckx and Wang (2004) developed IRT models using a hierarchical framework for polytomous items. However, these models were limited in at least one of two ways.

The first limitation was that it did not allow for modeling of predictor variables to help explain the variation in the item and person parameters (e.g., Donoghue & Hombo, 2003; Patz, 1996 as cited by Patz, Junker, and Johnson, 1999; Patz, Junker, and Johnson, 1999; Patz, Junker, Johnson, & Mariano, 2002). As mentioned above, it may be important to control for the influences of predictor variables in a psychometric testing environment (Cheong & Raudenbush, 2000; Fox, In press, a; Kamata, 1998, 2001). Although Adams et al.'s model may include predictor variables for the person parameter, to date, their model may not include predictors of item behaviors. In addition, although Maier's model (2000, 2002) may be extended to include predictor variables (e.g., Fox, In press, a), the ease at which this may be accomplished may be arguable. If a researcher believes that person covariates and predictors of item behaviors should be controlled for, then a more flexible model is not only desired but should be employed.

The other limitation was that the correlation between categories of a polytomous item may not be sufficiently accounted for in the model (e.g., Adams et al., 1997; Donoghue & Hombo, 2003; Maier, 2000, 2002; Patz, 1996 as cited by Patz, Junker, and Johnson, 1999; Patz, Junker, and Johnson, 1999; Patz, Junker, Johnson, & Mariano, 2002). That is, the aforementioned model treats the item response as being sampled from

a univariate distribution. However, in some cases, the categories of an item merely represent nominal variables; that is, the categories are simply labels. For example, an item with the categories 'negative', 'neutral', and 'positive', may be considered as three separate dichotomous, indicator variables labeled 'negative feeling', 'neutral feeling', and 'positive feeling', each with the possibilities 'yes' or 'no'. Viewed this way, each category represents a variable, and the response itself is a vector of 0s and a 1, and should be treated as if being sampled from a multivariate distribution (Fahrmeir & Tutz, 2001).

Below, a general framework is proposed that uses HLM to model various IRT models. This is accomplished by applying a multivariate generalized linear modeling framework within HLM. The model and framework is relatively new and is commonly seen in the statistical literature under the heading of 'Multivariate Generalized Linear Mixed Model' (MGLMM; e.g., Fahrmeir & Tutz, 2001; Gueorguieva, 2001; Hartzel, Agresti, & Caffo, 2001). Here, to be consistent with the majority of the educational literature, rather than describing the model as being 'mixed', the model is described as 'hierarchical' and label it a Hierarchical Multivariate Generalized Linear Model (HMGLM).

Additionally, although Tuerlinckx and Wang (2004) recently illustrated the application of the MGLMM to IRT models and although it can be shown that the models they define are similar to those that are defined here (in particular those in Chapter 3), the focus of this dissertation, unlike the aforementioned studies, is to expand IRT models using a particular framework—the hierarchical framework set forth by Goldstein (2003), Raudenbush and Bryk (2002), and Snijders and Bosker (1999): HLM. And, unlike Tuerlinckx and Wang (2004), HLM is used to expand IRT models by conceptualizing the

units that are measured (e.g., persons and items) as being nested within one another (see Chapter 2). Furthermore, this provides a more 'natural' way for conceptualizing hierarchical polytomous IRT models. Therefore, by using the HLM framework to apply the HMGLM to IRT, readers may better see the hierarchical relationships that exist in educational testing data.

However, the purpose of applying the HMGLM to IRT and HLM is not necessarily to develop an alternative framework for modeling and estimating IRT models *per se*; rather, the purpose of applying the HMGLM is to develop a framework in which the IRT models may be extended in various ways, such as adding person covariates and predictors of item behaviors. Specifically, the advantages of using the framework provided by the HMGLM are that (1) both of the aforementioned limitations are avoided, i.e., polytomous IRT models may be extended to include person covariates and predictors of item behaviors, and the correlation between categories of a polytomous item may be accounted for; (2) models using the HMGLM may currently be estimated using existing software (e.g., SAS, 2001; STATA, 2000); (3) IRT and HLM are unified using a common notation; (4) score functions and information matrices (which may be used for parameter estimation) are well-known under the HMGLM (e.g., see Fahrmeir & Tutz, 2001); and (5) a broad class of IRT models within the HLM framework may be estimated using a common method (e.g., maximum likelihood).

This paper consists of seven chapters. In Chapter 1, the motivation for unifying HLM and IRT are discussed, and two limitations with the current IRT models within the HLM framework already are identified. In addition, Chapter 1 describes four approaches for unifying HLM and polytomous IRT models, as well as the limitations associated with

each approach. Chapter 2 provides a detailed description of a new approach for unifying HLM and polytomous IRT models. This new approach applies a hierarchical multivariate generalized linear model. In addition, Chapter 2 presents a re-formulation of two polytomous IRT models, the Rating Scale Model (Andrich, 1978) and the Partial Credit Model (Masters, 1982), using the hierarchical multivariate generalized linear model. Chapter 3 provides a simulation study for the parameter recovery of these models, as well as an example analysis for illustrating the use and interpretation of the models. Chapter 4 simulates and illustrates the application of the hierarchical multivariate generalized linear model in which the Rating Scale Model is extended to include person covariates. Chapter 5 simulates and illustrates the application of the hierarchical multivariate generalized linear model in which the Rating Scale Model is extended to include a group level as a measure of DIF. Chapter 6 simulates and illustrates the application of the hierarchical multivariate generalized linear model in which the Rating Scale Model is extended to include item covariates to explain DIF. Finally, Chapter 7 discusses the general contributions of the hierarchical multivariate generalized linear model, both methodologically and substantively, to the fields of HLM, IRT, and educational research.

1-2. Overview of Previous Hierarchical IRT Models for Polytomous Items

As Kamata (2001) points out, the unification of IRT and HLM occurred several years ago across three separate fields: psychometrics (e.g., Adams et al., 1997), non-linear mixed-effects modeling methods (e.g., Hedeker & Gibbons, 1993, as cited by Kamata, 2001), and random-effect Bayesian modeling (e.g., Spiegelhalter, Thomas, Best, & Gilks, 1996, as cited by Kamata, 2001). Since each field essentially conducted their

work independently of one another, each pursued the unification using different perspectives. Kamata (1998, 2001) continued this tradition by using a generalized linear modeling approach in HLM. Below, each perspective is discussed in relation to IRT models for polytomous items.

However, before this endeavor is pursued, one first briefly describes two traditional, non-hierarchical IRT models for polytomous items: Masters' (1982) Partial Credit Model (PCM) and a special case of the PCM, the Rating Scale Model (Andrich, 1978). By doing so, the reader may recognize the transition that is made from modeling non-hierarchically to modeling hierarchically, and the reader may notice the similarities and differences between the current hierarchical IRT models for polytomous items. Furthermore, these models and each perspective are discussed below using a common example within a typical testing condition to illustrate how the concepts of IRT transfer over to HLM.

1-2-1. Traditional, Non-Hierarchical Partial Credit and Rating Scale Models

Masters' (1982) Partial Credit Model (PCM) defines the probability π_{ijk} that person k will respond to category i of item j as

$$\pi_{ijk} = \frac{\exp \sum_{i=0}^{i'} \left(\theta_k - \delta_{ij} \right)}{\sum_{i=0}^{I} \exp \sum_{i=0}^{i'} \left(\theta_k - \delta_{ij} \right)}, \qquad (1.1)$$

where θ_k is the location of person k on the underlying latent trait continuum; and δ_{ij} is the location of a particular category i $(i = 0,1,\ldots,i',\ldots I)$ for item j on the underlying latent trait continuum.

The PCM may be re-expressed in terms of logits; that is, as a model that describes the log-odds of the probability that person k will select category i rather than category $i-1$ for item j

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \theta_k - \delta_{ij}. \tag{1.2}$$

Although θ_k and δ_{ij} may take on several different interpretations depending on the testing environment (for example, in achievement testing θ_k is commonly referred to as proficiency), here a personality testing environment is assumed, and one continues with the example given in Section 1-1 in which each item contains three categories, 'negative', 'neutral', and 'positive'. The personality test attempts to measure the latent trait 'honesty' of each particular applicant. This is achieved by asking various types of honesty questions, in which the applicant responds by selecting one of the three categories, which represents his/her feelings toward the question. Hence, in our example, θ_k is the honesty of applicant k; and δ_{ij} is the 'attractiveness' of a particular category i, or feeling i, rather than $i-1$ for each question j.

Thus, in a testing environment, the PCM suggests that the probability that a person will select a particular category of a particular item depends not only on the person's location on the underlying latent trait continuum (in this case, honesty), but also

it depends on the item's category location on the underlying latent trait continuum (in this case, the attractiveness of each feeling for each item).

Notice that the traditional model does not consider the hierarchical relationship that exists between persons and items. To help illustrate this idea, it may be better to think of persons as being schools and items as being students. Using this example, it is easier to see that a set of students is nested within a particular school. Furthermore, if the same test was given to the students across the different schools, it seems reasonable to expect that student performance on the test would be more homogenous within a particular school, and, generally speaking, the performance of a school may be more heterogeneous than another school (e.g., school in a higher SES location may perform differently than a school in a lower SES location).

Thus, referring back to our original honesty example, it seems reasonable to argue that items are nested within persons. Hence, it seems reasonable that a particular person's set of responses will be more homogeneous than when compared to a set of responses for another person. Furthermore, it seems reasonable that overall a person's responses are heterogeneous when compared to another person's responses. Therefore, the traditional RSM and PCM do not consider the variation of the responses within persons and between persons. Hence, in HLM terms, θ_k and δ_{ij} do not vary across the person or item level and are considered fixed parameters. In other words, there is no Level-1 model for the items that is defined within a Level-2 model for the persons.

Continuing then, Andrich's (1978) Rating Scale Model (RSM), may be considered a special case of the PCM (as mentioned above). To obtain the RSM, the PCM is first re-expressed to model the overall location of each item on the underlying

latent trait continuum and the response threshold of selecting category i rather than $i-1$ (instead of modeling the item's category location on the underlying latent trait continuum as before), one obtains

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right)=\theta_k-\delta_j-\tau_{ij}, \tag{1.3}$$

where θ_k is given above; but now δ_{ij} is decomposed into two components, i.e., $\delta_{ij}=\delta_j+\tau_{ij}$, where δ_j is the overall attractiveness of item j ($\delta_j=\frac{1}{I}\sum_{i=1}^{I}\delta_{ij}$); and τ_{ij} is the response threshold of being attracted to category i rather than $i-1$, and are deviations from the overall attractiveness of item j (δ_j).

However, if the category thresholds are constrained to be equal across items, i.e., $\tau_{ij}=\tau_i$, then RSM may be considered a special case of the PCM

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right)=\theta_k-\delta_j-\tau_i, \tag{1.4}$$

where δ_j is defined above; and τ_i is the threshold of being attracted to category i rather than $i-1$ for all items.

Thus, in our example, the RSM suggests that the probability that a person will be attracted to select a particular feeling for a particular item depends not only on the person's honesty, but also the overall attractiveness of the item and the threshold of being attracted to feeling i rather than $i-1$. Again, notice now that the thresholds do not vary for each item; rather, the thresholds are common across items.

Additionally, notice like the PCM, the RSM does not consider the variation of the responses within persons and between persons. Hence for the traditional PCM and RSM,

the hierarchical nature is ignored, and all parameters are considered fixed parameters. That is, in HLM terms, there is no Level-1 model for the items that is defined within a Level-2 model for the persons.

(As an aside, note the PCM and RSM are also appropriate for modeling dichotomous items, in which the dichotomous response is treated as being two categories (i.e., the 1-parameter model). Lastly, similar relationships hold for the hierarchical analog of the RSM.)

1-2-2. Random Coefficients in a Multinomial Model Approach

One approach for modeling IRT models in HLM was spearheaded by Adams and Wilson (1996) and Adams et al. (1997). In their approach, they applied a multinomial model that incorporated random coefficients for the modeling of the person's location on the underlying continuum. Specifically, the Level-1 model for their aptly named Multidimensional Random Coefficient Multinomial Model (MRCMM) is defined as

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \eta_{ijk} = \mathbf{b}'_{ij}\boldsymbol{\theta}_k + \mathbf{a}'_{ij}\boldsymbol{\xi}, \tag{1.5}$$

where π_{ijk} is defined above; $\mathbf{b}'_{ij}$ is a vector of scores for the vector of multiple dimensions $(\boldsymbol{\theta}_k)$ for person k; and $\mathbf{a}'_{ij}$ is a design vector for the set of item parameters $(\boldsymbol{\xi})$, i.e., δ_j and τ_i. Notice that the item parameters $(\boldsymbol{\xi})$ may be considered fixed.

The Level-2 model specifies the random distribution of $\boldsymbol{\theta}_k$, which may linearly depend on predictor variables (e.g., SES, gender, etc.)

$$\boldsymbol{\theta}_k = \mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k, \tag{1.6}$$

where $\mathbf{x}_k$ is a vector for the covariate scores; $\boldsymbol{\beta}$ is matrix for the fixed regression coefficients for the covariates; and $\varepsilon_k \sim N\left(0, \sigma_\varepsilon^2\right)$.

If the model is constrained to be unidimensional (Adams & Wilson, 1996), and constraints are placed on the item parameters (ξ), then Adams et al. (1997) have shown this model to be a hierarchical generalization of the PCM (e.g., see Rijmen et al. (2003)) and RSM (as well as a generalization for the 1-parameter model, c.f., Lord, 1980; Adams & Wilson, 1996). Additionally, Adams and colleagues (Wang, Wilson, & Adams, 1998) showed that the MRCMM is a generalization of the models proposed by Andersen (1985) and Embretson (1991), in which covariates were used to measure change (in the person parameter θ_k).

Continuing our example then, the MRCMM suggests that the probability that an applicant will be attracted to select a particular feeling for a particular item depends not only on the applicant's honesty, but also the overall attractiveness of the item and the threshold of being attracted to feeling *i* rather than $i-1$. Additionally, if the researcher has reason to believe that the applicant's honesty may be influenced by other variables, such as his or her criminal history or the number of occasions he or she has taken the test, then these covariates may be controlled for as well (Equation (1.6)).

Furthermore, unlike the traditional PCM and RSM, the random coefficients in a multinomial model considers the variation of the responses within persons and between persons. This is seen in the Level-1 and -2 models (Equations (1.5) and (1.6)) when the item parameters (δ_j and τ_i) are treated as fixed effects and are nested within the random effect of persons (θ_k).

Unfortunately, as mentioned above, currently the MRCMM is limited in that the software for estimating the parameters (i.e., ConQuest, 1998), may only estimate models that contain predictor variables at the person level model, and the MRCMM may not be applied when modeling predictor variables for the item parameters; nor may they be applied when controlling for the correlated relationships of the multivariate response vectors.

1-2-3. Bayesian Modeling of Random-Effects Approach

Another approach for modeling polytomous IRT models in HLM was proposed by Maier (2000, 2002) and Fox (In press, b). In the approach, Bayesian procedures are applied to the modeling of the random effects of the PCM, which may be represented as a Means-as-Outcomes model in the HLM framework (Maier, 2000, 2002; Raudenbush & Bryk, 2002). Specifically, in logit form, Maier's model is given by

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \eta_{ijk} = \theta_{rk} - \delta_{ij}, \tag{1.7}$$

where π_{ijk} and δ_{ij} is the PCM parameterization of δ_j and τ_i; and θ_{rk} is the ability of person *i* for response set *r*. Note δ_{ij} is treated as a fixed parameter, and is interpreted as a location of a particular category *i* for item *j* on the underlying latent trait continuum.

The Level-1 and -2 models specify the hierarchical nature of θ_{rk}. Specifically, the Level-1 model states

$$\theta_{rk} = \alpha_k + \varepsilon_{rk}, \tag{1.8}$$

where ε_{rk} is the random error associated with the random intercept α_k of person k for response set r, $\varepsilon_{rk} \sim N\left(0, \sigma_\varepsilon^2\right)$

The Level-2 model defines α_k. It is given as

$$\alpha_k = \boldsymbol{W}_k'\boldsymbol{\gamma} + \nu_{0k}, \tag{1.9}$$

where $\boldsymbol{W}_k' = \left(1, W_{1k}, \ldots, W_{p-1,k}\right)$ is a matrix containing the p predictor variables; $\boldsymbol{\gamma}' = \left(\gamma_0, \gamma_1, \ldots, \gamma_{p-1}\right)$ is a matrix containing the fixed regression coefficients for the p predictor variables; and ν_{0k} is the random error associated with the fixed regression coefficients $\boldsymbol{\gamma}$ for person k, $\nu_{0k} \sim N\left(0, \sigma_\nu^2\right)$.

Referring back to the example, the Bayesian modeling of random-effects approach models applicant behavior similarly to the random coefficients in a multinomial model approach, so the concepts will not be repeated here. However, one of the primary differences between the two approaches (and the traditional RSM) is that the Bayesian approach specifically models the variation of the responses within persons in the Level-1 model (i.e., ε_{rk} in Equation (1.8)), and it specifically models the variation of the responses between persons in the Level-2 model (i.e., ν_{0k} in Equation (1.9)).

Unfortunately, the Bayesian modeling of random-effects approach does not adequately account for the correlated relationships of the multivariate response vectors. Additionally, the estimation of parameters using a fully Bayesian approach requires specification of a prior distribution. However, as models become more complex (which is the case as one includes predictor variables), an inappropriate choice for the prior distribution may lead to an improper posterior distribution, which may not be detected by

MCMC methods. Also, some researchers may not accept the fully Bayesian perspective and may believe in applying other theoretical perspectives, e.g., a frequentist perspective.

1-2-4. Rater Effects Approach

A third approach, a rater effects approach, was developed by Patz and colleagues (Patz, 1996 as cited by Patz, Junker, and Johnson, 1999; Patz, Junker, and Johnson, 1999; Patz, Junker, Johnson, & Mariano, 2002). It is fairly different from the previous approaches in that it applies a generalizability framework within an HLM framework to obtain a 'rater effect'. Specifically, the approach is given by the Hierarchical Rater Model (HRM), which is essentially a 3-Level model in which the ratings of a rater are nested within item responses, which in turn is nested within a person's location on the underlying continuum.

Specifically, at Level-1 (which Patz and colleagues describe as the first stage model), the model is defined by

$$\begin{aligned} \log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) &= \text{logit}\left[P\left(\zeta_{jk}=\zeta \middle| \theta_k, X_{jkm}\in\{\zeta,\zeta-1\}\right)\right] \\ &= \eta_{ijk} \\ &= \theta_k - \delta_j - \tau_{\zeta j}, \end{aligned} \tag{1.10}$$

where ζ_{jk} is an ideal, unobserved, latent trait rating variable that describes person k's performance on item j, which follows (any IRT model, but in this case) the PCM (where δ_{ij} is decomposed into two components, i.e., $\delta_{ij}=\delta_j+\tau_{ij}$, such that δ_j is the overall attractiveness of item j ($\delta_j=\frac{1}{I}\sum_{i=1}^{I}\delta_{ij}$); and τ_{ij} is the response threshold of being attracted to category i rather than $i-1$, and are deviations from the overall attractiveness

of item j (δ_j)); and X_{jkm} is the signal detection model (see, e.g., Table 1) for rater m who rates person k on item j, which follows the Level-2 model described below. Note $\tau_{\zeta j}$ now describes the threshold of the ideal, latent rating ζ for item j, rather than the observed rating i. Also, note that δ_j and $\tau_{\zeta j}$ are considered fixed effects.

Table 1. The Signal Detection Model for the Rating Probabilities $\left(p_{\zeta im}\right)$

		i			
		0	1	...	I
ζ	0	p_{00m}	p_{01m}	⋯	p_{0Im}
	1	p_{10m}	p_{11m}	⋯	p_{1Im}
	...	...	...	...	...
	I	p_{I0m}	p_{I1m}	⋯	p_{IIm}

Note. $p_{\zeta im}$ is the probability that rater m rates the observed rating i given the ideal rating ζ.

The Level-2 model (which Patz and colleagues described as the second stage model) describes the relationship between one or more raters' rating i and the ideal rating category ζ_{jk} $\left(\zeta = 0,1,\ldots,I\right)$. The model is a discrete signal detection problem using a matrix of rating probabilities $p_{\zeta im} \equiv P\left(\text{rater } m \text{ rates } i \middle| \zeta_{jk}\right)$, as seen in Table 1. Although the density of $p_{\zeta im}$ for each row in Table 1 make take any form, Patz and colleagues used a normal density (see Patz, Junker, and Johnson (1999) for the parameterization of the normal density).

Finally, the Level-3 model (which follows from the HLM framework) defines θ_k as a random effect that is distributed as $N\left(\mu, \sigma_\theta^2\right)$.

To better understand the rater effects approach, the personality testing example is referred to again. Recall in this example, that we have an applicant whom is responding

to an honesty exam, in which each item asks the applicant to select one of three categories, negative, neutral, or positive. However, instead of the applicant selecting the categories, for the rater effects approach, the applicant is asked to provide a response to the open-ended question. And, for this response, a rater (or multiple raters) is asked to rate the applicant's response for each item as being in one of the aforementioned categories. Thus, the rater effects approach suggests that the probability that an applicant will fall into a particular category of feeling for a particular item depends not only on the applicant's honesty, but also the overall attractiveness of the item, and the threshold that a rater assigns a particular feeling i rather than $i-1$ for each question j.

Additionally, unlike the previous approaches, the rater effects approach models the variation of the responses within persons and between persons by applying a generalizability approach. Specifically, this approach attempts to measure the nested effect of the rater's ratings on the person's item responses (see, e.g., the Level-2 model depicted in Table 1). Additionally, as mentioned above, this effect is nested within the Level-3 model, the person level model, which models the variation of the responses between persons as random effects (θ_k).

Unfortunately, although the rater effects approach effectively estimates the rater effect for simulated data (e.g., Donoghue & Hombo, 2003; Patz, Junker, and Johnson, 1999; Patz, Junker, Johnson, & Mariano, 2002), the approach does not consider the modeling of predictor variables for persons and items, and the approach does not adequately account for the correlated relationships of the multivariate response vectors. Additionally, researchers report that, when compared to non-hierarchical rater effects models, the precision of estimates afforded by the HLM framework was not observed

when applying the model to real data. Also, researchers complained that the estimation of the parameters was relatively "labor and time intensive" (Barr & Raju, 2003, p.41).

1-2-5. A Hierarchical, Univariate General Linear Model Approach

The last approach discussed here for modeling polytomous IRT models in HLM essentially extends the work of Kamata (1998, 2001), which proposed using a hierarchical, univariate generalized linear model (GLM) to parameterize an IRT model for dichotomous items (i.e., the 1-parameter model, Lord, 1980). To illustrate the approach, the models are first defined using the notation typically applied in hierarchical GLM. Then, the parameters are described in terms of how the model relates to the traditional IRT parameters.

The hierarchical, univariate GLM approach is defined by applying a multinomial model using a baseline-category logit link function (Raudenbush & Bryk, 2002). The reason for doing so is to illustrate the equivalence between the adjacent-category link function and the baseline-category link function, which is used in the popular text by Raudenbush and Bryk (2002) and briefly noted by Rijmen et al. (2003).

Specifically, the Level-1 model uses a regression-type formulation, and is defined by

$$\begin{aligned} \log\left(\frac{\pi_{ijk}}{\pi_{Ijk}}\right) &= \eta_{ijk} \\ &= \sum_{q=1}^{Q} \beta_{qijk} X_{qjk}, \end{aligned} \tag{1.11}$$

where π_{ijk} is the probability that the observed response of person k on item j falls in category i; π_{Ijk} is the probability that the observed response of person k on item j falls in the 'baseline' category I; X_{qjk} is the q^{th} dummy variable for person k, with values 1 when $q = j$, and 0 when $q \neq j$ for item j; and, for person k, β_{qijk} is the regression coefficient of category i for item j. Thus, for the Level-1 model, the category level model, the regression coefficient of category i for item j (β_{qijk}) measures the overall effect (i.e., mean effect) of category i for item j, which one may notice is assumed to be fixed for each category of each item (i.e., there are no random effects added to the Level-1 model).

To model how the category effects behave across items, the Level-2 model, the item level model, is defined. Specifically, for the PCM, the Level-2 model may be defined as

$$\begin{aligned}
\beta_{q0jk} &= \gamma_{q0jk} + \sum_{i=0}^{I-1} \gamma_{1ijk} w_{1ijk} \\
\beta_{q1jk} &= \gamma_{q0jk} + \sum_{i=0}^{I-1} \gamma_{1ijk} w_{1ijk} \\
&\cdots \\
\beta_{q,I-1,jk} &= \gamma_{q0jk} + \sum_{i=0}^{I-1} \gamma_{1ijk} w_{1ijk}
\end{aligned} \tag{1.12}$$

where, for person k, γ_{q0jk} is mean effect of item j across categories i; γ_{1ijk} is the effect of item j on a particular category i; and w_{1ijk} is a dummy variable with values 1 if $i' = i$ for the j^{th} item answered by person k, and 0 otherwise.

In contrast, for the RSM, it is assumed that the effect of item j on a particular category i is equal for all items; hence, the constraint that $\gamma_{1ijk} = \gamma_{1i1k} = \ldots = \gamma_{1iJk} = \gamma_{1ik}$ is made, and the Level-2 model for the RSM becomes

$$\beta_{qijk} = \gamma_{q0jk} + \sum_{i=0}^{I-1} \gamma_{1ik} w_{1ik} , \tag{1.13}$$

where γ_{q0jk} is defined above; γ_{1ik} is the effect of item j on a particular category i, which is common across the j items; and w_{1ik} is a dummy variable with values 1 if $i' = i$ for the j^{th} item answered by person k, and 0 otherwise.

Continuing with the RSM (where analogous definitions apply to the PCM), the Level-3 model, the person level model, models how the aforementioned effects behave at the person level. Specifically, the Level-3 model is defined as

$$\gamma_{q0jk} = \lambda_{q0j0} + u_{qjk} \tag{1.14}$$

$$\gamma_{1ik} = \lambda_{1i0} \tag{1.15}$$

where λ_{q0j0} is the mean effect of persons on item j; u_{qjk} is the unique, random effect of person k (i.e., u_{qjk} is the deviation of person k from the fixed, category intercept $\left(\lambda_{q0j0}\right)$); and λ_{1i0} is the mean change in λ_{q0j0} for a particular category i, for all persons.

However, in a testing environment, it is assumed that the unique effect $\left(u_{qjk}\right)$ of person k does not vary across the categories of an item j. Hence, the effects are constrained to be equal for each category i of each item j, i.e.,

$u_{qjk} = u_{q1k} = \ldots = u_{QJk} = u_k$, $u_k \sim N\left(0, \sigma_u^2\right)$. And, the Level-3 model for γ_{q0jk} becomes

$$\gamma_{q0jk} = \lambda_{q0j0} + u_k, \tag{1.16}$$

where γ_{q0jk} and λ_{q0j0} are defined above; and u_k is the random effect of person *k*.

Thus, for the person level model, the mean effect of category *i* for item *j* $\left(\lambda_{q0j0}\right)$ varies for each item, but is fixed for each person *k*. And, the unique effect of person *k* $\left(u_k\right)$ on the mean effect of category *i* for item *j* is constant for each λ_{q0j0}. Lastly, the effect of the item on a particular category *i* $\left(\lambda_{1i0}\right)$, varies for each category *i*, but is fixed for each person *k* (and constant across the *j* items).

