Using Explanatory Item Response Models to Analyze Group Differences in Science Achievement

Derek C. Briggs

University of Colorado, Boulder

January 3, 2007

ABSTRACT

This paper illustrates the use of an explanatory item response modeling (EIRM) approach in the context of measuring group differences in science achievement. The distinction between item response models and EIRMs, recently elaborated by De Boeck & Wilson (2004), is presented within the statistical framework of generalized linear mixed models. It is shown that the EIRM approach provides a powerful framework for both a psychometric and statistical analysis of group differences. This is contrasted with the more typical two-step approach, in which psychometric analysis (i.e., measurement) and statistical analysis (i.e., explanation) occur independently. The two approaches are each used to describe and explain racial/ethnic gaps on a standardized science test. It is shown that the EIRM approach results in estimated racial/ethnic achievement gaps that are larger than those found in the two-step approach. In addition, when science achievement is examined by subdomains, the magnitude of racial/ethnic gap estimates under the EIRM approach are more variable and sensitive to the inclusion of contextual variables. These differences stem from the fact that the EIRM approach allows for disattenuated estimates of group level parameters, while the two-step approach depends upon estimates of science achievement that are shrunken as a function of measurement error.

INTRODUCTION

Item response theory (IRT) models are primarily viewed as tools for measurement. In this capacity they have been, and continue to be, applied with much success. In recent years it has become increasingly apparent that IRT models exist as special cases of a broader statistical framework (c.f., Verhelst & Verstralen, 2001; Kamata, 2001; Rijmen, Tuerlinckx, De Boeck & Kuppens, 2003; De Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004). It has been shown that when IRT models are cast within the framework of generalized linear mixed models or nonlinear mixed models, it becomes possible to specify—with great flexibility—targeted research questions about both within-person differences in item response probabilities, and between-person differences in the latent construct(s) being measured. Traditionally, these latter sorts of questions are addressed after student-level measures have already been estimated. For example, when a sample of students are administered a standardized test in science, we might be interested both in (a) the extent to which student performance differs as a function of race/ethnicity, and (b) the extent to which these differences vary as a function of different subdomains