However, the baseline category parameterization implies that the regression coefficient of category *i* for item *j* is the mean effect of category *i* for the j^{th} item from the baseline category *I*, i.e.,

$$\beta_{qijk} = \tilde{\beta}_{qijk} - \tilde{\beta}_{qIjk}, \tag{1.17}$$

But, rather than a baseline category parameterization (such as that discussed by Raudenbush and Bryk, 2002), popular polytomous IRT models apply an adjacent-category parameterization (i.e., Agresti, 1996, 2002; Andrich, 1978; Hartzel et al., 2001; Masters, 1982; Wright & Masters, 1982), e.g., see the RSM in Equation (1.4). Therefore, the correct effect of interest is not the effect of category *i* for the j^{th} item from the baseline category *I* (Equation (1.17)); rather, the correct effect is the effect of category *i* for the j^{th} item from the adjacent category $i - 1$

$$\beta^{*}_{qijk} \equiv \tilde{\beta}_{qijk} - \tilde{\beta}_{q,i-1,jk}. \tag{1.18}$$

This implies that to obtain the adjacent-category effect $\left(\lambda^{*}_{qij0}\right)$ from the baseline-category parameterizations, one must do the following:

$$\begin{aligned} \beta^{*}_{qijk} &\equiv \tilde{\beta}_{qijk} - \tilde{\beta}_{q,i-1,jk} \\ &= \left(\tilde{\beta}_{qijk} - \tilde{\beta}_{qIjk}\right) - \left(\tilde{\beta}_{q,i-1,jk} - \tilde{\beta}_{qIjk}\right) \\ &= \beta_{qijk} - \beta_{q,i-1,jk}. \end{aligned} \tag{1.19}$$

Taking the equations above, this suggests the following. The mean effect of category i from the adjacent category $i-1$ $\left(\lambda_{q0j0}\right)$, in the HLM framework, is analogous to (the negative of) the location of a particular category i for item j on the underlying latent trait continuum $\left(-\delta_j\right)$, in the IRT framework. Additionally, the effect of the item on a particular category i $\left(\lambda_{1i0}\right)$, in the HLM framework, is analogous to (the negative of) the threshold of a particular category i $\left(-\tau_i\right)$, in the IRT framework. Lastly, the location of person k on the underlying latent trait continuum $\left(\theta_k\right)$, in the IRT framework, is analogous to the unique effect of person k $\left(u_k\right)$. In short, the parameters for the traditional RSM are equivalent to the parameters in the hierarchical GLM in the following manner:

$$\delta_j = -\lambda_{q0j0} \tag{1.20}$$

$$\tau_i = -\lambda_{1i0} \tag{1.21}$$

$$\theta_k = u_k. \tag{1.22}$$

Therefore, in the personality testing example, the hierarchical GLM approach is very similar to the random coefficients in a multinomial model approach in that the probability that an applicant will be attracted to a particular feeling for a particular item depends not only on the applicant's honesty, but also the attractiveness that an applicant will select a particular feeling i rather than $i-1$ for each question j. However, rather than modeling the parameters directly like the random coefficients approach, the hierarchical GLM approach models effects—that is, the overall attractiveness of an item as well as the effect of the item on a particular category (i.e., the Level-2 model; Equations (1.12) or (1.13)), while the honesty of an applicant is modeled using a unique effect that is treated as random at the Level-3 model (Equation (1.16)).

Furthermore, like the random coefficients approach, the hierarchical GLM approach can model person covariates; however, unlike the random coefficients approach, the hierarchical GLM approach can also model predictors of item behaviors. Since modeling person covariates and predictors of item behaviors are very similar for the hierarchical, univariate GLM approach and the hierarchical, multivariate GLM (which is the main focus of the paper), this discussion is left for Chapters 4, 5, and 6.

One limitation of the hierarchical univariate GLM approach is that the approach does not adequately account for the correlated relationships of the multivariate response vectors.

Chapter 2. A Hierarchical Multivariate Generalized Linear Modeling Framework for IRT

2-1. The Hierarchical Multivariate Generalized Linear Model

As stated earlier, the purpose of this paper is to develop a framework for modeling IRT models in HLM such that traditional IRT models may be extended in various manners. This framework will not only attempt to develop models that avoid the limitations of the previous models (i.e., polytomous IRT models may be extended to include person and item-specific covariates, and the correlation between categories of a polytomous item may be accounted for), but the model is also advantageous to apply because, as mentioned above, (1) models using HMGLM may currently be estimated using existing software (e.g., SAS, 2001; STATA, 2000); (2) IRT and HLM are unified using a common notation; (3) score functions and information matrices (which may be used for parameter estimation) are well-known under the HMGLM (e.g., see Fahrmeir & Tutz, 2001); and (4) a broad class of IRT models within the HLM framework may be estimated using a common method (e.g., maximum likelihood).

Using the notation typically applied in hierarchical GLM, the hierarchical models for the HMGLM, which has its roots in the multivariate framework provided by Fahrmeir and Tutz (2001), Gueorguieva (2001), and Hartzel et al. (2001), are defined. As mentioned previously the models defined here in Chapter 2 may resemble those defined recently by Tuerlinckx and Wang (2004); however, one reiterates that, unlike the aforementioned authors, the models below are defined by explicitly modeling the nested levels. Specifically, the Level-1 model defines the category level. The Level-2 model defines the item level. And, the Level-3 model defines the person level. Finally, the

combined model is defined. After the presentation of these models, the Rating Scale Model (RSM; Andrich, 1978) and Partial Credit Model (PCM; Masters, 1982) are defined within the HMGLM. For each of these definitions, to help ease the presentation, one continues with the previous honesty exam example, and one illustrates how the concepts behind each of the IRT models transfers over to the HMGLM.

2-1-1. The Level-1 Model for the HMGLM

As mentioned above, the Level-1 model for the HMGLM defines the Level-1 units, the categories of the items. To define the Level-1 model for the HMGLM, the categorical responses i ($i = 0, 1, 2, \ldots, I$) of person k ($k = 1, 2, 3, \ldots, K$) to item j ($j = 1, 2, 3, \ldots, J$) are re-expressed as a dummy-coded, multivariate response vector

$$\underset{\sim}{y}_k = \left(\underset{\sim}{y}'_{1k}, \underset{\sim}{y}'_{2k}, \underset{\sim}{y}'_{3k}, \ldots, \underset{\sim}{y}'_{Jk}\right)' \tag{2.1}$$

where

$$\begin{aligned} \underset{\sim}{y}_{1k} &= \left(y_{11k}, y_{21k}, \ldots, y_{I1k}\right)' \\ \underset{\sim}{y}_{2k} &= \left(y_{12k}, y_{22k}, \ldots, y_{I2k}\right)' \\ &\ldots \\ \underset{\sim}{y}_{Jk} &= \left(y_{1Jk}, y_{2Jk}, \ldots, y_{IJk}\right)', \end{aligned} \tag{2.2}$$

and

$$y_{ijk} = \begin{cases} 1 & \text{if response to item } j \text{ equals } i \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

Note that if the multivariate response vector $\underset{\sim}{y}_{jk}$ is a vector of 0's, then category 0 was chosen by person k for item j. Here, category 0 was chosen to be the reference category to be consistent with polytomous IRT models; however, other reference categories can be

utilized without loss of generality. Additionally, notice the multivariate response vectors are one of the primary differences between multivariate hierarchical GLM and univariate hierarchical GLM.

Another primary difference is that it is assumed that the y_{ijk} are conditionally independent given the multivariate (not univariate) random effect $\boldsymbol{u}_{jk}$. If the sum of the conditionally independent observations y_{ijk} for $\tilde{y}_{jk}$ is taken, i.e., $y_{jk} = \sum_{i=1}^{I} y_{ijk}$, then it is also assumed that $\boldsymbol{y}_{jk}$ are multinomially distributed with parameters $\boldsymbol{\pi}_{jk} = \left(\pi_{1jk}, \pi_{2jk}, \ldots, \pi_{IjK}\right)'$ (Hartzel, Agresti, & Caffo, 2001).

Thus, the conditional distribution $f\left(\boldsymbol{y}_{jk} \mid \boldsymbol{u}_{jk}\right)$ is a member of the multivariate exponential family with multivariate means $\boldsymbol{\mu}_{1k}, \boldsymbol{\mu}_{2k}, \boldsymbol{\mu}_{3k}, \ldots, \boldsymbol{\mu}_{Jk}$. That is,

$$\begin{aligned} \boldsymbol{\mu}_{1k} &= E\left(\boldsymbol{y}_{1k} \mid \boldsymbol{u}_1\right) = \boldsymbol{h}\left(\boldsymbol{\eta}_{1k}\right) \\ \boldsymbol{\mu}_{2k} &= E\left(\boldsymbol{y}_{2k} \mid \boldsymbol{u}_2\right) = \boldsymbol{h}\left(\boldsymbol{\eta}_{2k}\right) \\ \boldsymbol{\mu}_{3k} &= E\left(\boldsymbol{y}_{3k} \mid \boldsymbol{u}_3\right) = \boldsymbol{h}\left(\boldsymbol{\eta}_{3k}\right) \\ &\ldots \\ \boldsymbol{\mu}_{Jk} &= E\left(\boldsymbol{y}_{Jk} \mid \boldsymbol{u}_J\right) = \boldsymbol{h}\left(\boldsymbol{\eta}_{Jk}\right), \end{aligned} \tag{2.4}$$

where $\boldsymbol{h}\left(\boldsymbol{\eta}_{jk}\right)$ is a vector of inverse link functions

$$\begin{aligned} & h_1\left(\boldsymbol{\eta}_{jk}\right) \\ & h_2\left(\boldsymbol{\eta}_{jk}\right), \\ & \ldots \\ & h_I\left(\boldsymbol{\eta}_{jk}\right) \end{aligned} \tag{2.5}$$

where $\boldsymbol{\eta}_{jk}$ is a vector of functions that describe the linear relationship of the fixed and random parameters.

To obtain the desired form of a polytomous IRT model, the vector of inverse link functions are defined using the adjacent-category link function (Agresti, 1996, 2002) $\boldsymbol{h}\left(\boldsymbol{\eta}_{jk}\right)$

$$h_i\left(\boldsymbol{\eta}_{jk}\right) = \boldsymbol{\mu}_{jk} \equiv \pi_{ijk} = \frac{\exp\left(\sum_{i=0}^{i'} \eta_{ijk}\right)}{\sum_{i=0}^{I} \exp\left(\sum_{i=0}^{i'} \eta_{ijk}\right)}, \tag{2.6}$$

which is the probability π_{ijk} of person k selecting category i $\left(i = 0,1,\ldots,i',\ldots I\right)$ of item j, where $\eta_{0jk} \equiv 0$; hence, $\exp\left(\sum_{i=0}^{0} \eta_{ijk}\right) = \exp\left(\eta_{0jk}\right) \equiv 1$.

Re-expressing the link function as the log-odds of person k responding to category i rather than category $i-1$ for item j, the Level-1 model for the HMGLM is obtained:

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \log\left(\frac{\dfrac{\exp\left(\sum_{i=0}^{i'} \eta_{ijk}\right)}{\sum_{i=0}^{I} \exp\left(\sum_{i=0}^{i'} \eta_{ijk}\right)}}{\dfrac{\exp\left(\sum_{i=0}^{i'-1} \eta_{i-1,jk}\right)}{\sum_{i=0}^{I} \exp\left(\sum_{i=0}^{i'-1} \eta_{i-1,jk}\right)}}\right)$$

$$
\begin{aligned}
&= \log\left(\frac{\exp\left(\sum_{i=0}^{i'}\eta_{ijk}\right)}{\exp\left(\sum_{i=0}^{i'-1}\eta_{i-1,jk}\right)}\right) \\
&= \log\left(\frac{\exp\left(\eta_{0jk}+\eta_{1jk}+\ldots+\eta_{i-1,jk}+\eta_{ijk}\right)}{\exp\left(\eta_{0jk}+\eta_{1jk}+\ldots+\eta_{i-1,jk}\right)}\right) \\
&= \left(\eta_{0jk}+\eta_{1jk}+\ldots+\eta_{i-1,j,k}+\eta_{ijk}\right)-\left(\eta_{0jk}+\eta_{1jk}+\ldots+\eta_{i-1,jk}\right) \\
&= \eta_{ijk}.
\end{aligned}
\tag{2.7}
$$

Specifically, the Level-1 model for the HMGLM defines the log-odds of the probability that person *k* will select category *i* rather than category $i-1$ for item *j* as category effects

$$
\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \sum_{j=1}^{J}\beta_{jk}^{(i)}x_{jk}\,, \tag{2.8}
$$

where $\beta_{jk}^{(i)}$ is the mean category effect if person *k* selects category *i* of item *j*; and x_{jk} is a dummy variable with values 1 if person *k* answers item *j*, and 0 otherwise.

Thus, like the Level-1 model for the univariate, hierarchical GLM approach, the mean category effect $\left(\beta_{jk}^{(i)}\right)$ non-randomly varies across each category *i* of each item *j* for each person *k*. Furthermore, the mean category effect $\left(\beta_{jk}^{(i)}\right)$ is influenced by the effect of the particular item in which the categories are nested. The Level-2 model describes these effects.

2-1-2. The Level-2 Model for the HMGLM

Since the Level-1 model described the category effects for an item only if it has been answered by person k, then like the Level-1 model, the Level-2 model is defined in terms of the answered item as well. Specifically, the Level-2 model, the item-level model for the HMGLM, is generally defined as

$$\beta_{jk}^{(i)} = \gamma_{0jk} + \sum_{i=1}^{I} \gamma_{1jk}^{(i)} w_{1jk}, \tag{2.9}$$

where, for person k, γ_{0jk} is the mean effect of item j across categories i; $\gamma_{1jk}^{(i)}$ is the effect of item j on a particular category i; and w_{1jk} is a dummy variable with values 1 if $i' = i$ for the j^{th} item answered by person k, and 0 otherwise. Recall $\eta_{0jk} \equiv 0$. Thus, for identifiability, $\gamma_{1jk}^{(0)} \equiv 0$.

Thus, like the Level-2 model for the univariate, hierarchical GLM approach, the Level-2 model defines how the category effects $\left(\beta_{jk}^{(i)}\right)$ behave when they are nested within the item-level model. Specifically, the category effects vary non-randomly and depend upon the mean effect of the item across the categories $\left(\gamma_{0jk}\right)$ and the effect of the item on each category $\left(\gamma_{1jk}^{(i)}\right)$.

2-1-3. The Level-3 Model for the HMGLM

The person-level model for the HMGLM, the Level-3 model, defines how the item effects behave when nested within persons. Specifically, the Level-3 model is defined as

$$\gamma_{0jk} = \lambda_{0j0} + u_{jk}, \tag{2.10}$$

$$\gamma_{1jk}^{(i)} = \lambda_{1j0}^{(i)}, \tag{2.11}$$

where, for the j^{th} item that is answered by person k, λ_{0j0} is the mean effect of persons on item j; u_{jk} is the random effect of person k on the mean effect of item j; and $\lambda_{1j0}^{(i)}$ is the mean change in the λ_{0j0} for a particular category of item j, for all persons.

However in IRT, we assume that the person effects are constant across items. Thus, the following constraint is made

$$u_{1k} = u_{2k} = \ldots = u_{jk} = u_k,$$

and the Level-3 model for the mean item effect becomes

$$\gamma_{0jk} = \lambda_{0j0} + u_k, \tag{2.12}$$

where λ_{0j0} is defined above; and u_k is the random effect of person k across items.

Thus, like the Level-3 model for the univariate, hierarchical GLM approach, the Level-3 model defines the mean effect of the item $\left(\gamma_{0jk}\right)$ as depending upon the mean effect of the item across all persons $\left(\lambda_{0j0}\right)$, and depending upon the unique effect of a particular person k $\left(u_k\right)$. Additionally, the Level-3 model defines the effect of the item on a specific category $\left(\gamma_{1jk}^{(i)}\right)$ as being fixed for each person k $\left(\lambda_{1j0}^{(i)}\right)$.

2-1-4. The Combined Model for the HMGLM

To obtain the combined model for the HMGLM, the models for Levels 1, 2, and 3 are combined

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right)=\sum_{j=1}^{J}\left[\lambda_{0j0}+\sum_{i=1}^{I}\lambda_{1j0}^{(i)}w_{1jk}+u_k\right]x_{jk}. \tag{2.13}$$

To obtain the matrix representation of the combined model, the following matrices are defined:

$$\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_J)', \tag{2.14}$$

where

$$\boldsymbol{\beta}_j=\left(\lambda_{0j0},\lambda_{1j0}^{(1)},\lambda_{1j0}^{(2)},\ldots,\lambda_{1j0}^{(I)}\right). \tag{2.15}$$

Hence η_{jk} defines the following linear relationships

$$\begin{aligned}
\boldsymbol{\eta}_{1k}&=\mathbf{Z}_{1k}\boldsymbol{\beta}+\mathbf{W}_{1k}\boldsymbol{u}_k\\
\boldsymbol{\eta}_{2k}&=\mathbf{Z}_{2k}\boldsymbol{\beta}+\mathbf{W}_{2k}\boldsymbol{u}_k\\
\boldsymbol{\eta}_{3k}&=\mathbf{Z}_{3k}\boldsymbol{\beta}+\mathbf{W}_{3k}\boldsymbol{u}_k\\
&\cdots\\
\boldsymbol{\eta}_{Jk}&=\mathbf{Z}_{Jk}\boldsymbol{\beta}+\mathbf{W}_{Jk}\boldsymbol{u}_k,
\end{aligned} \tag{2.16}$$

where $\boldsymbol{\beta}$ is defined above and is a $(p\times 1)$-dimensional matrix for the unknown parameters (p) of the fixed effects; $\mathbf{Z}_{i1},\mathbf{Z}_{i2},\mathbf{Z}_{i3},\ldots,\mathbf{Z}_{iJ}$ are $(I\times p)$-dimensional design matrices for the fixed effects; $\boldsymbol{u}_k$ are $(p\times 1)$-dimensional matrices for the unknown parameters (p) of the random effects; and $\mathbf{W}_{1k},\mathbf{W}_{2k},\mathbf{W}_{3k},\ldots,\mathbf{W}_{Jk}$ are $(I\times p)$-dimensional design matrices for the random effects.

Lastly, the random effect $\boldsymbol{u}$ is assumed to be independent and identically distributed with density $g(\boldsymbol{u})$, which is not restricted to any form. Here the density $g(\boldsymbol{u})$ is chosen to analogously follow traditional IRT assumptions and previous formulations of hierarchical IRT models (e.g., Kamata, 1998, 2001; Lord, 1980; Miyazaki, 2000)

$$\boldsymbol{u} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{2.17}$$

(Note the dummy variable x_{jk} in the above discussion represents the situation where all persons respond to all items, i.e., the data are balanced. If the data are unbalanced, then x_{jk} may take on a similar coding scheme as that provided in Equation (1.11) for the hierarchical, univariate GLM approach. That is, x_{jk} becomes x_{qjk}, and represents the q^{th} dummy variable for person *k*, with values 1 when $q = j$, and 0 when $q \neq j$ for item *j*.)

2-2. A New Model 1: The Hierarchical Multivariate Generalized Linear-Partial Credit Model (HMGL-PCM)

To illustrate the relationship between the HMGLM and traditional IRT parameters, the PCM is defined within the HMGLM. Since the PCM is defined within the hierarchical framework of the HMGLM, the model can essentially be thought of as a new model. This new model is named the Hierarchical Multivariate Generalized Linear-Partial Credit Model (HMGL-PCM). For the HMGL-PCM, the reader should notice the application of the HLM framework (i.e., the definition of model levels), which is not used

by Tuerlinckx and Wang (2004) and provides a more natural way for conceptualizing the hierarchical PCM.

The Level-1 model (the category level) for the HMGL-PCM is defined as

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right)=\sum_{j=1}^{J}\beta_{jk}^{(i)}x_{jk}\,, \tag{2.18}$$

where all terms are defined above.

The Level-2 model (the item level) is defined as

$$\beta_{jk}^{(i)}=\gamma_{0jk}+\sum_{i=1}^{I}\gamma_{1jk}^{(i)}w_{1jk}\,, \tag{2.19}$$

where all terms are defined above.

Here, to see the relationship between the HMGL-PCM and traditional PCM, one refers back to the honesty example. Recall, that in this example, an applicant is responding to several polytomous honesty items by selecting a particular category, which represents his/her feelings toward the item. Hence, for the HMGL-PCM, the probability that an applicant is attracted to a particular feeling for a particular answered item depends upon the overall attractiveness of the item $\left(\gamma_{0jk}\right)$, and how the attractiveness of the item influences a particular feeling $\left(\gamma_{1jk}^{(i)}\right)$. Additionally, notice that the attractiveness of a feeling for an item is nested within that item, as modeled from Level-1 to Level-2.

Continuing with the HMGL-PCM, the Level-3 model (the person level) is defined as

$$\gamma_{0jk}=\lambda_{0j0}+u_k\,, \tag{2.20}$$

$$\gamma_{1jk}^{(i)}=\lambda_{1j0}^{(i)}, \tag{2.21}$$

where all terms are defined above.

The combined model for the HMGL-PCM is defined as

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right)=\sum_{j=1}^{J}\left[\lambda_{0j0}+\sum_{i=1}^{I}\lambda_{1j0}^{(i)}w_{1jk}^{(i)}+u_k\right]x_{jk}, \tag{2.22}$$

which reduces to the following for a particular category i of an item j

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right)=\lambda_{0j0}+\lambda_{1j0}^{(i)}+u_k. \tag{2.23}$$

Here we can clearly see how the category effects function as the categories are nested within items, which in turn are nested within persons. Specifically, the probability that an applicant is attracted to a particular feeling for a particular answered item not only depends upon the overall attractiveness of the item $\left(\gamma_{0jk}\right)$, but also how the attractiveness of the item influences a particular feeling $\left(\gamma_{1jk}^{(i)}\right)$. In addition, as the Level-3 model shows (Equations (2.20) - (2.21)), the overall attractiveness of the item $\left(\lambda_{0j0}\right)$ and the influence of an item on a particular feeling $\left(\lambda_{1j0}^{(i)}\right)$ is fixed across persons. Furthermore, as is commonly assumed in IRT, the unique effect $\left(u_k\right)$ of an applicant randomly varies across the different applicants (but remains fixed across items and across feelings).

In short, the parameters of the HMGL-PCM are related to the parameters of the traditional PCM in the following manner:

$$\delta_j=-\lambda_{0j0}, \tag{2.24}$$

$$\tau_{ij} = -\lambda_{1j0}^{(i)}. \tag{2.25}$$

and

$$\theta_k = u_k. \tag{2.26}$$

2-3. A New Model 2: The Hierarchical Multivariate Generalized Linear-Rating Scale Model (HMGL-RSM)

Now the RSM is defined within the HMGLM, and this new model is named the Hierarchical Multivariate Generalized Linear-Rating Scale Model (HMGL-RSM). Recall from Section 1-2-1, that the RSM is simply a special case of the PCM. Hence, the model definitions of the HMGL-RSM follow very closely to the HMGL-PCM. Again, the reader should notice the application of the HLM framework (i.e., the definition of model levels), which is not used by Tuerlinckx and Wang (2004) and provides a more natural way for conceptualizing the hierarchical RSM.

The Level-1 model (the category level) is

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \sum_{j=1}^{J} \beta_{jk}^{(i)} x_{jk}, \tag{2.27}$$

where all terms are defined above.

The Level-2 model (the item level) is obtained by constraining the effect of an item on a particular category to be equal for all items (i.e., $\gamma_{11k}^{(i)} = \gamma_{12k}^{(i)} = \ldots = \gamma_{1Jk}^{(i)} = \gamma_{1\cdot k}^{(i)}$)

$$\beta_{jk}^{(i)} = \gamma_{0jk} + \sum_{i=1}^{I} \gamma_{1\cdot k}^{(i)} w_{1\cdot k}^{(i)}, \tag{2.28}$$

where γ_{0jk} is defined above; $\gamma_{1\cdot k}^{(i)}$ is the effect of the item on a particular category, which again is equal for all items; and $w_{1\cdot k}^{(i)}$ is a dummy variable with values 1 if $i' = i$ for the j^{th} item answered by person k, and 0 otherwise.

The Level-3 model (the person level model) is

$$\gamma_{0jk} = \lambda_{0j0} + u_k, \tag{2.29}$$

$$\gamma_{1\cdot k}^{(i)} = \lambda_{1\cdot 0}^{(i)}, \tag{2.30}$$

where λ_{0j0} and u_k are defined above; and $\lambda_{1\cdot 0}^{(i)}$ is the mean change in the λ_{0j0} for a particular category i, for all persons.

In our example, the parameters of the HMGL- RSM may be interpreted accordingly. The probability that an applicant is attracted to a particular feeling for a particular answered item depends upon the overall attractiveness of the item $\left(\gamma_{0jk}\right)$, and the common influence of the attractiveness of the items on a particular feeling $\left(\gamma_{1\cdot k}^{(i)}\right)$. Additionally, the overall attractiveness of the item and the influence of the items on a particular feeling is fixed across persons (λ_{0j0} and $\lambda_{1\cdot 0}^{(i)}$, respectively). Lastly, the unique effect of an applicant on the item randomly varies across the different applicants $\left(u_k\right)$.

The combined model for the HMGL-RSM is defined as

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \sum_{j=1}^{J}\left[\lambda_{0j0} + \sum_{i=1}^{I}\lambda_{1\cdot 0}^{(i)} w_{1\cdot k}^{(i)} + u_k\right] x_{jk}, \tag{2.31}$$

which reduces to the following for a particular category i of an item j

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + u_k . \tag{2.32}$$

In short, the parameters of the HMGL-RSM are related to the parameters of the traditional RSM in the following manner:

$$\delta_j = -\lambda_{0j0}, \tag{2.33}$$

$$\tau_i = -\lambda_{1\cdot 0}^{(i)}. \tag{2.34}$$

and

$$\theta_k = u_k . \tag{2.35}$$

2-4. Assumptions

Like non-hierarchical, univariate GLM, there are distributional and structural assumptions of the HMGLM that need to be satisfied for the model to hold. As mentioned above, the distributional assumption is that the y_{ijk} are conditionally independent given the random effect $\boldsymbol{u}_k$ (i.e., $f\left(\boldsymbol{y}_{jk} \mid \boldsymbol{u}_k\right)$), and the conditional distribution $f\left(\boldsymbol{y}_{jk} \mid \boldsymbol{u}_k\right)$ is a member of the multivariate exponential family. Here, it is assumed to be multinomially distributed with parameters $\boldsymbol{\pi}_{jk} = \left(\pi_{1jk}, \pi_{2jk}, \ldots, \pi_{IjK}\right)'$.

The structural assumption is given by the Level 1 model; and, that is, the expectation of $f\left(\boldsymbol{y}_{jk} \mid \boldsymbol{u}_k\right)$ (i.e., $\boldsymbol{\mu}_{jk}$) is determined by a vector of linear predictors (Equation (2.16)) in the form of a vector of inverse link functions, $\boldsymbol{h}\left(\boldsymbol{\eta}_{jk}\right)$. For the purposes here, $\boldsymbol{h}\left(\boldsymbol{\eta}_{jk}\right)$ is chosen to be the logit form of the adjacent-categories link

function (Equation (2.6); Agresti, 1996, 2002; Hartzel, Agresti, & Caffo, 2001). Agresti (2002) shows this function to be the form of the RSM and PCM.

Regarding the hierarchical nature of the HMGLM, recall from above that the random component $\boldsymbol{u}$ requires certain distributional assumptions. One of the advantages of applying the HMGLM is that $\boldsymbol{u}$ is not restricted to be a specific distribution. For the purposes here, IRT parameters are being modeled, and, recall, $\boldsymbol{u}$ is equivalent to the location of person k on the underlying continuum. In IRT, it is customary that the locations of all persons on the underlying continuum are assumed to be normally distributed (e.g., Cheong & Raudenbush, 2000; Kamata, 1998, 2001; Lord, 1980; Miyazaki, 2000). It is also customary in HLM, to model the random components as being multivariate normally distributed (Raudenbush & Bryk, 2002). Thus, although not necessary for the HMGLM, here previous customs were followed and $\boldsymbol{u}$ was assumed to be multivariate normally distributed (Equation (2.17)).

Additionally in traditional IRT methodology, the scale of the person and item parameters is indeterminate (Lord, 1980). For the HMGLM, this is resolved in the following manner. Recall that the HMGLM begins by modeling category effects of person k on category i of item j. This suggests that $\beta_{jk}^{(i)}$ measures the effect of the category from the grand mean

$$\beta_{jk}^{(i)} = \alpha_0 + \alpha_{jk}^{(i)}, \tag{2.36}$$

where α_0 is the grand mean of the person measures; and for person k, $\alpha_{jk}^{(i)}$ is the regression coefficient for category i of item j.

Also recall that, after several hierarchical levels are modeled, the unique effect of person *k* is modeled. This suggests that the unique effect of person k is the residual of person *k* from the grand mean of the person measures

$$\beta_{jk}^{(i)} = \alpha_0 + \alpha_{jk}^{(i)} + u_k , \tag{2.37}$$

where α_0 and $\alpha_{jk}^{(i)}$ are defined above; and u_k is the unique effect of person *k*. In other words, u_k is the deviation of person *k* from α_0.

In order to resolve the indeterminacy of the scale for the HMGL-RSM and -PCM, u is assumed to be $N(0,\Sigma)$. Notice if the coefficients are assumed to be independent, this is equivalent to saying that $\beta \sim N(0,\Sigma)$. Furthermore, since the coefficients are measured effects from the grand mean, and the distribution and mean of β is chosen to be normal and zero, respectively, then this is equivalent to saying that the grand mean, which again is centered on person measures, is zero, and the distribution is normal. Therefore, this resolves the indeterminateness of the scale by centering on person measures, in which the center of the normally distributed measures is zero.

Also in IRT, it is assumed that, beyond the characteristics (i.e., parameters) of an item, success on an item only depends on the person's location on the underlying continuum $(\theta_k = u_k)$. In other words, it is assumed that the test is unidimensional—success depends on the one dimension (e.g., honesty), and not on other traits (i.e., the test is not multidimensional) (Lord, 1980). From unidimensionality, it follows that the items are assumed to be locally independent. That is, the conditional probability of success on one particular item, given the person's location on the underlying continuum, is equal to the conditional probability of success on all other items, given the person's location on

the underlying continuum (Lord, 1980). By using the HMGLM, the assumption of unidimensionality is relaxed. For example, below one presents extensions of the HMGL-RSM in which person covariates (Chapter 4) and predictors of item behaviors for the overall item location (Chapters 5 and 6) are modeled.

By modeling the aforementioned, this implies the definition of local independence is slightly altered for the HMGL-RSM and -PCM. That is, the definition of local independence is now the following: the conditional probability of success on one particular item, given the person's ability *and* the covariates, is equal to the conditional probability of success on all other items, given the person's ability *and* the covariates (c.f., the definition of local independence above).

Note local independence is satisfied for the HMGL-RSM and -PCM because the item locations are assumed to be fixed at the person level. In other words, if the item locations varied randomly or non-randomly, then the conditional probability of success on one particular item, given the person's ability *and* the covariates, would <u>not</u> necessarily equal the conditional probability of success on all other items, given the person's ability *and* the covariates. (This suggests that the HMGL-RSM and -PCM may be used to examine violations of local independence by modeling item covariates that examine how the item locations vary. Although this goes beyond the scope of this dissertation, this type of analysis is similar to those presented in the following Chapters.)

2-5. Estimation

Estimation of the parameters for the HMGL-RSM and -PCM may be accomplished using frequentist or Bayesian methods. For examples of Monte Carlo

methods see Fahrmeir and Tutz (2001) and Hartzel et al. (2001). For examples of Bayesian procedures see Fahrmeir and Tutz (2001), Fox and Glas (1998), and Maier (2000, 2002). Fortunately, if one prefers frequentist methods, then the parameters of the HMGL-RSM and -PCM may be estimated by readily available popular statistical software packages, such as SAS (using PROC NLMIXED) and STATA (using GLLAMM; Rabe-Hesketh, Pickles, & Skrondal, 2001). Specifically, estimates of the parameters are obtained by maximizing an approximation to the likelihood integrated over the random effects, where the integral approximations are obtained via adaptive Gaussian quadrature and the optimization technique is carried out using a dual quasi-Newton algorithm (SAS, 2001) or a modified Newton-Rapheson algorithm (Rabe-Hesketh, Pickles, & Skrondal, 2001). Approximate standard errors of the successfully converged parameter estimates are based on the second derivative matrix of the likelihood function (SAS, 2001) or the delta-method (Rabe-Hesketh, Pickles, & Skrondal, 2001).

Unfortunately, popular software such as PROC NLMIXED does not estimate multiple random effects. For example, for the models given above, only the person parameter (θ_k) may be considered random (u_k) while the item and category parameters (δ_j, τ_i) may be considered fixed $\left(\lambda_{0j0}^{(i)}, \lambda_{0j0}, \lambda_{1j0}^{(i)}\right)$. If one wishes to treat the item parameters as random, then one may use GLLAMM or other methods (such as MCMC or Bayesian estimation).

Chapter 3. Parameter Recovery and Example

3-1. Simulation Design

The following section describes the design for a simulation study. Specifically, observations were simulated using the RSM. Next, parameter estimates of the RSM and HMGL-RSM were obtained with Winsteps (1999) and SAS (2001), respectively. Finally, A comparison between the analyses of the parameter recovery rates follows. Because of computational constraints (i.e., see Section 7-2-3), the PCM was not simulated. However, because of the similarity between the RSM and PCM, similar results would be expected (e.g., see Section 3-3).

3-1-1. Design

The design of the simulation is as follows. Observations were simulated using the RSM. This model was chosen because it is commonly used when scaling polytomous data, such as those found in questionnaire data (e.g., Dodd, 1990; Smith & Johnson, 2000; Zhu, Updyke, & Lewandowski, 1997) and achievement data (e.g., Michigan Education Assessment Program, 2003). For the study, simulees (K = 100, 500, or 1000) responded to polytomous items (J = 10 or 25), where each item consisted of 3 categories i (i = 0, 1, 2). The number of simulees, items, and categories were chosen to follow typical data from a questionnaire (e.g., Dodd, 1990; Smith & Johnson, 2000; Zhu, Updyke, & Lewandowski, 1997) or a large-scale assessment (e.g., U.S. Department of Education, 1999).

Item parameters were also selected to represent parameter estimates from typical polytomous data. Specifically, item parameters were selected from the IRT scaling of a

confidential readiness assessment. For this assessment, there were three sub-scales that measured the personal and social development (16 items), language (12 items), and mathematical thinking (14 items) of a child. For each item, a particular scenario was observed with the child, and a rater would then proceed to score the child in one of three categories, a lower, middle, and higher category, each representing the performance of that child on that particular item. For the purposes of this dissertation, only the first 25 items were used. Table 2 displays the item parameters used in the simulation. (Note although τ_1 and τ_2 appear to be extreme, these are typical values seen in educational questionnaires because it is common in education that the middle categories, as opposed to the extreme categories, are frequently used. For example, see Dodd (1990), Smith and Johnson (2000), and Zhu, Updyke, and Lewandowski (1997).)

Table 2. Item Parameters Used in the Simulation

	RSM
Simulation Item	δ_j
1	-0.09
2	0.02
3	-0.92
4	-1.57
5	-0.81
6	-0.74
7	-0.81
8	-0.01
9	0.07
10	-0.85
11	-1.28
12	-1.02
13	-1.14
14	-1.39
15	0.54
16	-0.32

Table 2 (cont'd)

17	-0.09
18	0.11
19	-0.15
20	-0.42
21	0
22	0.51
23	0.52
24	0.73
25	0.79
τ_1	-2.24
τ_2	2.24

Note. δ_j = location for item j. $\{\tau_1$ and $\tau_2\}$ = thresholds 1 and 2.

To produce the simulated responses under the RSM, each simulee k was randomly assigned a location θ_k, $\theta \sim N(0, 1)$, and each item j was randomly assigned a set of item parameters. If $J = 10$, then the item parameters were randomly selected to be those that appear for the first 10 items in Table 2; otherwise, $J = 25$ and all items were used.

Using θ_k, δ_j, and τ_i, three response probabilities for each simulee by item combination were produced, $P_{0jk}(\theta)$, $P_{1jk}(\theta)$, and $P_{2jk}(\theta)$. If $\sum_{0}^{i'} P_{i'jk}(\theta) < Y_{jk} \leq \sum_{0}^{i'+1} P_{i'jk}(\theta)$, then simulee k was assigned a response of $i'+1$ for item j; otherwise a response of 0 was assigned. Note that $i' = 0, 1$; and Y_{jk} was a single, random number for each $j \times k$ combination, $Y \sim U(0,1)$.

The simulation procedure utilized a fully crossed 3 × 2 factorial design that simulated 6 conditions. Each administration was iterated 50 times producing 300 unique

response data matrices. The number of iterations was chosen because Kamata (1998) showed this to be a reasonable number for obtaining stable estimates. S-Plus (2000) was used to generate all data.

3-1-2. Analysis

PROC NLMIXED of SAS (2001) was used to estimate the person and item parameters for the HMGL-RSM, while WINSTEPS (1999) was used to estimate the person and item parameters for the RSM. An example of the SAS code for the HMGL-RSM is provided in Appendix A. (An example of the SAS code for the HMGL-PCM is provided in Appendix B.) An example for the input data structure is provided in Appendix C. To investigate the accuracy of the parameter estimates for the RSM and HMGL-RSM, the root mean square error (RMSE) for μ_θ, Σ, δ_j, and τ_i was obtained over the iterations for each condition. Specifically, the RMSE was obtained by

$$RMSE(\hat{\omega}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\omega}_n - \omega_n)^2}, \qquad (3.1)$$

where the maximum number of n iterations was $N = 50$; and ω is an arbitrary parameter representing either μ_θ, Σ, δ_j, or τ_i.

3-2. Parameter recovery results

Below, the descriptive statistics are presented for θ for the 50 iterations of each condition. Recall, that δ_j and τ_i were specified and shown in Table 2. Also, the results for the mean and standard deviations of the parameter estimates for 50 iterations for all

conditions are displayed and discussed. Lastly, the results of the analysis for recovering the parameters are presented.

3-2-1. Descriptive Statistics

The results of the descriptive statistics for θ and Σ of 100, 500, and 1000 persons are presented in Table 3. As can be seen, the sampling distribution of μ_θ was centered on or near zero with a small standard error (which decreased as persons increased, as would be expected). Additionally, the sampling distribution of μ_Σ was centered on or near one with a small standard error (which decreased as persons increased, as would be expected). These findings suggest that the distribution of θ was simulated very well for all conditions.

Table 3. Mean and Standard Error of θ and Σ for the Simulated 100, 500, and 1000 Persons

	θ		Σ	
K	M	SE	M	SE
100	-0.01	(0.11)	0.98	(0.06)
500	0.01	(0.05)	1.00	(0.03)
1000	0.00	(0.03)	1.00	(0.02)

Note. K = Number of simulated individuals. M = Mean. SE = Standard error.

Displayed in Tables 4 and 6, and 5 and 7 are the mean and standard deviations of the parameter estimates for the RSM and HMGL-RSM, respectively. As can be seen for both the RSM and HMGL-RSM, the standard deviations of the estimates are similar

across conditions. Furthermore, the standard deviations are fairly low and decrease as the number of persons increase. This suggests that WINSTEPS and PROC NLMIXED obtain relatively consistent estimates of the HMGL-RSM parameters.

As for the mean of the estimates, in general, the estimates obtained by WINSTEPS for the RSM appear to resemble the estimates obtained by PROC NLMIXED for the HMGL-RSM (c.f., Table 2). Below, in Section 3-2-2, the RMSE is examined.

Table 4. Mean and Standard Error of the Parameter Estimates for the RSM when J = 10

	100		500		1000	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.21	(0.28)	-0.13	(0.13)	-0.11	(0.08)
$\hat{\delta}_2$	0.01	(0.24)	-0.01	(0.10)	0.02	(0.08)
$\hat{\delta}_3$	-0.98	(0.31)	-1.03	(0.14)	-1.03	(0.09)
$\hat{\delta}_4$	-1.67	(0.28)	-1.75	(0.12)	-1.75	(0.09)
$\hat{\delta}_5$	-0.93	(0.23)	-0.93	(0.13)	-0.91	(0.08)
$\hat{\delta}_6$	-0.84	(0.23)	-0.83	(0.11)	-0.83	(0.10)
$\hat{\delta}_7$	-0.90	(0.27)	-0.91	(0.13)	-0.88	(0.09)
$\hat{\delta}_8$	-0.05	(0.23)	-0.05	(0.12)	0.00	(0.08)
$\hat{\delta}_9$	0.05	(0.25)	0.05	(0.13)	0.09	(0.08)
$\hat{\delta}_{10}$	-0.94	(0.23)	-0.94	(0.13)	-0.93	(0.08)
$\hat{\tau}_1$	-2.53	(0.12)	-2.53	(0.07)	-2.52	(0.04)
$\hat{\tau}_2$	2.53	(0.12)	2.53	(0.07)	2.52	(0.04)
$\mu_{\hat{\theta}}$	0.00	(0.01)	0.00	(0.01)	0.00	(0.00)
$\Sigma_{\hat{\theta}}$	1.36	(0.12)	1.37	(0.05)	1.36	(0.03)

Note. $\{100, 500, 1000\}$ = Number of simulated individuals. $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. $\mu_{\hat{\theta}}$ = Mean person location. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations. M = Mean. SE = Standard error.

Table 5. Mean and Standard Error of the Parameter Estimates for the HMGL-RSM when J = 10

	100		500		1000	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.18	(0.26)	-0.11	(0.12)	-0.09	(0.07)
$\hat{\delta}_2$	0.01	(0.22)	0.00	(0.09)	0.02	(0.07)
$\hat{\delta}_3$	-0.88	(0.28)	-0.92	(0.13)	-0.93	(0.08)
$\hat{\delta}_4$	-1.51	(0.25)	-1.57	(0.10)	-1.57	(0.08)
$\hat{\delta}_5$	-0.84	(0.21)	-0.84	(0.12)	-0.82	(0.07)
$\hat{\delta}_6$	-0.76	(0.21)	-0.75	(0.10)	-0.75	(0.08)
$\hat{\delta}_7$	-0.81	(0.24)	-0.82	(0.12)	-0.79	(0.08)
$\hat{\delta}_8$	-0.04	(0.21)	-0.04	(0.11)	0.00	(0.08)
$\hat{\delta}_9$	0.05	(0.22)	0.04	(0.11)	0.08	(0.07)
$\hat{\delta}_{10}$	-0.85	(0.21)	-0.84	(0.11)	-0.84	(0.07)
$\hat{\tau}_1$	-2.25	(0.10)	-2.25	(0.05)	-2.24	(0.03)
$\hat{\tau}_2$	2.25	(0.10)	2.25	(0.05)	2.24	(0.03)
$\mu_{\hat{\theta}}$	0.01	(0.00)	0.01	(0.00)	0.01	(0.00)
$\Sigma_{\hat{\theta}}$	0.99	(0.12)	1.00	(0.05)	1.00	(0.04)

Note. $\{100,500,1000\}$ = Number of simulated individuals. $\{\hat{\delta}_1,\hat{\delta}_2,\ldots,\hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1,\hat{\tau}_2\}$ = thresholds 1 and 2. $\mu_{\hat{\theta}}$ = Mean person location. $\Sigma_{\hat{\theta}}$= Standard deviation of the person locations. M = Mean. SE = Standard error.

Table 6. Mean and Standard Error of the Parameter Estimates for the RSM when J = 25

	100		500		1000	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.19	(0.27)	-0.12	(0.12)	-0.10	(0.08)
$\hat{\delta}_2$	0.02	(0.22)	0.00	(0.10)	0.02	(0.07)
$\hat{\delta}_3$	-0.92	(0.30)	-0.96	(0.13)	-0.96	(0.08)
$\hat{\delta}_4$	-1.57	(0.24)	-1.63	(0.10)	-1.64	(0.08)
$\hat{\delta}_5$	-0.87	(0.23)	-0.87	(0.12)	-0.85	(0.07)
$\hat{\delta}_6$	-0.79	(0.21)	-0.78	(0.10)	-0.78	(0.08)
$\hat{\delta}_7$	-0.84	(0.25)	-0.85	(0.12)	-0.82	(0.08)
$\hat{\delta}_8$	-0.04	(0.22)	-0.04	(0.11)	0.00	(0.08)
$\hat{\delta}_9$	0.05	(0.23)	0.05	(0.12)	0.08	(0.07)
$\hat{\delta}_{10}$	-0.88	(0.21)	-0.88	(0.11)	-0.87	(0.07)
$\hat{\delta}_{11}$	-1.29	(0.24)	-1.33	(0.12)	-1.33	(0.07)
$\hat{\delta}_{12}$	-1.10	(0.25)	-1.09	(0.12)	-1.05	(0.09)
$\hat{\delta}_{13}$	-1.14	(0.22)	-1.20	(0.14)	-1.19	(0.09)
$\hat{\delta}_{14}$	-1.41	(0.24)	-1.44	(0.08)	-1.45	(0.08)
$\hat{\delta}_{15}$	0.58	(0.26)	0.56	(0.10)	0.56	(0.07)
$\hat{\delta}_{16}$	-0.36	(0.26)	-0.33	(0.12)	-0.32	(0.10)
$\hat{\delta}_{17}$	-0.05	(0.20)	-0.09	(0.10)	-0.11	(0.07)
$\hat{\delta}_{18}$	0.10	(0.26)	0.10	(0.12)	0.11	(0.08)
$\hat{\delta}_{19}$	-0.13	(0.25)	-0.17	(0.13)	-0.16	(0.07)
$\hat{\delta}_{20}$	-0.44	(0.21)	-0.44	(0.11)	-0.45	(0.09)
$\hat{\delta}_{21}$	0.03	(0.25)	0.00	(0.11)	0.00	(0.08)
$\hat{\delta}_{22}$	0.50	(0.25)	0.54	(0.10)	0.50	(0.09)
$\hat{\delta}_{23}$	0.54	(0.24)	0.53	(0.11)	0.54	(0.09)
$\hat{\delta}_{24}$	0.80	(0.25)	0.76	(0.10)	0.77	(0.07)
$\hat{\delta}_{25}$	0.85	(0.26)	0.79	(0.10)	0.82	(0.08)
$\hat{\tau}_1$	-2.36	(0.06)	-2.35	(0.03)	-2.34	(0.02)
$\hat{\tau}_2$	2.36	(0.06)	2.35	(0.03)	2.34	(0.02)
$\mu_{\hat{\theta}}$	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
$\Sigma_{\hat{\theta}}$	1.13	(0.08)	1.14	(0.04)	1.13	(0.02)

Table 6 (cont'd)

Note. $\{100,500,1000\}$ = Number of simulated individuals. $\{\delta_1,\delta_2,\ldots,\delta_{25}\}$ = location for items 1 – 25. $\{\hat{\tau}_1,\hat{\tau}_2\}$ = thresholds 1 and 2. $\mu_{\hat{\theta}}$ = Mean person location. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations. SE = Standard error.

Table 7. Mean and Standard Error of the Parameter Estimates for the HMGL-RSM when J = 25

	100		500		1000	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.19	(0.26)	-0.09	(0.11)	-0.09	(0.07)
$\hat{\delta}_2$	0.02	(0.21)	0.02	(0.09)	0.02	(0.07)
$\hat{\delta}_3$	-0.89	(0.28)	-0.91	(0.13)	-0.93	(0.08)
$\hat{\delta}_4$	-1.52	(0.23)	-1.56	(0.10)	-1.58	(0.07)
$\hat{\delta}_5$	-0.84	(0.22)	-0.82	(0.12)	-0.82	(0.07)
$\hat{\delta}_6$	-0.76	(0.20)	-0.73	(0.10)	-0.75	(0.08)
$\hat{\delta}_7$	-0.81	(0.24)	-0.80	(0.12)	-0.79	(0.08)
$\hat{\delta}_8$	-0.04	(0.21)	-0.02	(0.11)	0.00	(0.08)
$\hat{\delta}_9$	0.05	(0.22)	0.06	(0.11)	0.08	(0.07)
$\hat{\delta}_{10}$	-0.85	(0.20)	-0.82	(0.11)	-0.84	(0.07)
$\hat{\delta}_{11}$	-1.24	(0.23)	-1.27	(0.11)	-1.28	(0.07)
$\hat{\delta}_{12}$	-1.06	(0.24)	-1.03	(0.11)	-1.01	(0.09)
$\hat{\delta}_{13}$	-1.10	(0.21)	-1.14	(0.13)	-1.15	(0.09)
$\hat{\delta}_{14}$	-1.36	(0.23)	-1.37	(0.07)	-1.39	(0.08)
$\hat{\delta}_{15}$	0.56	(0.25)	0.56	(0.09)	0.54	(0.07)
$\hat{\delta}_{16}$	-0.35	(0.25)	-0.30	(0.11)	-0.31	(0.09)
$\hat{\delta}_{17}$	-0.04	(0.19)	-0.07	(0.09)	-0.11	(0.07)
$\hat{\delta}_{18}$	0.10	(0.25)	0.12	(0.12)	0.11	(0.07)
$\hat{\delta}_{19}$	-0.12	(0.24)	-0.15	(0.12)	-0.16	(0.07)
$\hat{\delta}_{20}$	-0.42	(0.20)	-0.40	(0.11)	-0.43	(0.08)
$\hat{\delta}_{21}$	0.03	(0.24)	0.02	(0.11)	0.00	(0.08)

Table 7 (cont'd)

$\hat{\delta}_{22}$	0.48	(0.24)	0.55	(0.10)	0.49	(0.08)
$\hat{\delta}_{23}$	0.52	(0.23)	0.53	(0.10)	0.52	(0.09)
$\hat{\delta}_{24}$	0.77	(0.24)	0.76	(0.10)	0.75	(0.07)
$\hat{\delta}_{25}$	0.82	(0.25)	0.78	(0.10)	0.79	(0.08)
$\hat{\tau}_1$	-2.26	(0.06)	-2.25	(0.03)	-2.24	(0.02)
$\hat{\tau}_2$	2.26	(0.06)	2.25	(0.03)	2.24	(0.02)
$\mu_{\hat{\theta}}$	0.00	(0.00)	0.02	(0.02)	0.00	(0.00)
$\Sigma_{\hat{\theta}}$	0.99	(0.08)	1.00	(0.04)	1.00	(0.02)

Note. $\{100,500,1000\}$ = Number of simulated individuals. $\{\delta_1,\delta_2,\ldots,\delta_{25}\}$ = location for items 1 – 25. $\{\hat{\tau}_1,\hat{\tau}_2\}$ = thresholds 1 and 2. $\mu_{\hat{\theta}}$ = Mean person location. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations. SE = Standard error.

3-2-2. RMSE

The results of the RMSE for μ_θ, Σ, δ_j, and τ_i of the RSM and HMGL-RSM when persons respond to 10 and 25 items are provided in Tables 8 and 9. For both the RSM and HMGL-RSM, trends indicated that as persons increased from 100 to 1000, the RMSE generally decreased for μ_θ, Σ, δ_j, and τ_i. This is expected because as the persons increase there were more observations from which to estimate the person and item parameters.

Additionally as one case see, although the RMSE decreases for both the RSM and HMLG-RSM, the RSME is somewhat higher for the RSM estimates. This is particularly the case for τ_1, τ_2, and Σ_θ, when persons responded to 10 items. This probably occurs because, when using WINSTEPS to estimate these parameters, more items are needed to obtain more precise estimates. In contrast, notice that as more items are estimated the

RMSE does not decrease for Σ_θ for the HMGL-RSM; rather the RMSE remains fairly stable. This occurs because Σ_θ of the HMGL-RSM is the variation between the empirical Bayes estimates of the random effect of persons. As discussed by Raudenbush and Bryk (2002), and shown here, this estimate depends on the number of units of the random effects, not the number of fixed effects, in this case, the number of items.

Table 8. RMSE for the RSM and HMGL-RSM across 10 Items

	RSM			HMGL-RSM		
	K			K		
	100	500	1000	100	500	1000
δ_1	0.30	0.13	0.08	0.27	0.12	0.07
δ_2	0.24	0.10	0.08	0.21	0.10	0.07
δ_3	0.32	0.17	0.14	0.28	0.12	0.08
δ_4	0.29	0.21	0.20	0.25	0.10	0.08
δ_5	0.26	0.18	0.12	0.21	0.12	0.07
δ_6	0.25	0.14	0.13	0.20	0.10	0.08
δ_7	0.28	0.17	0.11	0.24	0.12	0.08
δ_8	0.24	0.12	0.08	0.21	0.11	0.08
δ_9	0.25	0.13	0.08	0.22	0.11	0.07
δ_{10}	0.25	0.15	0.11	0.21	0.11	0.07
τ_1	0.31	0.30	0.28	0.10	0.05	0.03
τ_2	0.31	0.30	0.28	0.10	0.05	0.03
μ_θ	0.01	0.01	0.00	0.01	0.01	0.01
Σ_θ	0.37	0.37	0.36	0.12	0.05	0.04

Note. K = Number of simulated persons. $\{\delta_1, \delta_2, \ldots, \delta_{10}\}$ = location for items 1 – 10. $\{\tau_1, \tau_2\}$ = thresholds 1 and 2. $\mu_{\hat{\theta}}$ = Mean person location. Σ_θ = Standard deviation of the person locations.

Table 9. RMSE for the RSM and HMGL-RSM across 25 Items

	RSM			HMGL-RSM		
	K			K		
	100	500	1000	100	500	1000
δ_1	0.28	0.12	0.08	0.27	0.11	0.07
δ_2	0.22	0.10	0.07	0.21	0.09	0.07
δ_3	0.29	0.13	0.09	0.28	0.13	0.08
δ_4	0.24	0.12	0.10	0.24	0.10	0.07
δ_5	0.24	0.14	0.08	0.22	0.12	0.07
δ_6	0.21	0.11	0.09	0.20	0.10	0.08
δ_7	0.25	0.12	0.08	0.23	0.12	0.08
δ_8	0.22	0.11	0.08	0.21	0.10	0.08
δ_9	0.23	0.12	0.07	0.22	0.11	0.07
δ_{10}	0.21	0.11	0.07	0.20	0.11	0.07
δ_{11}	0.24	0.13	0.09	0.23	0.11	0.07
δ_{12}	0.26	0.14	0.09	0.24	0.11	0.09
δ_{13}	0.22	0.15	0.10	0.22	0.13	0.09
δ_{14}	0.24	0.10	0.10	0.23	0.07	0.08
δ_{15}	0.26	0.10	0.07	0.25	0.09	0.07
δ_{16}	0.26	0.12	0.09	0.25	0.12	0.09
δ_{17}	0.20	0.10	0.08	0.20	0.09	0.07
δ_{18}	0.25	0.12	0.08	0.24	0.12	0.07
δ_{19}	0.25	0.13	0.07	0.24	0.12	0.07
δ_{20}	0.21	0.11	0.09	0.20	0.11	0.08
δ_{21}	0.25	0.11	0.08	0.24	0.11	0.08
δ_{22}	0.25	0.11	0.09	0.24	0.11	0.08
δ_{23}	0.24	0.11	0.09	0.23	0.10	0.08
δ_{24}	0.26	0.10	0.08	0.24	0.10	0.07
δ_{25}	0.26	0.10	0.08	0.25	0.10	0.07
τ_1	0.13	0.11	0.10	0.06	0.03	0.02
τ_2	0.13	0.11	0.10	0.06	0.03	0.02
μ_θ	0.00	0.00	0.00	0.01	0.01	0.01
Σ_θ	0.15	0.14	0.13	0.12	0.05	0.04

Table 9 (cont'd)

Note. K = Number of simulated persons. $\{\delta_1, \delta_2, \ldots, \delta_{25}\}$ = location for items 1 – 25. $\{\tau_1, \tau_2\}$ = thresholds 1 and 2. $\mu_{\hat{\theta}}$ = Mean person location. Σ_{θ} = Standard deviation of the person locations.

3-3. Example

Below, an example analysis is presented using both the HMGL-RSM and -PCM. The purpose is to illustrate the basic concepts underlying these two models, as well as to illustrate the differences between the two models.

3-3-1. Design

The design of the analysis is as follows. Five hundred respondents were randomly selected from a larger sample of students that responded to a confidential readiness assessment. (Note this was the same assessment that was simulated in Section 3-1.) In this sample, 46% had parents with high SES (SES = 1); 44% had parents with middle SES (SES = 2); and 10% had parents with low SES (SES = 3). 56% were male, and 44% were female. Additionally, approximately less than 1% were age 5; 23% were age 6; 65% were age 7; 12% were age 8; and less than 1% were age 9. Lastly, less than 1% were Asian; 42% were African-American; 2% were Hispanic; and 56% were Caucasian.

For the purposes of this illustration, only the first 10 items of the assessment were used. (Note each item measured the person's personal and social development.) Additionally, only those respondents who answered each item and whose parents

provided their SES were used. As illustrated above, the sample and item sizes were adequate to obtain relatively precise parameter estimates.

3-3-2. Analysis

To analyze the responses of the students, PROC NLMIXED of SAS (2001) was used to estimate the person and item parameters for the HMGL-RSM and -PCM. Comparison between model fit is achieved using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Manalo (2004) and Singer (1998) shows these measures to be adequate for judging model fit in HLM analyses.

3-3-3. Results

The results of the analysis for the HMGL-RSM and -PCM are presented in Table 10. As can be seen, $\hat{\delta}_1$ - $\hat{\delta}_{10}$, $\mu_{\hat{\theta}}$, and $\hat{\Sigma}_{\hat{\theta}}$ are similar between the two models. Additionally, $\hat{\tau}_{i1}$ - $\hat{\tau}_{i10}$ are similar across the ten items for the -PCM. Lastly, notice $\hat{\tau}_{i1}$ - $\hat{\tau}_{i10}$ are also generally similar to $\hat{\tau}_1$ and $\hat{\tau}_2$ for the -RSM.

To determine which model better fits the data, the AIC and BIC are examined. As shown, the AIC is lower for the HMGL-PCM than the -RSM, but the BIC is lower for the HMGL-RSM than the -PCM. This suggests that the AIC indicates the HMGL-PCM as being a better fit for the data, while the BIC indicates the HMGL-RSM as being the better fit. However, focusing on the information weights, which act similar to an effect size in that measures are normalized and models can be compared on a common (probabilistic) scale (formulas can be found in Burnham and Anderson (2002)), we see that the information weights for the HMGL-RSM and -PCM are .11 and .89 for the AIC, and 1.00

and 0.00 for the BIC. Since higher values indicate better fit, and given the larger disparity in the weights between the BIC than the AIC, and because the BIC compensates for the large sample size and the AIC does not, the BIC might give a better representation of the model fit for the two models. Hence, using the BIC, it appears that the HMGL-RSM fits the data better. This suggests that the thresholds $\left(\tau_{ij}\right)$ are common across items (i.e., $\tau_{ij} = \tau_i$), and items share common thresholds.

Table 10. Parameter Estimates for the HMGL-RSM and -PCM

	RSM		PCM				
	Est.	SE	Est.	SE	τ_{1j}	τ_{2j}	SE(τ_{2j})
δ_1	0.49	(0.16)	0.48	(0.14)	-2.10	2.10	(0.09)
δ_2	0.75	(0.16)	0.72	(0.15)	-2.03	2.03	(0.09)
δ_3	-0.28	(0.16)	-0.28	(0.14)	-2.00	2.00	(0.08)
δ_4	-0.92	(0.16)	-0.93	(0.14)	-2.22	2.22	(0.07)
δ_5	-0.12	(0.16)	-0.12	(0.14)	-2.05	2.05	(0.08)
δ_6	0.03	(0.16)	0.04	(0.14)	-2.22	2.22	(0.08)
δ_7	-0.22	(0.16)	-0.22	(0.14)	-2.69	2.69	(0.09)
δ_8	0.79	(0.16)	0.85	(0.15)	-2.39	2.39	(0.10)
δ_9	0.87	(0.16)	0.81	(0.15)	-1.87	1.87	(0.09)
δ_{10}	-0.04	(0.16)	-0.05	(0.14)	-2.09	2.09	(0.08)
τ_1	-2.15	.	-	-	-	-	-
τ_2	2.15	(0.03)	-	-	-	-	-
μ_θ	-0.01	.	-0.01	.			
Σ_θ	2.80	(0.12)	2.82	(0.12)			
AIC	7146.7		7142.6				
BIC	7201.5		7273.2				

Table 10 (cont'd)

Note. $\{\delta_1, \delta_2, \ldots, \delta_{10}\}$ = location for items 1 – 10. $\{\tau_1, \tau_2\}$ = thresholds 1 and 2 for the -RSM. $\{\tau_{1j}, \tau_{2j}\}$ = thresholds 1 and 2 for item j of the -PCM. μ_θ = Mean person location. $\sum_\theta$ = Standard deviation of the person locations. AIC = Akaike Information Criterion. BIC = Bayesian Information Criterion. Est. = Estimate. SE = Standard error.

To illustrate the interpretation of $\hat{\theta}$ for the HMGL-RSM (which is similar for the -PCM), one focuses on an arbitrarily chosen respondent. For this respondent, $\hat{\theta} = -2.36$ logits. Note although a rater selected the categories for the respondent, assume (for this example and the following examples) that the respondent made the selection for himself or herself. Thus, on the underlying continuum, notice this person's location is much lower on the scale than the overall attractiveness of, say, item 1 $\left(\hat{\delta}_1 = .49\right)$. As shown below, for this item this suggests that the respondent is more likely to be attracted to the lower categorical responses than the higher categorical responses.

To determine the probability that this respondent will select category 0, 1, or 2, one refers back to Equations (2.6) and (2.32)-(2.35). For item 1,

$$\pi_{01} = \frac{\exp(0)}{\psi} = .67$$

$$\pi_{11} = \frac{\exp(-2.36 - .49 - (-2.15))}{\psi} = .33$$

$$\pi_{21} = \frac{\exp([-2.36 - .49 - 2.15] + [-2.36 - .49 - (-2.15)])}{\psi} = .00,$$

where

$$
\begin{aligned}
\psi &= \exp(0) + \\
&\quad \exp(-2.36 - .49 - (-2.15)) + \\
&\quad \exp([-2.36 - .49 - 2.15] + [-2.36 - .49 - (-2.15)]) \\
&\quad = 1.50.
\end{aligned}
$$

This suggests that, for item 1, the probability that this respondent will select category 0 is .67, which is approximately double the probability of selecting category 1. As for category 2, the respondent has a probability of 0 of selecting this category.

Chapter 4. Extending the HMGL-RSM To Include Person Covariates

4-1. The HMGL-RSM with Person Covariates

As seen in Chapter 3, one advantage of applying the HMGLM to model the RSM is that the it affords the opportunity to obtain better precision for the estimates of the person and item parameters. However, this is not the only advantage. As mentioned previously, another advantage—the primary focus of this paper—is that by modeling the RSM in the HMGLM, the user may posit a model that includes covariates. In this chapter, the inclusion of covariates at the person level is discussed. This form of the HMGL-RSM may be especially important in accountability investigations in which the user is interested in the location of student, after controlling for the effects of a covariate (e.g., Stone and Lane (2003)).

To model the HMGL-RSM with person covariates, one follows the previous definitions of the HMGL-RSM (Section 2-2), in which the category is nested within the item, which in turn is nested within the person. However, now covariates at the person level are included.

4-1-1. The Level-1 Model with Person Covariates

The Level-1 model (the category level) is defined as

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \sum_{j=1}^{J} \beta_{jk}^{(i)} x_{jk}\,, \tag{4.1}$$

where $\beta_{jk}^{(i)}$ is the mean category effect if person k selects category i of item j; and x_{jk} is a dummy variable with values 1 if person k answers item j, and 0 otherwise.

4-1-2. The Level-2 Model with Person Covariates

The Level-2 model (the item level) is defined as

$$\beta_{jk}^{(i)} = \gamma_{0jk} + \sum_{i=1}^{I} \gamma_{1\cdot k}^{(i)} w_{1\cdot k}^{(i)}, \tag{4.2}$$

where, for person k, γ_{0jk} is the mean effect of item j across categories i; $\gamma_{1\cdot k}^{(i)}$ is the effect of an item on a particular category i; and $w_{1\cdot k}^{(i)}$ is a dummy variable with values 1 if $i' = i$, and 0 otherwise. For identifiability, $\gamma_{1\cdot k}^{(0)} \equiv 0$.

4-1-3. The Level-3 Model with Person Covariates

The Level-3 model (the person level model) is defined as

$$\gamma_{0jk} = \lambda_{0j0} + \sum_{t=1}^{T} \lambda_{0j,t} w_{0jk,t} + u_k, \tag{4.3}$$

$$\gamma_{1\cdot k}^{(i)} = \lambda_{1\cdot 0}^{(i)}, \tag{4.4}$$

where, for the j^{th} item that is answered by person k, λ_{0j0} is the mean effect of persons on item j; $\lambda_{0j,t}$ is the effect of person covariate t; $w_{0jk,t}$ is a dummy variable with values 1 if covariate t effects person k, and 0 otherwise; u_k is the random effect of person k on the mean effect of item j, after accounting for covariate t; and $\lambda_{1\cdot 0}^{(i)}$ is the mean change in λ_{0j0} for a particular category of the items, for all persons.

However, the effect of covariate t is assumed to effect person k equally for each item j; hence $\lambda_{01,t} = \lambda_{02,t} = \ldots = \lambda_{0J,t} = \lambda_{0\cdot,t}$. Thus, the Level-3 model for γ_{0jk} becomes

$$\gamma_{0jk} = \lambda_{0j0} + \sum_{t=1}^{T} \lambda_{0\cdot,t} w_{0\cdot k,t} + u_k, \tag{4.5}$$

where λ_{0j0} and u_k are defined above; $\lambda_{0\cdot,t}$ is the effect of person covariate t, which is now constant across items; and $w_{0\cdot k,t}$ is a dummy variable with values 1 if covariate t effects person k, and 0 otherwise.

Here, it is helpful to refer back to the honesty example, in which a particular feeling of an applicant in nested within an item, which in turn is nested within the person. As before, a particular answered item not only depends upon the overall attractiveness of the item $\left(\lambda_{0j0}\right)$, but it also depends on the attractiveness of the item influencing a particular feeling $\left(\lambda_{1\cdot 0}^{(i)}\right)$. In addition to the honesty of the person, the response to the item also depends upon the person covariate $\left(\lambda_{0\cdot,t}\right)$, such as SES. In other words, for example, the respondent may become more honest as SES increases.

4-1-4. The Combined Model with Person Covariates

The combined model of the HMGL-RSM with person covariates reduces to the following for a particular category i of the item j

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + \sum_{t=1}^{T} \lambda_{0\cdot,t} w_{0\cdot k,t} + u_k, \tag{4.6}$$

where all terms are defined above.

Therefore, the parameters of the HMGL-RSM with person covariates are related to and extend the parameters of the traditional RSM in the following manner:

$$\delta_j = -\lambda_{0j0}, \tag{4.7}$$

$$\tau_i = -\lambda_{1\cdot 0}^{(i)}. \tag{4.8}$$

and

$$\begin{aligned} \theta_{k,1} &= \lambda_{0\cdot,1} w_{0\cdot k,1} + u_k \\ \theta_{k,2} &= \lambda_{0\cdot,2} w_{0\cdot k,2} + u_k \\ &\cdots \\ \theta_{k,T} &= \lambda_{0\cdot,T} w_{0\cdot k,T} + u_k \end{aligned} . \tag{4.9}$$

where δ_j and τ_i are defined above; and $\theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,T}$ is the location of person k, when accounting for covariate t (t = 1, ..., T).

4-2. Simulation Study for the HMGL-RSM with Person Covariates

The following section describes a simulation study for the HMGL-RSM with person covariates. Since Section 3-2 already described a simulation study that examined the parameter recovery of the person and item parameters when person covariates were not added to the HMGL-RSM, the focus of this section is to examine the behaviors of the person parameters when being influenced by covariates.

4-2-1. Design

The design of the simulation is as follows. Observations were simulated using the HMGL-RSM. For the study, 100, 500, or 1000 simulees responded to 10 polytomous items, where each item consisted of 3 categories i (i = 0, 1, 2). The number of simulees, items, and categories were chosen to follow typical data from a questionnaire (e.g., Dodd, 1990; Smith & Johnson, 2000; Zhu, Updyke, & Lewandowski, 1997) or a large-scale assessment (e.g., Michigan Education Assessment Program, 2003; U.S. Department of Education, 1999). In addition, the number of simulees and items were chosen because, as

shown in Section 3-2, these sample sizes allow for reasonable precision (at least when covariates were not modeled).

To produce the simulated responses, each simulee k was randomly assigned to be in one of three levels of a person covariate $\left(\lambda_{0\cdot,1}\right)$. The probability of being selected to a given level was chosen to be .46, .46, and .08, respectively. Probabilities followed the actual frequencies of the levels of a covariate used in an actual administration of a confidential readiness assessment. Here, the covariate was SES.

Additionally, each simulee k was randomly assigned a u_k, $u \sim N(0,1)$. Thus, θ_k was obtained by using Equation (4.9). For the simulation, to examine the effect of the person covariate, $\lambda_{0\cdot,1}$ was selected to be .2, .5, and 1. These values were chosen to follow previous simulation designs of hierarchical IRT models using person covariates (Kamata, 1998). δ_j and τ_i were randomly selected to represent parameter estimates obtained from typical polytomous data (i.e., items 1-10 in Table 2).

Using θ_k, δ_j, and τ_i, three response probabilities for each simulee by item combination were produced, $P_{0jk}(\theta)$, $P_{1jk}(\theta)$, and $P_{2jk}(\theta)$. If $\sum_{0}^{i'} P_{i'jk}(\theta) < Y_{jk} \leq \sum_{0}^{i'+1} P_{i'jk}(\theta)$, then simulee k was assigned a response of $i'+1$ for item j; otherwise a response of 0 was assigned. Note that $i' = 0, 1$; and Y_{jk} was a single, random number for each $j \times k$ combination, $Y \sim U(0,1)$.

The simulation procedure utilized a fully crossed 3 × 3 factorial design that simulated 9 conditions. Each administration was iterated 50 times producing 450 unique response data matrices. The number of iterations was chosen because Kamata (1998)

showed this to be a reasonable number for obtaining stable estimates. S-Plus (2000) was used to generate all data. SAS (2001) was used to obtain parameter estimates and conduct significance tests.

4-2-2. Analysis

For the analysis regarding the parameter recovery of the HMGL-RSM with person covariates, the RMSE for u_k, $\lambda_{0\cdot,1}$, δ_j and τ_i was obtained over the iterations for each condition. Specifically, the RMSE was obtained by

$$RMSE(\hat{\omega}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\omega}_n - \omega_n)^2}, \tag{4.10}$$

where the maximum number of n iterations was $N = 50$; and ω is an arbitrary parameter representing either u_k, $\lambda_{0\cdot,1}$, δ_j and τ_i. A descriptive analysis of the RMSE was conducted for each condition.

4-2-3. Results: Descriptive Statistics

Displayed in Tables 11, 12, and 13 are the mean and standard deviations of the parameter estimates for the HMGL-RSM when $\lambda_{0\cdot,1}$ equaled .2, .5, and 1, respectively. As can be seen, the standard deviations of the estimates are similar across conditions. Additionally, the standard deviations are fairly low and decrease as the number of persons increase. This suggests that PROC NLMIXED obtains relatively consistent estimates of the HMGL-RSM parameters.

As for the mean of the estimates, in general, the estimates obtained by PROC NLMIXED for the HMGL-RSM appear to differ only slightly from their parameter values. Below, in Section 4-2-4, the RMSE is examined.

Table 11. Mean and Standard Error of the Parameter Estimates for the HMGL-RSM when $\lambda_{0\cdot,1} = .2$

	100		500		1000	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.10	(0.35)	-0.13	(0.14)	-0.09	(0.11)
$\hat{\delta}_2$	0.04	(0.37)	0.00	(0.16)	0.00	(0.14)
$\hat{\delta}_3$	-0.89	(0.39)	-0.93	(0.16)	-0.92	(0.12)
$\hat{\delta}_4$	-1.59	(0.29)	-1.61	(0.16)	-1.58	(0.14)
$\hat{\delta}_5$	-0.81	(0.38)	-0.83	(0.18)	-0.82	(0.14)
$\hat{\delta}_6$	-0.76	(0.34)	-0.76	(0.15)	-0.73	(0.13)
$\hat{\delta}_7$	-0.88	(0.35)	-0.85	(0.15)	-0.81	(0.12)
$\hat{\delta}_8$	-0.05	(0.31)	-0.05	(0.14)	0.00	(0.11)
$\hat{\delta}_9$	0.05	(0.37)	0.05	(0.16)	0.07	(0.12)
$\hat{\delta}_{10}$	-0.84	(0.38)	-0.87	(0.17)	-0.85	(0.12)
$\hat{\tau}_1$	-2.27	(0.12)	-2.25	(0.05)	-2.25	(0.03)
$\hat{\tau}_2$	2.27	(0.12)	2.25	(0.05)	2.25	(0.03)
$\lambda_{0\cdot,1}$	0.19	(0.18)	0.19	(0.07)	0.20	(0.06)
$\mu_{\hat{u}}$	0.01	(0.00)	0.01	(0.00)	0.01	(0.00)
$\Sigma_{\hat{u}}$	0.99	(0.13)	1.00	(0.06)	1.00	(0.03)

Note. $\{100, 500, 1000\}$ = Number of simulated individuals. $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. $\lambda_{0\cdot,1}$ = person covariate. $\mu_{\hat{u}}$ = Mean person location, after controlling for $\lambda_{0\cdot,1}$. $\Sigma_{\hat{\theta}}$= Standard deviation of the person locations, after controlling for $\lambda_{0\cdot,1}$. M = Mean. SE = Standard error.

Table 12. Mean and Standard Error of the Parameter Estimates for the HMGL-RSM when $\lambda_{0\cdot,1} = .5$

	100		500		1000	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.14	(0.39)	-0.12	(0.14)	-0.09	(0.11)
$\hat{\delta}_2$	0.06	(0.36)	0.01	(0.16)	0.00	(0.16)
$\hat{\delta}_3$	-0.87	(0.40)	-0.95	(0.14)	-0.93	(0.12)
$\hat{\delta}_4$	-1.57	(0.31)	-1.62	(0.16)	-1.58	(0.15)
$\hat{\delta}_5$	-0.80	(0.34)	-0.83	(0.16)	-0.81	(0.14)
$\hat{\delta}_6$	-0.75	(0.33)	-0.77	(0.14)	-0.73	(0.13)
$\hat{\delta}_7$	-0.83	(0.33)	-0.86	(0.16)	-0.80	(0.13)
$\hat{\delta}_8$	0.00	(0.32)	-0.06	(0.14)	-0.01	(0.12)
$\hat{\delta}_9$	0.07	(0.31)	0.06	(0.15)	0.07	(0.12)
$\hat{\delta}_{10}$	-0.82	(0.35)	-0.87	(0.16)	-0.86	(0.14)
$\hat{\tau}_1$	-2.24	(0.13)	-2.25	(0.05)	-2.24	(0.04)
$\hat{\tau}_2$	2.24	(0.13)	2.25	(0.05)	2.24	(0.04)
$\lambda_{0\cdot,1}$	0.50	(0.17)	0.49	(0.07)	0.50	(0.06)
$\mu_{\hat{u}}$	0.00	(0.00)	0.01	(0.00)	0.01	(0.00)
$\Sigma_{\hat{u}}$	0.97	(0.13)	1.00	(0.06)	1.00	(0.03)

Note. $\{100, 500, 1000\}$ = Number of simulated individuals. $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. $\lambda_{0\cdot,1}$ = person covariate. $\mu_{\hat{u}}$ = Mean person location, after controlling for $\lambda_{0\cdot,1}$. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations, after controlling for $\lambda_{0\cdot,1}$. M = Mean. SE = Standard error.

Table 13. Mean and Standard Error of the Parameter Estimates for the HMGL-RSM when $\lambda_{0\cdot,1} = 1$

	100		500		1000	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.10	(0.35)	-0.12	(0.14)	-0.10	(0.12)
$\hat{\delta}_2$	0.10	(0.38)	0.00	(0.17)	0.00	(0.15)
$\hat{\delta}_3$	-0.87	(0.39)	-0.94	(0.17)	-0.93	(0.14)
$\hat{\delta}_4$	-1.53	(0.34)	-1.60	(0.17)	-1.59	(0.16)
$\hat{\delta}_5$	-0.77	(0.37)	-0.83	(0.17)	-0.82	(0.15)
$\hat{\delta}_6$	-0.71	(0.38)	-0.77	(0.14)	-0.75	(0.16)
$\hat{\delta}_7$	-0.78	(0.36)	-0.86	(0.15)	-0.82	(0.13)
$\hat{\delta}_8$	0.04	(0.34)	-0.04	(0.15)	-0.01	(0.13)
$\hat{\delta}_9$	0.11	(0.34)	0.07	(0.16)	0.07	(0.14)
$\hat{\delta}_{10}$	-0.79	(0.38)	-0.87	(0.17)	-0.86	(0.14)
$\hat{\tau}_1$	-2.26	(0.16)	-2.25	(0.07)	-2.25	(0.05)
$\hat{\tau}_2$	2.26	(0.16)	2.25	(0.07)	2.25	(0.05)
$\lambda_{0\cdot,1}$	1.03	(0.18)	0.99	(0.07)	1.00	(0.07)
$\mu_{\hat{u}}$	-0.01	(0.01)	-0.01	(0.00)	-0.01	(0.00)
$\Sigma_{\hat{u}}$	0.98	(0.11)	1.01	(0.06)	1.00	(0.04)

Note. $\{100,500,1000\}$ = Number of simulated individuals. $\{\hat{\delta}_1,\hat{\delta}_2,\ldots,\hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1,\hat{\tau}_2\}$ = thresholds 1 and 2. $\lambda_{0\cdot,1}$ = person covariate. $\mu_{\hat{u}}$ = Mean person location, after controlling for $\lambda_{0\cdot,1}$. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations, after controlling for $\lambda_{0\cdot,1}$. M = Mean. SE = Standard error.

4-2-4. Results: RMSE

The results of the RMSE for $\mu_{\hat{u}}$, $\Sigma_{\hat{u}}$, $\lambda_{0\cdot,1}$, δ_j, and τ_i of the HMGL-RSM with a person covariate are provided in Table 14. Trends indicated that as persons increased

from 100 to 1000, the RMSE generally decreased for $\mu_{\hat{u}}$, $\Sigma_{\hat{u}}$, $\lambda_{0\cdot,1}$, δ_j, and τ_i. This is expected because as the persons increase there were more observations from which to estimate the person and item parameters.

Additionally as one case see, the magnitude of the covariate $\left(\lambda_{0\cdot,1}\right)$ does not influence the RMSE. This illustrates that regardless of the size of the covariate, the coefficient for the covariate is recovered fairly well, with increasing precision as the number of persons increase.

Table 14. RMSE for the HMGL-RSM with Person Covariates

	.2			.5			1		
	100	500	1000	100	500	1000	100	500	1000
δ_1	0.35	0.15	0.11	0.38	0.14	0.11	0.35	0.15	0.12
δ_2	0.37	0.15	0.14	0.36	0.15	0.16	0.39	0.17	0.15
δ_3	0.39	0.16	0.12	0.40	0.15	0.12	0.39	0.17	0.14
δ_4	0.29	0.16	0.13	0.31	0.17	0.15	0.34	0.17	0.16
δ_5	0.38	0.18	0.14	0.34	0.16	0.14	0.37	0.17	0.15
δ_6	0.34	0.15	0.12	0.33	0.15	0.13	0.37	0.15	0.16
δ_7	0.36	0.16	0.12	0.33	0.17	0.13	0.36	0.16	0.13
δ_8	0.31	0.14	0.11	0.32	0.15	0.12	0.34	0.16	0.13
δ_9	0.37	0.16	0.12	0.31	0.15	0.12	0.34	0.15	0.14
δ_{10}	0.38	0.17	0.12	0.35	0.16	0.14	0.39	0.17	0.14
τ_1	0.12	0.05	0.03	0.13	0.05	0.04	0.16	0.07	0.05
τ_2	0.12	0.05	0.03	0.13	0.05	0.04	0.16	0.07	0.05
$\lambda_{0\cdot,1}$	0.18	0.07	0.05	0.16	0.07	0.06	0.18	0.07	0.07
$\mu_{\hat{u}}$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
$\Sigma_{\hat{u}}$	0.12	0.06	0.03	0.13	0.06	0.03	0.11	0.06	0.03

Table 14 (cont'd)

Note. $\{100,500,1000\}$ = Number of simulated individuals. $\{.2,.5,1\}$ = Values of $\lambda_{0\cdot,1}$. $\{\delta_1,\delta_2,\ldots,\delta_{10}\}$ = location for items 1 – 10. $\{\tau_1,\tau_2\}$ = thresholds 1 and 2. $\lambda_{0\cdot,1}$ = person covariate. $\mu_{\hat{\theta}}$ = Mean person location, after controlling for $\lambda_{0\cdot,1}$. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations, after controlling for $\lambda_{0\cdot,1}$.

4-3. Example Analysis of the HMGL-RSM with Person Covariates

The purpose of this section is to provide an example analysis that illustrates the basic concepts of the HMGL-RSM. In particular, how to use the HMGL-RSM to model person covariates is illustrated.

4-3-1. Design

The design of the analysis is as follows. Five hundred respondents were randomly selected from a larger sample of students that responded to a confidential readiness assessment. Note this was the same assessment simulated in Sections 3-1 and 4-2, and notice this was the same sample and set of items illustrated in Section 3-3. Specifically, in this sample, 46% had parents with high SES (SES = 1); 44% had parents with middle SES (SES = 2); and 10% had parents with low SES (SES = 3). 56% were male, and 44% were female. Additionally, approximately less than 1% were age 5; 23% were age 6; 65% were age 7; 12% were age 8; and less than 1% were age 9. Lastly, less than 1% were Asian; 42% were African-American; 2% were Hispanic; and 56% were Caucasian.

For the purposes of this illustration, only the first 10 items of the assessment were used. (Note each item measured the person's personal and social development.)

Additionally, only those respondents who answered each item and whose parents provided their SES were used. As illustrated in Section 4-2, the sample and item sizes were adequate to obtain relatively precise parameter estimates.

4-3-2. Analysis

To analyze the responses of the students, PROC NLMIXED of SAS (2001) was used to estimate the person and item parameters for the HMGL-RSM with SES as the person covariate. For comparison, the MRCMM (Equation (1.5) and (1.6)) with SES as a covariate for the random person location (θ_k) was also estimated using PROC NLMIXED. Note that SAS was used and not Conquest because it was of interest to compare the models, not the estimation algorithms of the software.

Also note that for the MRCMM, typically the item response is a column vector, where the number of rows is equal to the number of categories, and where a row equals 1 if the person selected a particular category, and 0 otherwise. This creates a dummy, column vector with observations equal to $I \times J \times K$ rows. For the data here, when the observation vector was created this way, the adaptive Gaussian quadrature integral approximations did not converge. To reduce the number of observations, rather than using a column vector of 0s and 1s, the categorical response itself (e.g., if the person selected category 2, the response was 2) was used. This created a response column vector with observations equal to $J \times K$ rows. By doing this column vector, convergence was achieved.

4-3-3. Results

The results of the analysis for the HMGL-RSM and MRCMM with SES as a person covariate are presented in Table 15. As can be seen, the HMGL-RSM and MRCMM yield identical estimates for all parameters. This result is not surprising given that in order to comply with the assumptions of IRT, the HMGL-RSM is defined by constraining the person location to be equal across items and categories (see Section 2-1-3). Consequently, there is no variation in the person location across items and categories, as is the case with the MRCMM. Additionally, recall that in order to get the estimation algorithm to converge for the MRCMM, the number of observations was reduced. Therefore, because the general form of the HMGL-RSM and MRCMM are similar, and because the number of observations is equivalent, similar estimates are obtained.

Table 15. Parameter Estimates for the MRCMM and HMGL-RSM With SES as a Person Covariate

	MRCMM		HMGL-RSM	
	M	SE	M	SE
δ_1	0.09	(0.36)	0.09	(0.36)
δ_2	0.35	(0.36)	0.35	(0.36)
δ_3	-0.68	(0.36)	-0.68	(0.36)
δ_4	-1.31	(0.36)	-1.31	(0.36)
δ_5	-0.51	(0.36)	-0.51	(0.36)
δ_6	-0.37	(0.36)	-0.37	(0.36)
δ_7	-0.62	(0.36)	-0.62	(0.36)
δ_8	0.39	(0.36)	0.39	(0.36)
δ_9	0.48	(0.36)	0.48	(0.36)
δ_{10}	-0.44	(0.36)	-0.44	(0.36)
τ_1	2.15	.	2.15	.
τ_2	-2.15	(0.03)	-2.15	(0.03)
$\mu_{\theta 1}$	-0.12	(2.42)	-0.12	(2.42)
μ_u	0.12	(2.42)	0.12	(2.42)
$\lambda_{0\cdot,1}$	-0.24	(0.20)	-0.24	(0.20)
$\mu_{\theta 2}$	-0.74	(2.65)	-0.74	(2.65)
μ_u	-0.26	(2.65)	-0.26	(2.65)
$\lambda_{0\cdot,1}$	-0.24	(0.20)	-0.24	(0.20)
$\mu_{\theta 3}$	-0.15	(2.52)	-0.15	(2.52)
μ_u	0.57	(2.52)	0.57	(2.52)
$\lambda_{0\cdot,1}$	-0.24	(0.20)	-0.24	(0.20)
μ_θ	-0.40		-0.40	
μ_u	-0.01		-0.01	
Σ_θ	2.80	(0.12)	2.80	(0.12)
AIC	7147.3		7147.3	
BIC	7206.3		7206.3	

Note. $\{\delta_1, \delta_2, \ldots, \delta_{10}\}$ = location for items 1 – 10. $\{\tau_1, \tau_2\}$ = thresholds 1 and 2. $\lambda_{0\square,1}$ = Effect of SES. μ_θ = Overall Mean person location. $\{\mu_{\theta 1}, \mu_{\theta 2}, \mu_{\theta 3}\}$ = Mean for person locations of high, medium, and low SES groups. Σ_θ = Standard deviation of the person locations. AIC = Akaike Information Criterion. BIC = Bayesian Information Criterion. Est. = Estimate. SE = Standard error.

Nevertheless, the HMGL-RSM may be the preferred model, because it is expected that if the number of observations was increased for the MRCMM, as the developers intended it to be, then the estimates would be somewhat different. And as mentioned above, the MRCMM was not defined as being able to model additional hierarchical levels that predict how the item parameters behave, which may be important (e.g., see Section 5).

To illustrate the influence of SES, one now compares the HMGL-RSM with and without SES (Table 10 in Chapter 3). As one can see, when SES is not included in the model, the overall mean person location is centered near zero $\left(\mu_{\hat{\theta}} = -.01\right)$, as would be expected. Additionally, if SES is not modeled, the low, medium, and high SES groups have mean person locations equaling .27, -.34, and .26, respectively. Notice, then, that the mean person location of the high SES group is actually lower than the location for the low SES group. Also, the middle SES group is nearly one logit lower than both the high and low SES groups.

In contrast, when controlling for SES, the overall mean of the random effect of persons is centered near zero, $\mu_{\hat{u}} = -.01$. Notice this value is similar to the overall mean person location when SES is not accounted for. This is expected because, recall, the mean of the random effect of persons (u) is set to zero, and when SES is not modeled, $\theta = u$.

However, by modeling SES, we see that its effect on the person location $\left(\lambda_{0\cdot,1}\right)$ is -.24. Hence, as a person increases in one unit in SES (i.e., increases in poverty), his location decreases. Thus, by including SES, the overall person location decreases by almost half a logit $\left(\mu_{\hat{\theta}} = -.40\right)$. Hence, if the parent's SES is controlled for—that is, we

ignore the effects of the parent's SES—then, the average person's location on the underlying continuum is almost half a logit higher.

For example, the mean location $\left(\mu_{\hat{\theta}}\right)$ of the high, medium, and low SES groups is -.12, -.74 and -.15. However, notice after controlling for SES, the mean location $\left(\mu_{\hat{u}}\right)$ of the high, medium, and low SES groups now becomes .12, -.26, and .57. Although the rankings are the same to the rank orderings when SES is not controlled for, notice now that by controlling for SES, the groups' mean location increases. Additionally, the difference in mean locations between the groups becomes larger at nearly a half a logit.

So which is the better model for the data: the HMGL-RSM with SES or without SES? Examining the AIC and BIC values, we see that for both the AIC and BIC, the lower values are for the HMGL-RSM without SES. Furthermore, when inspecting the information weights, the AIC and BIC weights for the HMGL-RSM without SES are .57 and 1.00, while the AIC and BIC weights for the HMGL-RSM with SES are .43 and .08. Since the AIC and BIC are lower for the HMGL-RSM without SES, and since higher weights indicate the model is more likely, the evidence suggests that the HMGL-RSM without SES is the better fitting model.

Before this section is concluded, the reader should notice that the difference between the item locations for each item of the two the models is -.4. This difference does not necessarily indicate that by including SES in the model, the item location decreases by -.4; rather, it indicates the arbitrariness of the IRT scale. That is, recall from Section 2-4, the IRT scale is indeterminate, and the indeterminacy is resolved by centering on the normally distributed person measures, where the mean is equal to zero. By including the covariate SES in the model, the mean of the scale changes. Specifically,

in contrast to before, when not including SES, $\mu_\theta = \mu_u = 0$. However by including SES,

$$\mu_\theta = \mu_{\lambda_{0\cdot,1} w_{0\cdot k,1} + u} = \mu_{\lambda_{0\cdot,1} w_{0\cdot k,1}} + \mu_u = \mu_{\lambda_{0\cdot,1} w_{0\cdot k,1}} + 0 = -.4.$$

Chapter 5. Extending the HMGL-RSM To Include a Group Level

5-1. The Four-Level HMGL-RSM

As seen in Chapter 4, one advantage of applying the HMGLM to model the RSM is that the user may posit a model that includes person covariates. Another advantage is that the user may posit a model that includes a group level, which defines how the item parameters behave across groups. Hence, a Four-Level HMGL-RSM is defined. This form of the HMGL-RSM may be especially important in educational testing during investigations of differential item functioning (DIF).

To model the Four-Level HMGL-RSM, four models are defined. The Level-1, -2, and -3 models follow the previous definitions of the HMGL-RSM, in which the category is nested within the item, which in turn is nested within the person (Section 2-3). For the 4-Level HGL-RSM, the Level-4 model is defined for the group level, where persons are nested within groups.

5-1-1. The Level-1 Model

The Level-1 model (the category level) is defined as

$$\log\left(\frac{\pi_{ijkl}}{\pi_{i-1,jkl}}\right) = \sum_{j=1}^{J} \beta_{jkl}^{(i)} x_{jkl}\,, \tag{5.1}$$

where $\beta_{jkl}^{(i)}$ is the mean category effect if person k in group l selects category i of item j; and x_{jkl} is a dummy variable with values 1 if person k in group l answers item j, and 0 otherwise.

5-1-2. The Level-2 Model

The Level-2 model (the item level) is defined as

$$\beta_{jkl}^{(i)} = \gamma_{0jkl} + \sum_{i=1}^{I} \gamma_{1\cdot kl}^{(i)} w_{1\cdot kl}^{(i)}, \tag{5.2}$$

where, for person k in group l, γ_{0jkl} is the mean effect of item j across categories i; $\gamma_{1\cdot kl}^{(i)}$ is the effect of an item on a particular category i; and $w_{1\cdot kl}^{(i)}$ is a dummy variable with values 1 if $i' = i$ for the j^{th} item answered by person k in group l, and 0 otherwise. For identifiability, $\gamma_{1\cdot kl}^{(0)} \equiv 0$.

Here, notice that before, the item effects only varied across persons. Now, not only do the item effects vary for each person k, but the item effects vary for each group l as well. To see how the effects vary, the person level model (Level 3) and the group level model (Level 4) are defined.

5-1-3. The Level-3 Model

The Level-3 model (the person level) is defined as

$$\gamma_{0jkl} = \lambda_{0j0l} + u_{kl}, \tag{5.3}$$

$$\gamma_{1\cdot kl}^{(i)} = \lambda_{1\cdot 0l}^{(i)}, \tag{5.4}$$

where, for the j^{th} item that is answered by person k in group l, λ_{0j0l} is the mean effect of persons for group l on item j; u_{kl} is the random effect of person k in group l on the mean effect of item j; and $\lambda_{1\cdot 0l}^{(i)}$ is the mean change in λ_{0j0l} for a particular category of the items.

However in IRT, we assume that the person effects are not only constant across items, but constant regardless of group as well. Thus, the following constraint is made

$$u_{k1} = u_{k2} = \ldots = u_{kl} = u_k ,$$

and the Level-3 model for the mean item effect becomes

$$\gamma_{0jkl} = \lambda_{0j0l} + u_k , \tag{5.5}$$

where λ_{0j0l} is defined above; and u_k is the random effect of person k (regardless of group) across items.

Here, it is helpful to refer back to the honesty example, for we can clearly see how the category effects function as the categories are nested within items, which in turn are nested within persons. Specifically, as mentioned above, the probability that an applicant is attracted to a particular feeling for a particular answered item not only depends upon the overall attractiveness of the item $\left(\lambda_{0j0l}\right)$, but also how the attractiveness of the item influences a particular feeling $\left(\lambda_{1\cdot 0l}^{(i)}\right)$. In addition, as the Level-3 model shows, the overall attractiveness of the item $\left(\lambda_{0j0l}\right)$ and the influence of an item on a particular feeling $\left(\lambda_{1\cdot 0l}^{(i)}\right)$ is fixed across persons, but may vary across l groups. Lastly, as is commonly assumed in IRT, the unique effect of an applicant randomly varies across the different applicants.

5-1-4. The Level-4 Model

Lastly, the Level-4 model (the group level) is defined as

$$\lambda_{0j0l} = \xi_{0j00} + \sum_{l=1}^{l-1} \xi_{0j0l} z_{0j0l}, \tag{5.6}$$

$$\gamma_{1\cdot kl}^{(i)} = \xi_{1\cdot 00}^{(i)}, \tag{5.7}$$

where, for the j^{th} item that is answered by person k in group l, ξ_{0j00} is the mean effect of groups on item j; ξ_{0j0l} is the mean change in ξ_{0j00} as group membership changes; $\xi_{1\cdot 00}^{(i)}$ is the mean change in ξ_{0j00} for a particular category of item j; and z_{0j0l} is a dummy variable with values 1 if person k is a member of a particular group l, and 0 otherwise.

Again, one refers back to the honesty example. In the group level model, we can see how the overall attractiveness of the item $\left(\lambda_{0j0l}\right)$ depends on group membership. For example, if an applicant belongs to the baseline group, such as Caucasian, then the overall attractiveness of the item for Caucasians is given as ξ_{0j00}. However, if an applicant belongs to a comparison group, such as Asians, then the overall attractiveness of the item for Asians is given as $\xi_{0j00} + \xi_{0j01}$. Additionally, notice the attractiveness of the item for a particular feeling $\left(\lambda_{1\cdot 0l}^{(i)}\right)$ remains fixed not only for different persons, but for different groups as well $\left(\xi_{1\cdot 00}^{(i)}\right)$.

5-1-5. <u>The Combined Model</u>

The combined model of the 4-Level HMGL-RSM reduces to the following for a particular category i of item j

$$\log\left(\frac{\pi_{ijkl}}{\pi_{i-1,jkl}}\right)=\xi_{0j00}+\sum_{l=1}^{l-1}\xi_{0j0l}z_{0j0l}+\xi_{1\cdot00}^{(i)}+u_k, \tag{5.8}$$

where all terms are defined above.

Therefore, the parameters of the HMGL-RSM are related to and extend the parameters of the traditional RSM in the following manner:

$$\begin{aligned}
&\delta_{j0}=-\xi_{0j00}\\
&\delta_{j1}=-\left(\xi_{0j00}+\xi_{0j01}\right)\\
&\delta_{j2}=-\left(\xi_{0j00}+\xi_{0j02}\right)\\
&\cdots\\
&\delta_{j,l-1}=-\left(\xi_{0j00}+\xi_{0j0,l-1}\right)
\end{aligned}, \tag{5.9}$$

$$\tau_i=-\xi_{1\cdot00}^{(i)}, \tag{5.10}$$

and

$$\theta_k=u_k, \tag{5.11}$$

where τ_i and θ_k are defined above; δ_{j0} is the location of the item on the underlying continuum for the baseline group; and δ_{jl} is the location of the item on the underlying continuum for a particular group l.

5-2. Simulation Study for the Four-Level HMGL-RSM

The following section describes a simulation study for the Four-Level HMGL-RSM. Since Section 3-2 already described a simulation study that examined the parameter recovery of the person and item parameters when a fourth level was not added to the HMGL-RSM, the focus of this section is to examine the behaviors of the item parameters when being influenced by the additional level. Specifically, the purposes of

the following section is (1) to determine the precision of the parameter recovery for the person and item parameters—in particular, the item parameters at the group level, and (2) to determine the accuracy of a statistical test to detect the influence of a group-level coefficient as a measure of DIF.

5-2-1. Design

The design of the simulation is as follows. Observations were simulated using the HMGL-RSM. For the study, 500 simulees from 2 groups ($l = 0, 1$) responded to 10 polytomous items, where each item consisted of 3 categories i ($i = 0, 1, 2$). The number of groups, simulees, items, and categories were chosen to follow typical data from a questionnaire (e.g., Dodd, 1990; Smith & Johnson, 2000; Zhu, Updyke, & Lewandowski, 1997) or a large-scale assessment (e.g., Michigan Education Assessment Program, 2003; U.S. Department of Education, 1999). In addition, the number of simulees and items were chosen because, as shown in Section 3-2, these sample sizes allow for reasonable precision (at least when a four-level model was not employed).

To produce the simulated responses, each simulee k in group l was randomly assigned a location θ_{kl}, $\theta \sim N(0,1)$. Additionally, each item j was randomly assigned a set of item parameters. These item parameters were selected to represent parameter estimates from typical polytomous data, and follow those that are presented in Table 2 for a confidential readiness assessment. The items that were selected to be simulated were randomly chosen to be the first 10 items of the confidential readiness assessment that did not exhibit DIF between males and females (Table 16). By selecting only non-DIF items (in regards to gender DIF), the influence of DIF by the non-focus items was minimized.

Table 16. DIF results for the Mantel-Haenszel test

Original Item	Simulation Item	M^2	p
1	1	0.52	0.471
2	2	0.37	0.545
3		76.52	0.000*
4		74.91	0.000*
5		36.46	0.000*
6	3	0.15	0.699
7	4	0.77	0.379
8	5	0.16	0.688
9		38.89	0.000*
10	6	8.17	0.004
11	7	0.12	0.731
12	8	0.39	0.532
13		13.28	0.000*
14		31.21	0.000*
15		16.13	0.000*
16		18.60	0.000*
17	9	7.75	0.005
18		17.38	0.000*
19	10	0.70	0.403
20		2.94	0.086
21		9.23	0.002*
22		0.09	0.760
23		2.85	0.091
24		0.03	0.871
25		9.79	0.002

Note. M^2 = Mantel-Haenszel test statistic. p = p-value. * = statistically significant at $\alpha = .05/25 = .002$. p = .000 implies p < .0001.

The Mantel-Haenszel (MH) test (Mantel, 1963) was used as the original test for DIF. This test was selected because it has been well-studied (e.g., Kim, 2000), and has

been typically used in DIF analyses of polytomous data (e.g., U.S. Department of Education, 1999).

Thus, using θ_{kl}, δ_{jl}, and τ_i, three response probabilities for each simulee by item combination were produced, $P_{0jkl}(\theta)$, $P_{1jkl}(\theta)$, and $P_{2jkl}(\theta)$. If $\sum_{0}^{i'} P_{i'jkl}(\theta) < Y_{jk} \leq \sum_{0}^{i'+1} P_{i'jkl}(\theta)$, then simulee k in group l was assigned a response of $i'+1$ for item j; otherwise a response of 0 was assigned. Note that $i' = 0, 1$; and Y_{jk} was a single, random number for each $j \times k$ combination, $Y \sim U(0,1)$.

The simulation manipulated three variables: (1) the proportion of simulees in the focus group, (2) the difference in the mean location of the person parameters for the reference group $(\bar{\theta}_0)$ and the focal group $(\bar{\theta}_1)$, and (3) the level of DIF in the focus item. Each variable and each condition (described below) was chosen because previous research found these to influence DIF detection (Luppescu, 2002).

The conditions for the proportion of simulees in the focus group varied between 10% (50 simulees) and 25% (125 simulees). This represented a testing situation where the focus group was small or moderate in size.

The conditions for the difference in mean location varied such that θ_0 was randomly sampled from $N(0,1)$, and θ_1 was randomly sampled from $N(-1,1)$ or $N(-.5,1)$. This represented a testing situation where, on average, the focus group had a moderately lower or somewhat lower person location than the reference group.

Lastly, the conditions for the level of DIF in the focus item (which was arbitrarily chosen to be item 1 in Table 16) varied for the focus group by a positive difference of 1

standard error (.07) or 2 standard errors (.14). This represented a testing situation where the focus item displayed a small or moderate effect of DIF; that is, the focus item was somewhat or moderately less attractive to endorse for the focus group. (Note the standard error for item 1 was found in Table 5 of Section 3-2-1 and chosen to be the standard error when 1000 persons responded to 10 items.)

The simulation procedure utilized a fully crossed 2 × 2 × 2 factorial design that simulated 8 conditions. Each administration was iterated 50 times producing 400 unique response data matrices. The number of iterations was chosen because Kamata (1998) showed this to be a reasonable number for obtaining stable estimates. S-Plus (2000) was used to generate all data. SAS (2001) was used to obtain parameter estimates and conduct significance tests.

5-2-2. Analysis

For the analysis regarding the parameter recovery of the Four-Level HMGL-RSM, the RMSE for δ_{jl} and τ_i was obtained over the iterations for each condition. Specifically, the RMSE was obtained by

$$RMSE(\hat{\omega}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\omega}_n - \omega_n)^2}, \qquad (5.12)$$

where the maximum number of n iterations was $N = 50$; and ω is an arbitrary parameter representing either δ_{jl} or τ_i. A descriptive analysis of the RMSE was conducted for each condition.

For the analysis regarding the accuracy of a statistical test to detect DIF: a t-test with $\alpha = .05$ is applied to examine the following hypotheses:

$$H_0 : \xi_{0101} = 0$$

$$H_1 : \xi_{0101} \neq 0$$

Thus, if H_0 is not rejected, then there is statistical evidence to suggest that ξ_{0101} does not significantly differ from zero, and no DIF exists. That is, the location of item 1 for each group is equal

$$\begin{aligned} \delta_{1,1} &= -(\xi_{0100} + \xi_{0101}) \\ &= -(\xi_{0100}) \\ &= \delta_{1,0}. \end{aligned}$$

If H_0 is rejected, then there is statistical evidence to suggest that ξ_{0101} significantly differs from zero, and DIF exists. That is, the location of item 1 for each group is not equal

$$\begin{aligned} \delta_{1,1} &= -(\xi_{0100} + \xi_{0101}) \\ &\neq -(\xi_{0100}) \\ &\neq \delta_{1,0}. \end{aligned}$$

Thus to examine the accuracy, if H_0 was rejected, then a 'hit' was made; otherwise a 'miss' was made. The number of hits across iterations for a condition was defined as the hit rate, i.e., the accuracy of the t-test for detecting DIF (when DIF exists) under the aforementioned conditions. A descriptive analysis of the hit rate was conducted for each condition.

Note Cheong and Raudenbush (2000), Kamata (1998), Luppescu (2002), and Kim (2003) describe and illustrate similar DIF analyses using a two-level, hierarchical IRT model for dichotomous data, in which the covariates for the item parameters were added at the item level rather than a group level. Although the model presented above will

reduce to an analogous formulation of the aforementioned models, the model that was defined may be preferable because users are given the option of specifying a random component at the group level. Although one did not include the random component here since it was not of interest, other users may wish to examine this component as a measure of the group location across the items.

5-2-3. Results: Descriptive Statistics

Displayed in Tables 17 and 18 are the mean and standard deviations of the parameter estimates for the Four-Level HMGL-RSM when the proportion of simulees in the focus group was 10% and 25%, respectively. As can be seen, the standard deviations of the estimates are similar and fairly low across conditions, except for $\hat{\delta}_{1,1}$. For $\hat{\delta}_{1,1}$, as the proportion of simulees in the focus group increased from 10% to 25%, the standard deviation decreased from a moderate to somewhat moderate magnitude, as would be expected. This suggested that PROC NLMIXED obtained relatively consistent estimates of the HMGL-RSM parameters, especially as the group size increased.

Table 17. Mean and Standard Error of the Parameter Estimates for the Four-Level HMGL-RSM for Proportion = 10%

	$\bar{\theta}_2$ = -.5				$\bar{\theta}_2$ = -1			
	1 SD		2 SD		1 SD		2 SD	
	M	SE	M	SE	M	SE	M	SE
$\hat{\delta}_{1,0}$	-0.05	(0.12)	-0.05	(0.12)	-0.02	(0.12)	-0.02	(0.12)
$\hat{\delta}_{1,1}$	0.06	(0.31)	0.13	(0.31)	0.28	(0.33)	0.34	(0.30)
$\hat{\delta}_2$	0.08	(0.09)	0.08	(0.09)	0.12	(0.09)	0.12	(0.09)
$\hat{\delta}_3$	-0.84	(0.11)	-0.84	(0.11)	-0.79	(0.12)	-0.79	(0.12)
$\hat{\delta}_4$	-1.45	(0.11)	-1.45	(0.11)	-1.40	(0.10)	-1.40	(0.10)
$\hat{\delta}_5$	-0.67	(0.11)	-0.67	(0.11)	-0.63	(0.11)	-0.63	(0.11)
$\hat{\delta}_6$	-0.71	(0.10)	-0.71	(0.10)	-0.65	(0.10)	-0.65	(0.10)
$\hat{\delta}_7$	-0.78	(0.10)	-0.78	(0.10)	-0.73	(0.10)	-0.73	(0.10)
$\hat{\delta}_8$	-0.06	(0.11)	-0.06	(0.11)	-0.01	(0.11)	-0.01	(0.11)
$\hat{\delta}_9$	0.12	(0.11)	0.12	(0.11)	0.17	(0.11)	0.17	(0.11)
$\hat{\delta}_{10}$	-0.71	(0.09)	-0.71	(0.09)	-0.65	(0.09)	-0.65	(0.09)
$\hat{\tau}_1$	-2.23	(0.04)	-2.23	(0.04)	-2.23	(0.04)	-2.23	(0.04)
$\hat{\tau}_2$	2.23	(0.04)	2.23	(0.04)	2.23	(0.04)	2.23	(0.04)

Note. $\bar{\theta}_2$ = mean location of focus group. SD = standard deviation shift in item 1 for focus group. $\{\hat{\delta}_{1,0}, \hat{\delta}_{1,1}\}$ = location for item 1 for the reference (0) and focal (1) groups. $\{\hat{\delta}_2, \hat{\delta}_3, \ldots, \hat{\delta}_{10}\}$ = location for items 2 – 10 for both groups. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. M = mean. SE = standard error.

Table 18. Mean and Standard Error of the Parameter Estimates for the Four-Level HMGL-RSM for Proportion = 25%

	$\bar{\theta}_2$ = -.5				$\bar{\theta}_2$ = -1			
	1 SD		2 SD		1 SD		2 SD	
	M	SE	M	SE	M	SE	M	SE
$\hat{\delta}_{1,0}$	0.01	(0.14)	0.01	(0.14)	0.10	(0.14)	0.10	(0.14)
$\hat{\delta}_{1,1}$	0.16	(0.19)	0.23	(0.20)	0.38	(0.21)	0.45	(0.19)
$\hat{\delta}_2$	0.15	(0.08)	0.15	(0.08)	0.27	(0.08)	0.27	(0.08)
$\hat{\delta}_3$	-0.75	(0.11)	-0.75	(0.11)	-0.63	(0.11)	-0.63	(0.11)
$\hat{\delta}_4$	-1.25	(0.11)	-1.25	(0.11)	-1.12	(0.10)	-1.12	(0.10)
$\hat{\delta}_5$	-0.46	(0.10)	-0.46	(0.10)	-0.34	(0.10)	-0.34	(0.10)
$\hat{\delta}_6$	-0.64	(0.10)	-0.64	(0.10)	-0.52	(0.10)	-0.52	(0.10)
$\hat{\delta}_7$	-0.77	(0.10)	-0.77	(0.10)	-0.65	(0.10)	-0.65	(0.10)
$\hat{\delta}_8$	-0.13	(0.11)	-0.13	(0.11)	0.00	(0.11)	0.00	(0.11)
$\hat{\delta}_9$	0.17	(0.11)	0.17	(0.11)	0.28	(0.11)	0.28	(0.11)
$\hat{\delta}_{10}$	-0.54	(0.10)	-0.54	(0.10)	-0.41	(0.10)	-0.41	(0.10)
$\hat{\tau}_1$	-2.21	(0.04)	-2.21	(0.04)	-2.21	(0.04)	-2.21	(0.04)
$\hat{\tau}_2$	2.21	(0.04)	2.21	(0.04)	2.21	(0.04)	2.21	(0.04)

Note. $\bar{\theta}_2$ = mean location of focus group. SD = standard deviation shift in item 1 for focus group. $\{\hat{\delta}_{1,0}, \hat{\delta}_{1,1}\}$ = location for item 1 for the reference (0) and focal (1) groups. $\{\hat{\delta}_2, \hat{\delta}_3, \ldots, \hat{\delta}_{10}\}$ = location for items 2 – 10 for both groups. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. M = mean. SE = standard error.

As for the mean of the estimates: In general, the estimates obtained by PROC NLMIXED appeared to differ slightly from the parameter values (c.f., Table 2). Specifically, trends indicated that the level of DIF did not influence the mean of the estimates. However, it appeared that as the proportion of simulees in the focus group increased and as the mean location of the focus group decreased, the mean of the

estimates generally deviated from the parameter values by a positive magnitude. Below, the RMSE is examined.

5-2-4. Results: RMSE

The results of the RMSE for the item parameters of the Four-Level HMGL-RSM are provided in Table 19. As alluded to above, trends indicated that as the level of DIF increased, the RMSE did not vary across the conditions substantially. This is expected because, as shown in Section 3-2-2, the location of the item does not influence the RMSE.

Table 19. RMSE for the Four-Level HMGL-RSM

	10%				25%			
	$\bar{\theta}_{.2}$ = -.5		$\bar{\theta}_{.2}$ = -1		$\bar{\theta}_{.2}$ = -.5		$\bar{\theta}_{.2}$ = -1	
SD	1	2	1	2	1	2	1	2
$\hat{\delta}_{1,0}$	0.13	0.13	0.14	0.14	0.17	0.17	0.23	0.23
$\hat{\delta}_{1,1}$	0.32	0.32	0.44	0.42	0.26	0.27	0.45	0.44
$\hat{\delta}_2$	0.11	0.11	0.14	0.14	0.15	0.15	0.26	0.26
$\hat{\delta}_3$	0.14	0.14	0.17	0.17	0.21	0.21	0.31	0.31
$\hat{\delta}_4$	0.16	0.17	0.20	0.20	0.34	0.34	0.46	0.46
$\hat{\delta}_5$	0.17	0.17	0.21	0.21	0.36	0.36	0.48	0.48
$\hat{\delta}_6$	0.11	0.11	0.13	0.13	0.14	0.14	0.25	0.25
$\hat{\delta}_7$	0.10	0.10	0.12	0.12	0.11	0.11	0.19	0.19
$\hat{\delta}_8$	0.12	0.12	0.11	0.11	0.16	0.16	0.10	0.10
$\hat{\delta}_9$	0.12	0.12	0.15	0.15	0.15	0.15	0.24	0.24
$\hat{\delta}_{10}$	0.17	0.17	0.22	0.22	0.33	0.33	0.45	0.45
$\hat{\tau}_1$	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.05
$\hat{\tau}_2$	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.05

Table 19 (cont'd)

Note. $\{10\%, 25\%\}$ = percentage of sample in focus group. $\bar{\theta}_2$ = mean location of focus group. SD = standard deviation shift in item 1 for focus group. $\{\hat{\delta}_{1,0}, \hat{\delta}_{1,1}\}$ = location for item 1 for the reference (0) and focal (1) groups. $\{\hat{\delta}_2, \hat{\delta}_3, \ldots, \hat{\delta}_{10}\}$ = location for items 2 – 10 for both groups. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. M = mean. SE = standard error.

Additionally, as the proportion of simulees in the focus group increased and as the mean location of the focus group decreased, the RMSE generally increased. The one exception occurs for $\hat{\delta}_{1,1}$ when $\bar{\theta}_2$ = -.5. In this case, as the proportion of simulees in the focus group increased, the RMSE decreased.

Additionally, as the proportion of simulees in the focus group increased, the magnitude of the RMSE generally increased from a low range (.04 to .22) to a moderate range (.32 to .42). For $\hat{\delta}_{1,1}$, the RMSE increased from a range of .32 to .44 to a range of .26 to .45. These trends and magnitudes suggest that the sample characteristics of the focal group influence the empirical Bayes estimates of not only the focal group, but the non-focal group as well.

5-2-5. Results: Accuracy

As for the accuracy of the t-test for detecting DIF (when DIF exists), the results show the hit rates were low (Table 20), but still moderately higher than the MH test (Table 21). Also, trends indicated that the hit rates increased as (1) the level of DIF increased, (2) the mean location of the focus group decreased, and (3) the proportion in the focus group increased. Thus, although the hit rates for detecting DIF with the HMGL-

RSM were low, it is expected that increasing the sample and group size should increase hit rates as well, and further set itself apart from the MH test. This provides some evidence for the use of the HMGL-RSM as a test for DIF.

Table 20. Hit Rates for Detecting DIF with the HMGL-RSM

	10%		25%	
	1 SD	2 SD	1 SD	2 SD
$\bar{\theta}_2 = -5$	0.06	0.10	0.10	0.16
$\bar{\theta}_2 = -1$	0.16	0.22	0.26	0.38

Note. $\{10\%, 25\%\}$ = percentage of sample in focus group. $\bar{\theta}_2$ = mean location of focus group. SD = standard deviation shift in item 1 for focus group.

Table 21. Hit Rates for Detecting DIF with the MH test

	10%		25%	
	1 SD	2 SD	1 SD	2 SD
$\bar{\theta}_2 = -5$	0.04	0	0.10	0.10
$\bar{\theta}_2 = -1$	0.02	0	0.12	0.08

Note. $\{10\%, 25\%\}$ = percentage of sample in focus group. $\bar{\theta}_2$ = mean location of focus group. SD = standard deviation shift in item 1 for focus group.

5-3. Example Analysis of the Four-Level HMGL-RSM

The purpose of this section is to provide an example analysis that illustrates the basic concepts of the 4-Level HMGL-RSM. In particular, one illustrates how to use the model to detect DIF between males and females.

5-3-1. Design

The design of the analysis is as follows. Five hundred respondents were randomly selected from a larger sample of students that responded to a confidential readiness assessment. Note this was the same assessment simulated in Sections 3-1, 4-2, and 5-2. In this sample, 53% were male, and 47% were female, as was the case in the original sample. Additionally, approximately 1% were age 5; 26% were age 6; 67% were age 7; and 6% were age 8. Lastly, approximately 1% were Asian; 48% were African-American; 8% were Hispanic; and 42% were age Caucasian.

For the purposes of this illustration, only the first 10 items of the assessment were used. (Note each item measured the person's personal and social development.) Additionally, only those respondents who answered each item and provided their gender were used. As illustrated in Sections 3-1 and 5-2, the sample and item sizes were adequate to obtain relatively precise parameter estimates and moderately accurate DIF tests.

5-3-2. Analysis

To analyze the responses of the students, PROC NLMIXED of SAS (2001) was used to estimate the person and item parameters for the Four-Level HMGL-RSM. Recall that the four levels of this model are given above. The group predictor that was used was Gender, in which Males was the reference group (0), and Females was the focus group (1).

For each item, the following hypotheses are examined

$$H_0 : \xi_{0j01} = 0$$

$$H_1 : \xi_{0j01} \neq 0,$$

where j = 1, ..., 10.

Thus for a particular item j, if H_0 was not rejected, then there was statistical evidence to suggest that DIF does not exist. Likewise, if H_0 was rejected, then there was statistical evidence to suggest that DIF exists.

Additionally, for comparative purposes, the MH test was conducted. As mentioned above, this test was selected because it has been well studied (e.g., Kim, 2000), and has been typically used in DIF analyses of polytomous data (e.g., U.S. Department of Education, 1999). Also, note previous simulation research has suggested that similar findings occur if no purification procedures were used, two stage purification procedures were used, or an iterative purification process was used (Wang & Su, 2004). Hence, because similar DIF results are obtained regardless of purification procedures, and because research has shown that the two stage and iterative purification procedures become inefficient when used in conditions similar to those studied here (Donoghue, Holland, & Thayer, 1993 as cited by Wang & Su, 2004), the decision to not apply any purification was made.

To examine, if the t-test for ξ_{0j01} and MH test for item j was accurate, the results of the analyses was compared to the DIF results found for the larger sample (Table 16). As shown, of the first 10 items, it was found that items 3-5 and 9 exhibit DIF between Males and Females.

5-3-3. Results

The results of the analysis are presented in Table 22. As can be seen, the t-test was fairly conservative at flagging DIF, while the MH test was not. Specifically, the t-test correctly identified items 3-5 and 9 as exhibiting DIF. However, the t-test also incorrectly identified items 7, 8, and 10 as exhibiting DIF. In contrast, the MH test only correctly identified item 9, and incorrectly identified item 1 as exhibiting DIF. Although the Type I error may be high for the HMGL-RSM, this may be preferable because the consequences may be greater if DIF was not flagged rather than flagged. Thus, although the Type I error may be high, it appears that the t-test was more powerful at detecting DIF than the MH test.

One reason the HMGL-RSM may be more powerful at detecting DIF than the MH test is that the HMGL-RSM is based on parametric methods, while the MH test is not. That is, the HMGL-RSM is based on the HMGLM framework which attempts to explicitly model the parameters that characterize the DIF. And, as shown above, the HMGL-RSM is estimated rather precisely; hence the parameters that characterize the DIF may be estimated rather precisely as well.

Table 22. Item Analysis of a Real Data Set

	HMGL-RSM					MH	
Item	Par.	Est.	SE	t	p	M^2	p
1	$\hat{\xi}_{0100}$	-1.49	0.22	-6.66	0.00	6.20	0.01 [b]
	$\hat{\xi}_{0101}$	0.31	0.33	0.95	0.34		
2	$\hat{\xi}_{0200}$	-1.87	0.23	-8.30	0.00	4.03	0.04
	$\hat{\xi}_{0201}$	0.63	0.33	1.91	0.06		
3	$\hat{\xi}_{0300}$	-1.21	0.22	-5.46	0.00	5.46	0.02
	$\hat{\xi}_{0301}$	1.52	0.33	4.61	0.00 [a]		
4	$\hat{\xi}_{0400}$	-0.68	0.22	-3.12	0.00	5.83	0.02
	$\hat{\xi}_{0401}$	1.52	0.33	4.61	0.00 [a]		
5	$\hat{\xi}_{0500}$	-0.94	0.22	-4.25	0.00	0.01	0.93
	$\hat{\xi}_{0501}$	1.13	0.33	3.45	0.00 [a]		
6	$\hat{\xi}_{0600}$	-0.82	0.22	-3.73	0.00	3.34	0.07
	$\hat{\xi}_{0601}$	0.59	0.33	1.80	0.07		
7	$\hat{\xi}_{0700}$	-1.13	0.22	-5.11	0.00	0.04	0.84
	$\hat{\xi}_{0701}$	0.92	0.33	2.81	0.01 [b]		
8	$\hat{\xi}_{0800}$	-1.90	0.23	-8.39	0.00	0.04	0.84
	$\hat{\xi}_{0801}$	0.99	0.33	3.00	0.00 [b]		
9	$\hat{\xi}_{0900}$	-2.25	0.23	-9.83	0.00	6.23	0.01 [a]
	$\hat{\xi}_{0901}$	1.62	0.33	4.87	0.00 [b]		
10	$\hat{\xi}_{0,10,00}$	-1.11	0.22	-5.03	0.00	0.24	0.62
	$\hat{\xi}_{0,10,01}$	0.93	0.33	2.82	0.01 [b]		
	$\hat{\tau}_1$	-2.11	.	.	.		
	$\hat{\tau}_2$	2.11	0.04	-60.31	0.00		

Note. Par. = parameter. Est. = estimate. SE = standard error. t = t-statistic. p = p-value. M^2 = Mantel-Haenszel test statistic. $\{\hat{\xi}_{0j00}, \hat{\xi}_{0j01}\}$ = overall attractiveness of item j for Males and Females, respectively. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. $\alpha = \cdot 1/10 = .01$. p = 0.00 implies p < .001. [a] = correct flag for DIF. [b] = incorrect flag for DIF.

To interpret the item parameters, recall that if the item does not exhibit DIF, then $\xi_{0j01} = 0$ and

$$\begin{aligned}\delta_{j1} &= \delta_{j0} \\ &\equiv \delta_j \\ &= -\xi_{0100}.\end{aligned}$$

If the item exhibits DIF, then $\xi_{0j01} \neq 0$ and

$$\begin{aligned}\delta_{j0} &= -\xi_{0j00} \\ \delta_{j1} &= -\left(\xi_{0j00} + \xi_{0j01}\right).\end{aligned}$$

For example, for item 1, the t-test was not statistically significant for $\hat{\xi}_{0101}$; hence, $\hat{\delta}_{j0} = \hat{\delta}_{j1} = \hat{\delta}_1 = -\hat{\xi}_{0100} = 1.49$. Similarly, for item 2, the t-test was not statistically significant for $\hat{\xi}_{0101}$; hence $\hat{\delta}_2 = -\hat{\xi}_{0200} = 1.87$. In other words, for item 1, the log-odds of the overall attractiveness of the item is 1.49 for a typical respondent, while for item 2, the log-odds of the overall attractiveness of the item is 1.87. Thus, item 1 has a lower overall attractiveness than item 2, which suggests that the polytomous alternatives for item 1 are more easier to endorse than those for item 2, for a typical respondent regardless of gender.

In contrast to items 1 and 2, for item 3 the t-test was statistically significant for $\hat{\xi}_{0101}$; hence for Males, $\hat{\delta}_{3,0} = -\xi_{0300} = 1.21$, and for Females, $\hat{\delta}_{3,1} = -\left(\xi_{0300} + \hat{\xi}_{0301}\right) = -\left(-1.21 + 1.52\right) = -.31$. Thus, the overall attractiveness of item 3 is substantially lower for Females than for Males. This suggests that the polytomous alternatives for item 3 are easier to endorse for Females than for Males.

(As an aside, the reader should note that the item location for all items is lower for Females than for Males. In other words, the items are easier for Females than for Males. However, this does makes sense because each of the studied items measures a person's personal and social development, and it is well known that Female children are more advanced in terms of personal and social development than Male children. Hence, the items are expected to be easier for Females than for Males.)

Chapter 6. Extending the HMGL-RSM To Include Item Covariates

6-1. The HMGL-RSM with Person Covariates

As seen in the preceding chapters, the major advantages of applying the HMGLM to model the RSM is that the it affords the user the opportunity to (1) obtain better precision for the estimates of the person and item parameters; (2) posit a model with person covariates; and (3) posit a model with a group level. In addition, the HMGLM affords one the opportunity to posit a model with item covariates. This form of the HMGL-RSM may be especially important in DIF studies in which the user attempts to explain why DIF exists.

To model the HMGL-RSM with item covariates, one follows the previous definitions of the HMGL-RSM (Section 2-2), in which the category is nested within the item, which in turn is nested within the person. But now, one includes covariates at the item level.

6-1-1. The Level-1 Model with Item Covariates

The Level-1 model (the category level) is defined as

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \sum_{j=1}^{J} \beta_{jk}^{(i)} x_{jk} , \qquad (6.1)$$

where $\beta_{jk}^{(i)}$ is the mean category effect if person k selects category i of item j; and x_{jk} is a dummy variable with values 1 if person k answers item j, and 0 otherwise.

6-1-2. The Level-2 Model with Item Covariates

The Level-2 model (the item level) is defined as

$$\beta_{jk}^{(i)} = \gamma_{0jk} + \sum_{i=1}^{I} \gamma_{1\cdot k}^{(i)} w_{1\cdot k}^{(i)} + \sum_{i=1}^{I} \gamma_{2jk}^{(i)} w_{2jk}^{(i)} + \ldots + \sum_{i=1}^{I} \gamma_{Tjk}^{(i)} w_{Tjk}^{(i)} \qquad (6.2)$$

where, for person k, γ_{0jk} is the mean effect of item j across categories i; $\gamma_{1\cdot k}^{(i)}$ is the effect of an item on a particular category i; $w_{1\cdot k}^{(i)}$ is a dummy variable with values 1 if $i' = i$, and 0 otherwise; $\gamma_{tjk}^{(i)}$ is the effect of covariate t $(t = 2,\ldots,T-1)$ on a particular category i for item j; and $w_{tjk}^{(i)}$ is a the value of the t^{th} covariate of category i for item j. For identifiability, $\gamma_{1\cdot k}^{(0)} \equiv 0$ and $\gamma_{tjk}^{(0)} \equiv 0$.

6-1-3. The Level-3 Model with Item Covariates

The Level-3 model (the person level) is defined as

$$\gamma_{0jk} = \lambda_{0j0} + u_k, \qquad (6.3)$$

$$\gamma_{1\cdot k}^{(i)} = \lambda_{1\cdot 0}^{(i)}, \qquad (6.4)$$

$$\gamma_{tjk}^{(i)} = \lambda_{tj0}^{(i)}, \qquad (6.5)$$

where, for the j^{th} item that is answered by person k, λ_{0j0} is the mean effect of persons on item j; u_k is the random effect of person k on the mean effect of item j; $\lambda_{1\cdot 0}^{(i)}$ is the mean change in λ_{0j0} for a particular category of the items; and $\lambda_{tj0}^{(i)}$ is the mean change in λ_{0j0} for a particular covariate t of category i for item j.

Here, it is helpful to refer back to the honesty example, in which a particular feeling of an applicant in nested within an item, which in turn is nested within the person.

As before, a particular answered item not only depends upon the overall attractiveness of the item $\left(\lambda_{0j0}\right)$, but also how the attractiveness of the item influences a particular feeling $\left(\lambda_{1\cdot 0}^{(i)}\right)$. However, in addition to the honesty of the person, the response to the item also depends upon an item covariate $\left(\lambda_{tj0}^{(i)}\right)$, such as age. In other words, for example, the respondent may select one feeling over another more frequently because of his or her age.

6-1-4. The Combined Model with Item Covariates

The combined model of the HMGL-RSM with person covariates reduces to the following for a particular category i of the item j

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + \sum_{t=2}^{T} \lambda_{tj0}^{(i)} w_{ijk}^{(i)} + u_k , \tag{6.6}$$

where all terms are defined above.

Therefore, the parameters of the HMGL-RSM with item covariates are related to and extend the parameters of the traditional RSM in the following manner:

$$\delta_j = -\lambda_{0j0} , \tag{6.7}$$

$$\tau_i = -\lambda_{1\cdot 0}^{(i)} , \tag{6.8}$$

$$\theta_k = u_k , \tag{6.9}$$

and

$$\upsilon_{tij} = -\lambda_{tj0}^{(i)} , \tag{6.10}$$

where δ_j, τ_i, and θ_k are defined above; and υ_{tij} is the location of covariate t of category i for item j on the underlying continuum, which increases one unit as $w_{tjk}^{(i)}$ increases one unit.

Notice the HMGL-RSM with item covariates allows the item covariates to vary not only for each item, but for each threshold within each item as well. Currently, the aforementioned models in Chapter 1 do not allow for such flexibility in item covariate modeling.

6-2. Simulation Study for the HMGL-RSM with Item Covariates

The following section describes a simulation study for the HMGL-RSM with item covariates. The focus of this section is to examine the behaviors of the person and item parameters of the HMGL-RSM when being influenced by an item covariate.

6-2-1. Design

The design of the simulation is as follows. Observations were simulated using the HMGL-RSM. For the study, 500 simulees responded to 10 polytomous items, where each item consisted of 3 categories i (i = 0, 1, 2). The number of simulees, items, and categories were chosen to follow typical data from a questionnaire (e.g., Dodd, 1990; Smith & Johnson, 2000; Zhu, Updyke, & Lewandowski, 1997) or a large-scale assessment (e.g., Michigan Education Assessment Program, 2003; U.S. Department of Education, 1999). In addition, the number of simulees and items were chosen because, as shown in Section 3-2, these sample sizes allow for reasonable precision (at least when covariates were not modeled).

To produce the simulated responses, each simulee k was randomly assigned to be in one of four levels of the item covariate $\left(\lambda_{210}^{(i)}\right)$. The probability of being selected to a given level was chosen to be .01, .25, .66, and .08, respectively. Probabilities followed the actual frequencies of the levels of a covariate used in an operational administration of a confidential readiness assessment. Here, the covariate was age. For the simulation, the covariate influenced an arbitrarily chosen item, item 1.

Additionally, each simulee k was randomly assigned a θ_k, $\theta \sim N(0,1)$. δ_j and τ_i were randomly selected to represent parameter estimates obtained from a confidential readiness assessment (i.e., items 1-10 in Table 2).

Using θ_k, δ_j, and τ_i, three response probabilities for each simulee by item combination were produced, $P_{0jk}(\theta)$, $P_{1jk}(\theta)$, and $P_{2jk}(\theta)$. If $\sum_{0}^{i'} P_{i'jk}(\theta) < Y_{jk} \leq \sum_{0}^{i'+1} P_{i'jk}(\theta)$, then simulee k was assigned a response of $i'+1$ for item j; otherwise a response of 0 was assigned. Note that $i' = 0, 1$; and Y_{jk} was a single, random number for each $j \times k$ combination, $Y \sim U(0,1)$.

For the simulation, three different models were simulated for item 1. They were:

Model 1

$$\begin{aligned} \log\left(\frac{\pi_{11k}}{\pi_{01k}}\right) &= \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + \lambda_{210}^{(1)} + u_k \\ \log\left(\frac{\pi_{21k}}{\pi_{11k}}\right) &= \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + \lambda_{210}^{(2)} + u_k \end{aligned} ; \qquad (6.11)$$

Model 2

$$\log\left(\frac{\pi_{11k}}{\pi_{01k}}\right) = \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + \lambda_{210} + u_k$$
$$\log\left(\frac{\pi_{21k}}{\pi_{11k}}\right) = \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + \lambda_{210} + u_k \quad , \tag{6.12}$$

where the following constraint is made: $\lambda_{210}^{(1)} = \lambda_{210}^{(2)} = \lambda_{210}$;

and Model 3

$$\log\left(\frac{\pi_{11k}}{\pi_{01k}}\right) = \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + \lambda_{210}^{(1)} + u_k$$
$$\log\left(\frac{\pi_{21k}}{\pi_{11k}}\right) = \lambda_{0j0} + \lambda_{1\cdot 0}^{(i)} + 0 + u_k \quad ; \tag{6.13}$$

where all terms are defined above.

For the other items, the model was

$$\log\left(\frac{\pi_{ijk}}{\pi_{i-1,jk}}\right) = \lambda_{0j0} + \lambda_{1\square 0}^{(i)} + u_k \,, \tag{6.14}$$

where $j = 2, \ldots, 10$.

Note that $\lambda_{210}^{(1)}$ and $\lambda_{210}^{(2)}$ were arbitrarily set to .25 and .5, respectively, and that λ_{210} was arbitrarily set to .25. The reason for arbitrarily selecting the values for the coefficients was because the simulation studies above illustrated that the magnitude of the coefficient did not affect the RMSE, so little would be gained by manipulating the magnitude. Additionally, the values appeared to represent typical coefficient values of a covariate when using the HMGL-RSM (see Chapter 4).

Also note that the sample, item, and group sizes were not manipulated. The reason for this is that previous simulation studies from the previous sections have already

examined this issue. It seems that similar results would follow for the current model if a similar design to those above were used.

The simulation procedure simulated the 3 aforementioned conditions. Each administration was iterated 50 times producing 150 unique response data matrices. The number of iterations was chosen because Kamata (1998) showed this to be a reasonable number for obtaining stable estimates. S-Plus (2000) was used to generate all data. SAS (2001) was used to obtain parameter estimates and conduct significance tests.

6-2-2. Analysis

The purpose of the analysis is not only to examine the RMSE of estimating the parameters for the HMGL-RSM with item covariates, but the purpose is to examine the RMSE of estimating the parameters for the HMGL-RSM with item covariates when the incorrect model is specified. The reason being is that typically the user does not know the true model that explains the data. By examining the RMSE for the incorrect model, one can better understand how incorrect model specification affects precision.

Therefore, for the analysis, the three models described above were simulated. For a particular dataset, SAS (2001) was then used to estimate Models 1-3. Hence, one model would yield estimations for the correct model, while the two other models would yield estimations for the incorrect models.

Next, the parameter recovery of the HMGL-RSM with item covariates was conducted. Specifically, the RMSE for θ_k, $\lambda_{210}^{(1)}$, $\lambda_{210}^{(2)}$, λ_{210}, δ_j and τ_i was obtained over the iterations for each condition. The RMSE was

$$RMSE(\hat{\omega}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\omega}_n - \omega_n)^2}, \qquad (6.15)$$

where the maximum number of n iterations was $N = 50$; and ω is an arbitrary parameter representing either θ_k, $\lambda_{210}^{(1)}$, $\lambda_{210}^{(2)}$, λ_{210}, δ_j and τ_i. A descriptive analysis of the RMSE was conducted for each condition.

6-2-3. Results: Descriptive Statistics

Displayed in Tables 23, 24, and 25 are the mean and standard deviations of the parameter estimates for the HMGL-RSM when the correct model for item 1 was Model 1, 2, and 3, respectively. As can be seen, the standard deviations of the estimates are generally low and similar across conditions. However, the standard deviations are relatively higher for item 1 when covariates are added to the model. This suggests that PROC NLMIXED obtains relatively consistent estimates of the HMGL-RSM parameters, but the consistency decreases for an item when covariates are added.

As for the mean of the estimates, in general, the estimates obtained by PROC NLMIXED for the HMGL-RSM appear to differ only slightly from their parameter values. Below, in Section 6-2-4, the RMSE is examined.

Table 23. Mean and Standard Error of the Parameter Estimates for the HMGL-RSM for Model 1

	1		2		3	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.03	(0.57)	0.13	(0.49)	1.16	(0.22)
$\hat{\lambda}_{210}^{(1)}$	0.22	(0.20)	0.23	(0.17)	-0.18	(0.09)
$\hat{\lambda}_{210}^{(2)}$	0.51	(0.24)				
$\hat{\delta}_2$	0.01	(0.10)	0.01	(0.10)	0.01	(0.10)
$\hat{\delta}_3$	-0.92	(0.10)	-0.93	(0.10)	-0.92	(0.10)
$\hat{\delta}_4$	-1.59	(0.12)	-1.62	(0.12)	-1.60	(0.12)
$\hat{\delta}_5$	-0.82	(0.11)	-0.83	(0.11)	-0.82	(0.11)
$\hat{\delta}_6$	-0.74	(0.12)	-0.76	(0.12)	-0.75	(0.12)
$\hat{\delta}_7$	-0.84	(0.11)	-0.86	(0.11)	-0.85	(0.11)
$\hat{\delta}_8$	-0.03	(0.11)	-0.03	(0.11)	-0.03	(0.11)
$\hat{\delta}_9$	0.05	(0.12)	0.05	(0.13)	0.05	(0.13)
$\hat{\delta}_{10}$	-0.86	(0.13)	-0.87	(0.13)	-0.86	(0.13)
$\hat{\tau}_1$	-2.24	(0.05)	-2.28	(0.05)	-2.25	(0.05)
$\hat{\tau}_2$	2.24	(0.05)	2.28	(0.05)	2.25	(0.05)
$\mu_{\hat{u}}$	0.01	(0.00)	0.01	(0.00)	0.01	(0.00)
$\Sigma_{\hat{u}}$	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)

Note. {1, 2, 3} = estimated Models 1, 2, and 3. $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. $\{\hat{\lambda}_{210}^{(1)}, \hat{\lambda}_{210}^{(2)}\}$ = item covariates. $\hat{\lambda}_{210}^{(1)}$ for Model 2 = $\hat{\lambda}_{210}$. $\mu_{\hat{u}}$ = Mean person location. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations. M = Mean. SE = Standard error.

Table 24. Mean and Standard Error of the Parameter Estimates for the HMGL-RSM for Model 2

	1		2		3	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.02	(0.52)	-0.01	(0.51)	0.56	(0.22)
$\hat{\lambda}_{210}^{(1)}$	0.22	(0.19)	0.22	(0.19)	0.02	(0.10)
$\hat{\lambda}_{210}^{(2)}$	0.23	(0.21)				
$\hat{\delta}_2$	0.01	(0.10)	0.01	(0.10)	0.01	(0.10)
$\hat{\delta}_3$	-0.92	(0.10)	-0.92	(0.10)	-0.92	(0.10)
$\hat{\delta}_4$	-1.59	(0.12)	-1.59	(0.12)	-1.60	(0.12)
$\hat{\delta}_5$	-0.82	(0.11)	-0.82	(0.11)	-0.82	(0.11)
$\hat{\delta}_6$	-0.74	(0.12)	-0.75	(0.12)	-0.75	(0.12)
$\hat{\delta}_7$	-0.84	(0.11)	-0.84	(0.11)	-0.84	(0.11)
$\hat{\delta}_8$	-0.03	(0.11)	-0.03	(0.11)	-0.03	(0.11)
$\hat{\delta}_9$	0.05	(0.12)	0.05	(0.12)	0.05	(0.12)
$\hat{\delta}_{10}$	-0.86	(0.13)	-0.86	(0.13)	-0.86	(0.13)
$\hat{\tau}_1$	-2.24	(0.05)	-2.25	(0.05)	-2.25	(0.05)
$\hat{\tau}_2$	2.24	(0.05)	2.25	(0.05)	2.25	(0.05)
$\mu_{\hat{u}}$	0.01	(0.00)	0.01	(0.00)	0.01	(0.00)
$\Sigma_{\hat{u}}$	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)

Note. {1, 2, 3} = estimated Models 1, 2, and 3. $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. $\{\hat{\lambda}_{210}^{(1)}, \hat{\lambda}_{210}^{(2)}\}$ = item covariates. $\hat{\lambda}_{210}^{(1)}$ for Model 2 = $\hat{\lambda}_{210}$. $\mu_{\hat{u}}$ = Mean person location. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations. M = Mean. SE = Standard error.

Table 25. Mean and Standard Error of the Parameter Estimates for the HMGL-RSM For Model 3

	1		2		3	
	M	SE	M	SE	M	SE
$\hat{\delta}_1$	-0.02	(0.48)	0.01	(0.10)	0.01	(0.10)
$\hat{\lambda}_{210}^{(1)}$	0.22	(0.17)	0.13	(0.20)	0.25	(0.08)
$\hat{\lambda}_{210}^{(2)}$	-0.03	(0.17)				
$\hat{\delta}_2$	0.01	(0.10)	-0.89	(0.10)	-0.92	(0.10)
$\hat{\delta}_3$	-0.92	(0.10)	-1.56	(0.12)	-1.59	(0.12)
$\hat{\delta}_4$	-1.59	(0.12)	-0.80	(0.10)	-0.82	(0.11)
$\hat{\delta}_5$	-0.82	(0.11)	-0.73	(0.12)	-0.74	(0.12)
$\hat{\delta}_6$	-0.74	(0.12)	-0.82	(0.10)	-0.84	(0.11)
$\hat{\delta}_7$	-0.84	(0.11)	-0.03	(0.11)	-0.03	(0.11)
$\hat{\delta}_8$	-0.03	(0.11)	0.05	(0.12)	0.05	(0.12)
$\hat{\delta}_9$	0.05	(0.12)	-0.83	(0.13)	-0.86	(0.13)
$\hat{\delta}_{10}$	-0.86	(0.13)	-2.20	(0.05)	-2.24	(0.05)
$\hat{\tau}_1$	-2.24	(0.05)	2.20	(0.05)	2.24	(0.05)
$\hat{\tau}_2$	2.24	(0.05)	-0.08	(0.56)	-0.10	(0.17)
$\mu_{\hat{u}}$	0.01	(0.00)	0.01	(0.00)	0.01	(0.00)
$\Sigma_{\hat{u}}$	1.00	(0.05)	1.00	(0.05)	1.00	(0.05)

Note. {1, 2, 3} = estimated Models 1, 2, and 3. $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. $\{\hat{\lambda}_{210}^{(1)}, \hat{\lambda}_{210}^{(2)}\}$ = item covariates. $\hat{\lambda}_{210}^{(1)}$ for Model 2 = $\hat{\lambda}_{210}$. $\mu_{\hat{u}}$ = Mean person location. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations. M = Mean. SE = Standard error.

6-2-4. Results: RMSE

The results of the RMSE for $\mu_{\hat{\theta}}$, $\Sigma_{\hat{\theta}}$, $\lambda_{210}^{(1)}$, $\lambda_{210}^{(2)}$, δ_j, and τ_i of the HMGL-RSM with item covariates are provided in Table 26. Trends indicated that the RMSE generally remained the same, which were low, for $\mu_{\hat{\theta}}$, $\Sigma_{\hat{\theta}}$, $\delta_2 - \delta_{10}$, and τ_i, even if an incorrect model was estimated.

However, the RMSE generally increased for δ_1, $\lambda_{210}^{(1)}$, and $\lambda_{210}^{(2)}$ when an incorrect model was estimated. This especially occurs if the correct model was Model 1 or 2 and the incorrect estimated model was Model 3. Nevertheless, except if the correct model was Model 1 and the incorrect estimated model was Model 3, the RMSE tended to remain within reasonable levels below or around .55. Thus, the analysis provides some evidence that if the model was correctly specified, the parameters were estimated extremely well unless it were influenced by an item covariate. In this case, only when the model did not specify an item covariate when there should have been one does the precision become unreasonable. Otherwise, the precision is somewhat low, yet reasonable.

Table 26. RMSE for the HMGL-RSM with Item Covariates

T	1			2			3		
E	1	2	3	1	2	3	1	2	3
δ_1	0.56	0.53	1.26	0.52	0.52	0.69	0.48	0.55	0.17
$\lambda_{210}^{(1)}$	0.20	0.17	0.44	0.19	0.19	0.25	0.18	0.23	0.08
$\lambda_{210}^{(2)}$	0.24			0.34			0.55		
δ_2	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
δ_3	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
δ_4	0.12	0.13	0.12	0.12	0.12	0.12	0.12	0.12	0.12
δ_5	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.11
δ_6	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
δ_7	0.11	0.12	0.11	0.11	0.11	0.11	0.11	0.10	0.11
δ_8	0.11	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.11
δ_9	0.12	0.13	0.13	0.12	0.12	0.12	0.12	0.12	0.12
δ_{10}	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
τ_1	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.06	0.05
τ_2	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.06	0.05
$\mu_{\hat{u}}$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
$\Sigma_{\hat{u}}$	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Note. T = True model. E = Estimated model. {1, 2, 3} = Models 1, 2, and 3. $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_{10}\}$ = location for items 1 – 10. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. $\{\hat{\lambda}_{210}^{(1)}, \hat{\lambda}_{210}^{(2)}\}$ = item covariates. $\mu_{\hat{u}}$ = Mean person location. $\Sigma_{\hat{\theta}}$ = Standard deviation of the person locations. M = Mean. SE = Standard error.

6-3. Example Analysis of the HMGL-RSM with Item Covariates

The purpose of this section is to provide an example analysis that illustrates the basic concepts of the HMGL-RSM with item covariates. In particular, one will illustrate how to use the model to assist in explaining DIF.

6-3-1. Design

The design of the analysis is as follows. Five hundred respondents were randomly selected from a larger sample of students who responded to a confidential readiness assessment. Note these respondents were the same respondents used in Section 5-3. However, those respondents who did not provide their Age were not used here. Thus, the final sample consisted of 473 respondents. Their demographics are provided below in Table 27. As one can see, there appears to be an equal distribution of Males and Females in each of the demographic categories.

Table 27. Demographic Information

		Males	Females	Total
SES				
	Hi	106	103	209
	Mid	123	100	223
	Lo	22	17	39
Age				
	5	1	2	3
	6	66	64	130
	7	171	161	332
	8	22	6	28
Ethnicity				
	Asian	2	5	7
	Af.-Am.	132	104	236
	Hisp.	23	18	41
	Cauc.	98	105	203

Note. Af.-Am. = African-American. Hisp. = Hispanic. Cauc. = Caucasian. 9 Males and 13 Females did not provide their parent's SES. 5 Males and 1 Female did not provide Ethnicity.

As in Section 5-3, this illustration only utilized the first 10 items of a confidential readiness assessment (which again measured the person's personal and social development). The item covariate that was used was Age. Age was selected as the covariate because there was some reason to believe that the older respondents may have interpreted the categories differently than the younger respondents. Lastly, recall that items 3-5 and 9 contained DIF. Hence, Age was used to explain the DIF that appeared for items 3-5 and 9 for Males and Females. (Note although the HMGL-RSM identified additional items as containing DIF, they were not modeled as being influenced by Gender or Age. The reason for doing so was because the effects of the non-DIF items on the item covariates were not of interest here.)

6-3-2. Analysis

To analyze the responses of the students, PROC NLMIXED of SAS (2001) was used to estimate the person and item parameters for the HMGL-RSM with Gender and Age as the item covariate for items 3-5 and 9, and no item covariates for the remaining items. Hence, the final model is the HMGL-RSM with a group level and item covariates.

6-3-3. Results

The results of the analysis are presented below in Table 28. To interpret the HMGL-RSM with item covariates, recall from above that if the item exhibits DIF, then $\xi_{0j01} \neq 0$ and

$$\delta_{j0} = -\xi_{0j00}$$
$$\delta_{j1} = -\left(\xi_{0j00} + \xi_{0j01}\right).$$

For example, for item 3, when item covariates are added to explain DIF, the overall attractiveness of the item for Males is $\hat{\delta}_{3,0} = -\xi_{0300} = 4.34$, while the overall attractiveness of the item for Females is

$$\hat{\delta}_{3,1} = -\left(\xi_{0300} + \hat{\xi}_{0301}\right) = -\left(-4.34 + 0.82\right) = 3.52.$$

Table 28. Parameter Estimates for the HMGL-RSM With Age as an Item Covariate

		Age Included				Age Not Included			
Item	Par.	Est.	SE	t	p	Est.	SE	t	p
1	$\hat{\xi}_{0100}$	-1.34	0.16	-8.21	0.00	-1.32	0.16	-8.12	0.00
2	$\hat{\xi}_{0200}$	-1.60	0.16	-9.68	0.00	-1.57	0.16	-9.58	0.00
3	$\hat{\xi}_{0300}$	-4.34	1.42	-3.05	0.00	-0.89	0.19	-4.64	0.00
	$\hat{\xi}_{0301}$	0.82	0.23	3.66	0.00	0.81	0.22	3.64	0.00
	$\hat{\xi}^{(1)}_{0302}$	0.49	0.21	2.34	0.02				
	$\hat{\xi}^{(2)}_{0302}$	0.54	0.21	2.60	0.01				
4	$\hat{\xi}_{0400}$	-1.53	1.42	-1.08	0.28	-0.37	0.19	-1.97	0.05
	$\hat{\xi}_{0401}$	0.84	0.23	3.69	0.00	0.82	0.22	3.69	0.00
	$\hat{\xi}^{(1)}_{0402}$	0.16	0.21	0.79	0.43				
	$\hat{\xi}^{(2)}_{0402}$	0.18	0.21	0.85	0.40				
5	$\hat{\xi}_{0500}$	-2.62	1.40	-1.87	0.06	-0.65	0.19	-3.41	0.00
	$\hat{\xi}_{0501}$	0.46	0.22	2.09	0.04	0.48	0.22	2.17	0.03
	$\hat{\xi}^{(1)}_{0502}$	0.26	0.20	1.27	0.21				
	$\hat{\xi}^{(2)}_{0502}$	0.33	0.20	1.61	0.11				
6	$\hat{\xi}_{0600}$	-0.58	0.16	-3.62	0.00	-0.57	0.16	-3.59	0.00
7	$\hat{\xi}_{0700}$	-0.71	0.16	-4.42	0.00	-0.70	0.16	-4.37	0.00
8	$\hat{\xi}_{0800}$	-1.44	0.16	-8.79	0.00	-1.42	0.16	-8.70	0.00

Table 28 (cont'd)

9	$\hat{\xi}_{0900}$	-1.37	1.45	-0.94	0.35	-1.95	0.20	-9.79	0.00
	$\hat{\xi}_{0901}$	0.88	0.23	3.86	0.00	0.95	0.23	4.17	0.00
	$\hat{\xi}^{(1)}_{0902}$	-0.11	0.21	-0.52	0.60				
	$\hat{\xi}^{(2)}_{0902}$	-0.04	0.21	-0.18	0.86				
10	$\hat{\xi}_{0,10,00}$	-0.67	0.16	-4.16	0.00	-0.66	0.16	-4.11	0.00
	$\hat{\tau}_1$	-2.18	.	.	.	2.10	.	.	.
	$\hat{\tau}_2$	2.18	0.05	-46.18	0.00	-2.10	0.03	-60.25	0.00

Note. Par. = parameter. Est. = estimate. SE = standard error. t = t-statistic. p = p-value. $\{\hat{\xi}_{0j00}, \hat{\xi}_{0j01}\}$ = overall attractiveness of item j for Males and Females, respectively. $\{\hat{\tau}_1, \hat{\tau}_2\}$ = thresholds 1 and 2. p = 0.00 implies p < .01.

To explain the difference in the attractiveness between Males and Females, the model suggests that Age may influence the genders. That is, older Males and Females may interpret the item categories differently than younger Males and Females. Additionally, this influence is not constant across category thresholds. For example, the location of Age on the underlying continuum as the category increases from 0 to 1 is $\hat{\upsilon}_{213} = -\hat{\xi}^{(1)}_{0302} = -.49$, while the location of Age as the category increases from 1 to 2 is $\hat{\upsilon}_{223} = -\hat{\xi}^{(2)}_{0302} = -.54$. Thus, if a Male or Female is age 5 then, then the location of Age on the underlying continuum as the category increases from 0 to 1 is $\hat{\upsilon}_{213} \times w^{(1)}_{23k} = -.49 \times 5 = -2.45$. If the age is 6, then the location is $\hat{\upsilon}_{213} \times w^{(1)}_{23k} = -.49 \times 6 = -2.94$. And so on, for Ages 7 and 8, where similar

interpretations hold for the location of Age as the category increases from 1 to 2. This suggests that as Age increases, the location of Age decreases for Males and Females.

To answer the question of whether Age adequately explains the DIF exhibited in the items, one examines the model fit of the current model compared to the model without Age as a covariate using the AIC and BIC. When Age is included in the model, the AIC and BIC are 6955.7 and 7060.7, respectively. When Age is not included in the model, the AIC and BIC are 6955.6 and 7027.0, respectively. Furthermore, when inspecting the information weights, the AIC and BIC weights for the HMGL-RSM without Age are .51 and 1.00, while the AIC and BIC weights for the HMGL-RSM with Age are .49 and .00. Since the AIC and BIC are lower for the HMGL-RSM without Age, and since higher weights indicate the model is more likely, the evidence suggests that the HMGL-RSM without Age is the better fitting model. Thus, although the HGML-RSM aids in the explanation of DIF, it was found that Age does not explain the existence of DIF for this particular example.

Chapter 7. Conclusions and Future Directions

7-1. Conclusions

As shown in the preceding chapters, the parameters of the HMGL-RSM were recovered fairly well. In addition, simulations and example analyses illustrated the three primary advantages of utilizing the HMGLM to model the RSM and PCM. Specifically, the HMGL-RSM and -PCM were able to extend existing models to include person covariates, a group level, and item covariates.

In addition, the dissertation illustrated several advantages of utilizing the HMGL-RSM and -PCM for analyzing educational testing data. Specifically, in Chapter 1, it was discussed that traditional methods, such as the RSM and PCM, do not account for the variation between persons and variation of responses within a person. By applying the HMGL-RSM and -PCM, this is accounted for. Additionally, in Chapter 1, one discussed how the HMGL-RSM and -PCM allow for a singular method that utilizes a hierarchical framework (HLM) that extends polytomous IRT models to include person covariates and predictors of item behaviors, and accounts for the correlation between categories of a polytomous item. No other method applies this specific framework to do so.

In Chapter 2, the HMGLM framework is used to define the HMGL-RSM and -PCM. It was noted, and should be re-stated, that although Tuerlinckx and Wang (2004) present similar models, the reader should be aware that the models presented here are not the same models as those presented by the aforementioned authors. The models defined here use the HMGLM framework; this framework defines a separate model for each hierarchical level. As argued, this allows for a more 'natural' way of not only modeling educational testing data, but also understanding educational testing data.

In Chapter 3, the HMGLM framework is used to illustrate how the HMGL-RSM performs in comparison to traditional IRT methods such as the RSM. As shown and discussed, the primary advantage is that the HMGL-RSM estimates have smaller standard errors than the RSM estimates. This, of course, becomes important as the user places higher stakes on the interpretation of those estimates. For example, if the user interprets the estimate of the person parameter as being the person proficiency, and if the user utilizes this estimate to make the high-stakes decision of whether or not the person passes high-school, then the less error associated with this estimate, the more confident the user will be in making this high-stake decision.

In Chapter 4, the HMGLM framework is used to illustrate how the HMGL-RSM can be extended to include person covariates. As shown, by applying the HMGL-RSM with person covariates the user can control for the influence of a covariate at the person level. This form of the HMGL-RSM may be especially important in accountability investigations in which the user is interested in the location of a student, after controlling for the effects of a covariate (e.g., Stone and Lane (2003)). For example, assume in the example analysis in Section 4-3 that test-takers obtain a monetary reward for performing well. As shown, SES was negatively related to performance. Thus, we can see that if the monetary cut-off were .5 logits, then the lower SES group would receive the monetary reward—only if SES was controlled for. Compare this to not controlling for SES: the lower SES group would not receive any monetary reward.

Additionally, as was implied in Chapter 4, the HMGL-RSM with person covariates has its advantages over traditional methods using covariates such as the analysis of covariance (ANCOVA). For instance, to apply ANCOVA as a measure for

controlling the effects of the covariates on student performance, then the user must first estimate the person and item locations using IRT. Next, the user applies ANCOVA procedures. To do so, one must estimate a model where the dependent variable is the total test score; and the independent variables are the IRT person estimate and covariate. By estimating this model, the user may be able to examine how the covariate influences the person's performance on the test. However, this process has its limitations. One limitation is that the estimates of the covariate and the estimates of the person performance are not necessarily placed on the same scale. This issue becomes a problem as the user attempts to interpret the estimates: Does 1 unit in the covariate scale mean the same thing as 1 unit in the person performance scale? Another limitation is that the process is somewhat time inefficient since two separate steps are used to obtain the aforementioned estimates, the IRT step and the ANCOVA step.

The advantage of applying the HMGLM to extend IRT models is that the procedure for controlling the effects of the covariates on student performance is simplified to only one step (i.e., estimating one model as opposed to two, which as mentioned earlier may be a more natural way of conceptualizing the data), and the estimates are placed on the same scale (Lord, 1980).

In Chapter 5, the HMGLM framework is used to illustrate how the HMGL-RSM can be extended to include a group level. As shown, this model was a somewhat powerful test for detecting DIF. Additionally, when compared to another popular DIF procedure, the MH test, the HMGL-RSM was not only more powerful, but it afforded a few advantages the MH test did not. For instance, although a purification procedure was not used here with the MH test because the purification procedure would not greatly

influence the DIF results for the simulated testing conditions (e.g., Wang & Su, 2004), there may be other operational conditions that a purification procedure may be necessary. By utilizing the HMGL-RSM, a purification procedure is not necessary and this issue is avoided. That is, by modeling the testing environment with the HMGL-RSM, the model controls for the effects of DIF and non-DIF items and simultaneously investigates for DIF. Hence, no purification is necessary since the effect of the other items is controlled for.

In Chapter 6, the HMGLM framework is used to illustrate how the HMGL-RSM can be extended to include item covariates. As shown, this extension may provide a way to explain why DIF exists. As briefly discussed, after a DIF examination occurs in an operational setting, the user must now attempt to explain why DIF occurs, and a decision regarding the item must be made. That is, the user must decide: even though DIF exists, does the item display any characteristics that would create a bias for a particular group? If so, should the item be modified or removed from the test? By applying the HMGL-RSM with item covariates, the guesswork is minimized for the first part of the decision. That is, rather than providing a subject judgment for whether or not the item displays any biasing characteristics, the HMGL-RSM allows the user to explicitly create a model that examines the user's hypothesis. For example, rather than the user suggesting a math item may be exhibiting DIF because it is a trigonometry item and the other items are not trigonometry items, the user may explicitly define a model that includes whether or not an item is a trigonometry item, and then he may examine this model for its ability to explain the occurrence of DIF.

Lastly, it is re-iterated: the HMGLM allows the user to accomplish all of the aforementioned advantages—simultaneously. Again, there is currently no other procedure that applies this particular hierarchical framework to do so. Below one discusses additional contributions of this framework and these models.

7-1-1. Contributions

Beyond extending the RSM and PCM, there are four main contributions that result by applying the HMGLM to unify HLM and polytomous IRT models. As stated before, they include (1) models using HMGLM may currently be estimated using existing software (e.g., SAS, 2001; STATA, 2000); (2) IRT and HLM are unified using a common notation; (3) score functions and information matrices (which may be used for parameter estimation) are well-known under the HMGLM (e.g., see Fahrmeir & Tutz, 2001); and (4) a broad class of IRT models within the HLM framework may be estimated using a common method (e.g., maximum likelihood).

7-1-1.1. Special Estimation Software is Not Necessary

By applying the HMGLM, estimation of IRT models does not require special software (e.g., HLM for Windows, 2001). To estimate the HMGLM, all one needs is any of the mass software that estimates generalized mixed models, such as SAS or STATA. Consequently, this suggests that users do not have to learn additional software to estimate these models. Although this may seem like a trivial point, it becomes a strong point once one considers the amount of time and money saved by not expending one's energies and finances necessary in purchasing and learning new software.

7-1-1.2. Common Notation

Another contribution of applying the HMGLM is that a common notation system may be used to describe the models that are unified from two different areas of research. Although this may seem trivial, it actually is not once one considers that each area of research, HLM and IRT, has its own notation. Furthermore, each researcher may bring his own 'style' to the notation system. Additionally, if one considers that each separate notation system may be considered a separate language, then it becomes cumbersome and confusing when researchers attempt to discuss similar concepts and theories in different languages, i.e., notations. For example, notice in the discussion above, that the ability parameter is represented by θ in IRT, but the ability parameter is represented by u in HLM. By applying the HMGLM, HLM and IRT may be unified in such a way that avoids this issue. And, at the same time, the interpretation of the parameters remains consistent. Furthermore, since the HMGLM is an extension of univariate GLM, which already has a strong history and accepted notation, users may simply incorporate IRT and HLM within a knowledge structure that already exists for GLM without confusing oneself any further.

7-1-1.3. Well-Known Score Functions and Information Matrices

By applying the HMGLM to IRT, the score functions and information matrices are well known for the hierarchical IRT models (see Fahrmeir and Tutz, 2001, Chapter 3). Since these are well known, it is not necessary for the user to derive these such that they can be used during maximum likelihood estimation of the parameters. Compare this

to the Bayesian approach. In this approach, for each new model that is developed, the user may have to derive a new prior and posterior distribution so that the parameters can be estimated. Although this may be a simple task for some, this may be an extremely difficult feat for others. By applying the HMGLM to IRT, this can be avoided, and most researchers who have a general understanding of GLM, HLM, and IRT can enjoy its application.

7-1-1.4. Common Estimation Method

As the reader can see, there are numerous possibilities for postulating hierarchical IRT models when the HMGLM is applied. Fortunately, since the HMGLM is simply an extension of GLM, which has well-studied and well-understood properties (e.g., score functions and information matrices), the HMGLM also has well-studied and well-understood properties. The advantage of this is that the numerous hierarchical IRT models that can be developed under the HMGLM may be estimated using a common estimation method. For instance, here, recall that estimates of the parameters are obtained by maximizing an approximation to the likelihood integrated over the random effects, where the integral approximations are obtained via adaptive Gaussian quadrature and the optimization technique is carried out using a dual quasi-Newton algorithm (SAS, 2001) or a modified Newton-Rapheson algorithm (Rabe-Hesketh, Pickles, & Skrondal, 2001). Again, compare this to the Bayesian approach. For this approach, if a new model is developed, characteristics such as the conditional probability distributions for the variances may differ for each new model. Consequently, if the characteristics change for each new model, then it may be necessary to alter the algorithm of the estimation method

for each new model. Obviously, this may prove to be laborious, and consequently the application of the new model may be avoided. Again, this is not the case for the HMGLM.

7-2. Limitations

Below, one describes five limitations that were encountered during this dissertation, some of which was the result of using popular estimation software such as PROC NLMIXED in SAS. They include: (1) the item discrimination parameter is not modeled; (2) data preparation is cumbersome; (3) potentially long estimation times; (4) unbalanced data was not considered; and (5) a non-normal distribution of random effects was not investigated.

7-2-1. Item Discrimination Parameter is Not Modeled

The first limitation is that the item discrimination parameters were not modeled. That is, Muraki (1992) presented an extension of the PCM in which each item has its own discrimination (i.e., slope). As suggested by this model, this may be an important parameter to consider if one cannot assume the discrimination of the test items equals one. Notice that this assumption was made in order to simulate responses for the HMGL-PCM and -RSM. Fortunately, this does not affect the generality of the HMGL-RSM or -PCM. That is, although it may be necessary to model the discrimination parameter for some achievement tests or questionnaires, this may not hold for all tests or questionnaires. For example, the Michigan Education Assessment Program does not apply a model with a discrimination parameter for estimating the parameters of the state's

achievement test (Michigan Education Assessment Program, 2003). Additionally, Dodd (1990), Smith and Johnson (2000), and Zhu, Updyke, and Lewandowski (1997) also do not model discrimination parameters for estimating the parameters of a questionnaire.

7-2-2. Data Preparation is Cumbersome

A second limitation is that data preparation is fairly cumbersome. That is, before using estimating the HMGL-RSM and -PCM, the user must structure the raw data such that the categorical response is a multivariate vector (rather than the category selection itself, which is typically the case when estimating non-hierarchical polytomous models, e.g., see the software WINSTEPS (1999)). Additionally, the user must create $J-1$ dummy variables that identify the item under investigation (see Appendix C). As can be guessed, this process becomes rather tiresome as the number of items and categories increases. Nevertheless, the author feels that the time invested in pursuing the application of the HMGLM in IRT is far outweighed by the benefits gained (see Section 7-1-1).

7-2-3. Possibly Long Estimation Times

Another limitation is that, if adaptive Gaussian quadrature is used (as is done in this dissertation), then the estimation of the HMGL-PCM and -RSM may require long estimation times. For example when using a PC with a 3.2 GHz, Intel Pentium 4 processor, parameter estimation of the HMGL-RSM took approximately 12 hours when the number of persons and items was 1000 and 25, respectively. This occurs because adaptive Gaussian quadrature requires finding the mode of the function being integrated. This means that as the number of random effects increases—in the case for IRT

modeling, as the number of persons increases—adaptive Gaussian quadrature finds the mode for each unique random effect for each iteration of the estimation algorithm.

Thus, alternative methods to the HMGL-PCM and -RSM may be more worthwhile if long estimation times are to be avoided. For example, if the user wants an estimate of the effect of a covariate for a group of students, an ANCOVA can be applied. Or, if the user wants to test for DIF, then the MH test can be applied. Of course, these alternatives also have their disadvantages, which were discussed above. Hence, the user must choose the preferred method based on which advantages and disadvantages are most important to him/her.

Nevertheless, the long estimation times does not appear to be a major hurdle in applying the HMGLM to IRT, at least in the near future, considering that computers are becoming increasingly faster, which may decrease estimation times. Additionally, as mentioned in Section 2-5, other estimation procedures, which are possibly faster than adaptive Gaussian quadrature, may be employed.

7-2-4. Unbalanced Data

A fourth limitation encountered in this dissertation is that the simulation study did not investigate the accuracy of the parameter estimates of unbalanced data (i.e., all persons do not respond to all items). Of course, in real data, unbalanced data is more likely the rule rather than the exception. Nevertheless, this dissertation provides insight on how well the parameters for the HMGL-PCM and -RSM are estimated under ideal conditions. Consequently, this ideal scenario can now be used as a benchmark for comparison with future studies that investigate the effects of unbalanced data.

7-2-5. Non-Normal Distribution for Random Effects Not Investigated

A fifth limitation is that non-normal random effects were not investigated. Although it is possible that the random effects may not be normal in actual data, in educational research it is commonly assumed that the distribution of the effects is normal (e.g., Cheong & Raudenbush, 2000; Kamata, 1998, 2001; Lord, 1980; Miyazaki, 2000). Here, customary assumptions were used, and it is expected that this should not affect the generality of the model itself. However, if the user is interested in non-normal effects, then one may posit a non-normal distribution and estimate the model using approaches other than those discussed here. For example, Hartzel et al. (2001) and Aitkin (1999, as cited by Hartzel et al., 2001) present a semi-parametric estimation method that does not rely on a multivariate normal specification of the random effects. Additionally, GLLAMM for STATA allows one to apply binomial, gamma, or Poisson (Rabe-Hesketh, Pickles, & Skrondal, 2001). Fahrmeir and Tutz (2001, Chapter 7) present estimation methods based on posterior modes or Bayesian techniques, which also do not require the distribution of the random effect to be normal. Lastly, Breslow and Clayton (1993, as cited by Gueorguieva, 2001) and Wolfinger and O'Connell (1994, as cited by Gueorguieva, 2001) present a penalized quasi-likelihood method, which also does not require the distribution of the random effect to be normal

7-3. Future directions

Future researchers may direct their efforts toward addressing the limitations described above. For instance, researchers can develop software specifically designed for

estimating the HMGLM. If accomplished, limitations of data preparation and estimation times would be avoided. However, that is not to say utilizing PROC NLMIXED in SAS is not worthwhile. Typical everyday users who are not adept at developing computer estimation software should fine SAS useful as it provides an easily understandable and readily available method to estimate the models discussed here.

Additionally, researchers may attempt to apply the HMGLM to a polytomous IRT model with a discrimination parameter (e.g., Muraki, 1992). This may be possible if one extends the work of Miyazaki (2000) to polytomous models. Additionally, researchers may examine the parameter recovery rate for more 'real-like' simulated data in which the data is unbalanced. Lastly, researchers may examine the estimates if non-normal random effects were utilized.

Other research may direct their efforts toward extending the contributions described above. For instance, researchers may wish to model a hierarchical FACETS model (Linacre, 1994). One application of this model is found in the literature regarding rater effects (e.g., Wolfe, Moulder, & Myford, 2001). It would be interesting to see how accurately rater effects would be measured by the FACETS model by applying the HMGLM.

Finally, future researchers may direct their efforts in comparing the HMGLM to the Bayesian modeling of random-effects approach (Section 1-2-3), the rater effects approach (Section 1-2-4), and the hierarchical univariate general linear model approach (Section 1-2-5). Although each of these approaches attempts to obtain similar information, they do so in different manners, as discussed above. It would be interesting

to examine the equivalence in the parameter estimates obtained from each approach. It is possible that one approach provides better estimates than the other approaches.

APPENDICES

APPENDIX A.

Example SAS Code for Estimating the HMGL-RSM for a Polytomous Test with 10 Items

```
*~~~ INPUT DATA ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

data RSM;
   infile "C:\WINDOWS\Start Menu\temp\data.dat" ;
   input y0 y1 y2 y3 person_id item_id x1-x10 ;
run;

proc sort ;
   by person_id;
run;

*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

*~~~ RUN NLMIXED FOR INITIAL ESTIMATES ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

proc nlmixed data=RSM ;

  *PRE-INITIAL ESTIMATES;
  parms beta1-beta10 gamma1-gamma3 = 0;

  *CODE LINEAR PREDICTORS;
  gamma3 = -1*(gamma1+gamma2);

  eta1 = x1* beta1 + x2* beta2 + x3* beta3 + x4* beta4 + x5* beta5 +
             x6* beta6 + x7* beta7 + x8* beta8 + x9* beta9 + x10* beta10 + gamma1;
  eta2 = x1* beta1 + x2* beta2 + x3* beta3 + x4* beta4 + x5* beta5 +
             x6* beta6 + x7* beta7 + x8* beta8 + x9* beta9 + x10* beta10 + gamma2;

  *RATING SCALE MODEL;
  pi0 = 1 / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );
  pi1 = exp(eta1) / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );
  pi2 = exp(eta1+eta2) / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );

  *DEFINE LIKELIHOOD;
  z = (pi0**y0)*(pi1**y1)*(pi2**y2)*(pi3**y3);
  if (z > 1e-8) then ll = log(z);
  else ll=-1e100;
  model y0 ~ general(ll);
```

```
  *SPECIFY RANDOM EFFECT DISTRIBUTION;
  *none;

  *OBTAIN INITIAL ESTIMATES;
  ods output ParameterEstimates = parest ;

run;

*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

*~~~ RUN NLMIXED FOR FINAL ESTIMATES ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

proc nlmixed data=RSM ;

  *READ IN INITIAL ESTIMATES;
  parms / data = parest;

  *CODE LINEAR PREDICTORS;
  theta = u1 ;

  gamma3 = -1*(gamma1+gamma2);

  eta1 = x1* beta1 + x2* beta2 + x3* beta3 + x4* beta4 + x5* beta5 +
            x6* beta6 + x7* beta7 + x8* beta8 + x9* beta9 + x10* beta10 + gamma1 +
theta;
  eta2 = x1* beta1 + x2* beta2 + x3* beta3 + x4* beta4 + x5* beta5 +
            x6* beta6 + x7* beta7 + x8* beta8 + x9* beta9 + x10* beta10 + gamma2 +
theta;

  *RATING SCALE MODEL;
  pi0 = 1 / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );
  pi1 = exp(eta1) / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );
  pi2 = exp(eta1+eta2) / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );

  *DEFINE LIKELIHOOD;
  z = (pi0**y0)*(pi1**y1)*(pi2**y2)*(pi3**y3);
  if (z > 1e-8) then ll = log(z);
  else ll=-1e100;
  model y0 ~ general(ll);

  *SPECIFY RANDOM EFFECT DISTRIBUTION AND OBTAIN EMPIRICAL
BAYES ESTIMATES;
  random  u1 ~ normal(0, s1*s1) subject = person_id OUT=bayesest;

run;
```

```
*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

***************************************************************************;

NOTE. THIS PROGRAM WAS OBTAINED AND MODIFIED FROM
    HARTZEL, AGRESTI, AND CAFFO (2001).

ALSO NOTE THEY STATE THE FOLLOWING:

"With Gauss-Hermite quadrature, computer
underflow can be a problem mainly when there are many within-cluster
observations. For most data sets in our experience, however, it is
the number of clusters that is large and not the number of
observations within a cluster. In using NLMIXED, we addressed this
problem by assigning the likelihood to a very small number within the
limits of computer precision. Specifically we entered

if (z > 1e-8) then ll = log(z); else ll=-1e100

for this purpose."

***************************************************************************;
```

APPENDIX B.

Example SAS Code for Estimating the HMGL-PCM for a Polytomous Test with 10 Items

```
*~~~ INPUT DATA ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

data PCM;
   infile "C:\WINDOWS\Start Menu\temp\data.dat" ;
   input y0 y1 y2 y3 person_id item_id x1-x10 ;
run;

proc sort ;
   by person_id;
run;

*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

*~~~ RUN NLMIXED FOR INITIAL ESTIMATES ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

proc nlmixed data=PCM ;

  *PRE-INITIAL ESTIMATES;
  parms beta1-beta10
     gamma11-gamma12
     gamma21-gamma22
     gamma31-gamma32
     gamma41-gamma42
     gamma51-gamma52
     gamma61-gamma62
     gamma71-gamma72
     gamma81-gamma82
     gamma91-gamma92
     gamma101-gamma102 = 0;

  *CODE LINEAR PREDICTORS;
  gamma12 = -1*(gamma11);
  gamma22 = -1*(gamma21);
  gamma32 = -1*(gamma31);
  gamma42 = -1*(gamma41);
  gamma52 = -1*(gamma51);
  gamma62 = -1*(gamma61);
  gamma72 = -1*(gamma71);
```

```
gamma82 = -1*(gamma81);
gamma92 = -1*(gamma91);
gamma102 = -1*(gamma101);

beta11 = beta1 + gamma11 ;
beta12 = beta1 + gamma12 ;

beta21 = beta2 + gamma21 ;
beta22 = beta2 + gamma22 ;

beta31 = beta3 + gamma31 ;
beta32 = beta3 + gamma32 ;

beta41 = beta4 + gamma41 ;
beta42 = beta4 + gamma42 ;

beta51 = beta5 + gamma51 ;
beta52 = beta5 + gamma52 ;

beta61 = beta6 + gamma61 ;
beta62 = beta6 + gamma62 ;

beta71 = beta7 + gamma71 ;
beta72 = beta7 + gamma72 ;

beta81 = beta8 + gamma81 ;
beta82 = beta8 + gamma82 ;

beta91 = beta9 + gamma91 ;
beta92 = beta9 + gamma92 ;

eta1 = x1* beta11 + x2* beta21 + x3* beta31 + x4* beta41 + x5* beta51 +
       x6* beta61 + x7* beta71 + x8* beta81 + x9* beta91 + x10* beta101;
eta2 = x1* beta12 + x2* beta22 + x3* beta32 + x4* beta42 + x5* beta52 +
       x6* beta62 + x7* beta72 + x8* beta82 + x9* beta92 + x10* beta102;

*PARTIAL CREDIT  MODEL;
pi0 = 1 / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );
pi1 = exp(eta1) / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );
pi2 = exp(eta1+eta2) / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );


*DEFINE LIKELIHOOD;
z = (pi0**y0)*(pi1**y1)*(pi2**y2)*(pi3**y3);
if (z > 1e-8) then ll = log(z);
else ll=-1e100;
```

```
  model y0 ~ general(ll);

  *SPECIFY RANDOM EFFECT DISTRIBUTION;
  *none;

  *OBTAIN INITIAL ESTIMATES;
  ods output ParameterEstimates = parest ;

run;

*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

*~~~ RUN NLMIXED FOR FINAL ESTIMATES ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

proc nlmixed data= PCM;

  *READ IN INITIAL ESTIMATES;
  parms / data = parest;

  *CODE LINEAR PREDICTORS;
  theta = u1 ;

  gamma12 = -1*(gamma11);
  gamma22 = -1*(gamma21);
  gamma32 = -1*(gamma31);
  gamma42 = -1*(gamma41);
  gamma52 = -1*(gamma51);
  gamma62 = -1*(gamma61);
  gamma72 = -1*(gamma71);
  gamma82 = -1*(gamma81);
  gamma92 = -1*(gamma91);
  gamma102 = -1*(gamma101);

  beta11 = beta1 + gamma11 ;
  beta12 = beta1 + gamma12 ;

  beta21 = beta2 + gamma21 ;
  beta22 = beta2 + gamma22 ;

  beta31 = beta3 + gamma31 ;
  beta32 = beta3 + gamma32 ;

  beta41 = beta4 + gamma41 ;
  beta42 = beta4 + gamma42 ;

  beta51 = beta5 + gamma51 ;
```

```
  beta52 = beta5 + gamma52 ;

  beta61 = beta6 + gamma61 ;
  beta62 = beta6 + gamma62 ;

  beta71 = beta7 + gamma71 ;
  beta72 = beta7 + gamma72 ;

  beta81 = beta8 + gamma81 ;
  beta82 = beta8 + gamma82 ;

  beta91 = beta9 + gamma91 ;
  beta92 = beta9 + gamma92 ;

  eta1 = x1* beta11 + x2* beta21 + x3* beta31 + x4* beta41 + x5* beta51 +
            x6* beta61 + x7* beta71 + x8* beta81 + x9* beta91 + x10* beta101 + theta;
  eta2 = x1* beta12 + x2* beta22 + x3* beta32 + x4* beta42 + x5* beta52 +
            x6* beta62 + x7* beta72 + x8* beta82 + x9* beta92 + x10* beta102 + theta;

  *PARTIAL CREDIT  MODEL;
  pi0 = 1 / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );
  pi1 = exp(eta1) / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );
  pi2 = exp(eta1+eta2) / (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2) );


  *DEFINE LIKELIHOOD;
  z = (pi0**y0)*(pi1**y1)*(pi2**y2)*(pi3**y3);
  if (z > 1e-8) then ll = log(z);
  else ll=-1e100;
  model y0 ~ general(ll);

  *SPECIFY RANDOM EFFECT DISTRIBUTION AND OBTAIN EMPIRICAL
BAYES ESTIMATES;
  random  u1 ~ normal(0, s1*s1) subject = person_id OUT=bayesest;

run;

*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

*************************************************************************;
```

NOTE. THIS PROGRAM WAS OBTAINED AND MODIFIED FROM
HARTZEL, AGRESTI, AND CAFFO (2001).

ALSO NOTE THEY STATE THE FOLLOWING:

"With Gauss-Hermite quadrature, computer
underflow can be a problem mainly when there are many within-cluster
observations. For most data sets in our experience, however, it is
the number of clusters that is large and not the number of
observations within a cluster. In using NLMIXED, we addressed this
problem by assigning the likelihood to a very small number within the
limits of computer precision. Specifically we entered

if (z > 1e-8) then ll = log(z); else ll=-1e100

for this purpose."

**********;

APPENDIX C.

Example of the Input Data Structure

obs	y0	y1	y2	y3	id	item	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	1	2	1	1	0	0	0	0	0	0	0	0	0
3	0	0	0	1	3	1	1	0	0	0	0	0	0	0	0	0
4	0	0	0	1	4	1	1	0	0	0	0	0	0	0	0	0
5	0	0	0	1	5	1	1	0	0	0	0	0	0	0	0	0
6	0	0	1	0	6	1	1	0	0	0	0	0	0	0	0	0
7	0	0	0	1	7	1	1	0	0	0	0	0	0	0	0	0
8	0	0	0	1	8	1	1	0	0	0	0	0	0	0	0	0
9	0	0	0	1	9	1	1	0	0	0	0	0	0	0	0	0
10	0	0	0	1	10	1	1	0	0	0	0	0	0	0	0	0
...																
401	0	1	0	0	1	5	0	0	0	0	1	0	0	0	0	0
402	0	0	1	0	2	5	0	0	0	0	1	0	0	0	0	0
403	0	0	1	0	3	5	0	0	0	0	1	0	0	0	0	0
404	0	0	1	0	4	5	0	0	0	0	1	0	0	0	0	0
405	0	0	0	1	5	5	0	0	0	0	1	0	0	0	0	0
406	0	0	1	0	6	5	0	0	0	0	1	0	0	0	0	0
407	0	0	1	0	7	5	0	0	0	0	1	0	0	0	0	0
408	0	1	0	0	8	5	0	0	0	0	1	0	0	0	0	0
409	0	1	0	0	9	5	0	0	0	0	1	0	0	0	0	0
410	1	0	0	0	10	5	0	0	0	0	1	0	0	0	0	0
...																
991	0	1	0	0	91	10	0	0	0	0	0	0	0	0	0	1
992	0	0	1	0	92	10	0	0	0	0	0	0	0	0	0	1
993	0	1	0	0	93	10	0	0	0	0	0	0	0	0	0	1
994	1	0	0	0	94	10	0	0	0	0	0	0	0	0	0	1
995	1	0	0	0	95	10	0	0	0	0	0	0	0	0	0	1
996	1	0	0	0	96	10	0	0	0	0	0	0	0	0	0	1
997	1	0	0	0	97	10	0	0	0	0	0	0	0	0	0	1
998	0	1	0	0	98	10	0	0	0	0	0	0	0	0	0	1
999	0	1	0	0	99	10	0	0	0	0	0	0	0	0	0	1
1000	1	0	0	0	100	10	0	0	0	0	0	0	0	0	0	1

REFERENCES

References

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), Objective measurement: Theory and practice (Vol. 3, pp. 143-166). Norwood: Ablex.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21(1), 1-23.

Agresti, A. (1996). An introduction to categorical data analysis. New York: John Wiley & Sons, Inc.

Agresti, A. (2002). Links between binary and multi-category logit item response models and quasi-symmetric loglinear models. Annales de la Faculte des Sciences de Toulouse Mathematiques, 11(4), 443-454.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. Biometrics, 55, 117-128.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. Psychometrika, 43, 3-16.

Andrich, D. (1978). A rating scale formulation for ordered response categories. Psychometrika, 43, 561-573.

Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. Organizational Research Methods, 6(1), 15-43.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M braille edition. Journal of Educational Measurement, 26(1), 67-79.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88, 9-25.

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd. ed.). New York: Springer.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. Psychological Methods, 5(4), 477-495.

ConQuest. (1998). ACER ConQuest: Generalised item response modelling software. Camberwell, Melbourne, Victoria: ACER Press.

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the Rating Scale Model. Applied Psychological Measurement, 14, 355-366.

Doherty, K. M., & Skinner, R. A. (2003). State of the states. In Quality Counts 2003: If I Can't Learn From You. Education Week, 22(17), 75-76, 78.

Donoghue, J.R., Holland, P.W., & Thayer, D.T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P.W. Holland & H. Wainer (eds.), Differential item functioning (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Donoghue, J. R., & Hombo, C. M. (2003). An extension of the hierarchical raters' model to polytomous items. Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. Psychometrika, 56, 495-515.

Fahrmeir, L., & Tutz, G. (2001). Multivariate statistical modelling based on generalized linear models (2nd. ed.). New York: Springer-Verlag.

Fox, J. P. (In press, a). Applications of multilevel IRT modeling. .

Fox, J. P. (In press, b). Multilevel IRT using dichotomous and polytomous response data.

Goldstein, H. (2003). Multilevel statistical models (3rd. ed.). New York: Oxford University Press.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. Statistical Modelling, 1(3), 177-193.

Hargrove, L. L., & Mao, M. X. (1997). Three-level HLM modeling of academic and contextual variables related to SAT scores in Texas. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Hargrove, L. L., Mao, M. X., & Barkanic, G. (1996). HLM modeling of coursework, AP, and other academic contextual variables related to SAT scores in Texas. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Hargrove, L. L., & Mellor, L. T. (1994). An HLM exploration of between-school effects related to within-school SAT score differences in Texas: Accountability

implications. Paper presented at the National Council on Measurement, New Orleans, LA.

Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. Statistical Modelling, 1(2), 81-102.

Hedeker, D., & Gibbons, R. D. (1993). MIXOR: A computer program for mixed-effects ordinal, probit, and logistic regression analysis. University of Illinois at Chicago.

Kamata, A. (1998). Some generalizations of the Rasch model: An application of the Hierarchical Generalized Linear Model. Unpublished doctoral dissertation, Michigan State University.

Kamata, A. (2001). Item analysis by the Hierarchical Generalized Linear Model. Journal of Educational Measurement, 38(1), 79-93.

Kim, S. H. (2000). An investigation of the Likelihood Ratio Test, the Mantel Test, and the Generalized Mantel-Haenszel Test of DIF. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kim, W. (2003). Development of a Differential Item Functioning (DIF) procedure using the Hierarchical Generalized Linear Model: A comparison study with logistic regression procedure. Unpublished doctoral dissertation, Pennsylvania State University.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. Journal of the Royal Statistical Society, Series B, Methodological, 58, 619-656.

Linacre, J. M. (1994). Many-facet Rasch measurement. Chicago: MESA Press.

Lord, F. M. (1980). Applications of Item Response Theory to practical testing problems. Hillsdale: Lawrence Erlbaum Associates, Inc.

Luppescu, S. (2002). DIF detection in HLM. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Maier, K. S. (2000). Applying Bayesian methods to hierarchical measurement models. Unpublished doctoral dissertation, University of Chicago.

Maier, K. S. (2001). A Rasch hierarchical measurement model. Journal of Educational and Behavioral Statistics, 26(3), 307-330.

Maier, K. S. (2002). Modeling Incomplete Scaled Questionnaire data with a Partial Credit Hierarchical Measurement Model. Journal of Educational and Behavioral Statistics, 27(3), 271-289.

Manalo, J. R. (2004). The accuracy and application of the AIC, BIC, and CAIC in hierarchical linear modeling. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association, 58, 690-700.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

Michigan Education Assessment Program (2003). *Design and Validity of the Test.* Retrieved March, 2004, from http://www.meap.org/.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. Applied Psychological Measurement, 11(1), 81-91.

Miyazaki, Y. (2000). Incorporating factor analysis into hierarchical models. Unpublished doctoral dissertation, Michigan State University.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159-176.

Patz, R. J. (1996). Markov Chain Monte Carlo methods for Item Response Theory models with applications for the National Assessment of Educational Progress. Unpublished doctoral dissertation, Carnegie Mellon University.

Patz, R. J., Junker, B. W., Johnson, M. A., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. Journal of Educational and Behavioral Statistics, 27, 341-384.

Patz, R. J., Junker, B. W., & Johnson, M. S. (1999). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2001). GLLAMM Manual. Department of Biostatistics and Computing, Institute of Psychiatry, Kings College, University of London.

Raudenbush, S., Bryk, A., & Congdon, R. (2001). HLM for Windows: Hierarchical Linear and Non-linear Modelling (Version 5.04). Lincolnwood: Scientific Software International.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods (2nd. ed.). London: Sage Publications, Inc.

Reckase, M. D. (1991). The discriminating power of items that measure more than one dimension. Applied Psychological Measurement, 15(4), 361-373.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. Applied Psychological Measurement, 21, 25-36.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. Multivariate Behavioral Research, 35(4), 543-568.

Rijmen, F., Tuerlinckx, F., De Boek, P., & Kuppens, P. (2003). A Nonlinear mixed model framework for Item Response Theory. Psychological Methods, 8(2), 185-205.

S-PLUS. (2000). S-Plus 2000. Cambridge: Mathsoft, Inc.

SAS. (2001). Statistical Analysis Software. Cary: SAS Institute.

Singer, J.D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. Journal of Educational and Behavioral Statistics, 24(4), 323-355.

Smith, E. V., & Johnson, B. D. (2000). Attention deficit hyperactivity disorder scaling and standard setting using Rasch measurement. Journal of Applied Measurement, 1(1), 3-24.

Snijders, T. A. B., & Bosker, R. J. (1999). Multilevel analysis : An introduction to basic and advanced multilevel modeling. London: Sage Publications.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). BUGS 0.5 Examples (Vol. 1). Cambridge, U.K.: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

STATA. (2000). Stata Statistical Software (Version 6). College Station, TX.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. Applied Measurement in Education, 16(1), 1-26.

Tuerlinckx, F. & Wang, W.C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Eds.), Explanatory item response models: A generalized linear and nonlinear approach (pp. 75-109). New York, NJ: Springer-Verlag.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. The NAEP 1996 Technical Report, NCES 1999–452, by Allen, N.L., Carlson, J.E., & Zelenak, C.A. (1999). Washington, DC: National Center for Education Statistics.

Wang, W. C., Wilson, M., & Adams, R. J. (1998). Measuring individual differences in change with Multidimensional Rasch Models. Journal of Outcome Measurement, 2(3), 240-265.

Wang, W. C. & Su, Y.H. (2004). Factors influencing the Mantel and Generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. Applied Psychological Measurement, 28(6), 450-480.

WINSTEPS (1999). *Rasch-Model Computer Program.* Chicago: MESA Press.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting Differential Rater Functioning over Time (DRIFT) using a Rasch multi-faceted Rating Scale Model. Journal of Applied Measurement, 2(3), 256-280.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. Journal of Statistical Computation and Simulations, 48, 233-243.

Wright, B. D., & Masters, G. N. (1982). Rating Scale Analysis. Chicago: Mesa Press.

Zhang, Y., & Zhang, L. (2002). Modeling school and district effects in the math achievement of Delaware students measured by DSTP: A preliminary application of Hierarchical Linear Modeling in accountability study. Paper presented at the American Educational Research Association, New Orleans, LA.

Zhu, W., Updyke, W. F., & Lewandowski, C. (1997). Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. Journal of Outcome Measurement, 1(4), 286-304.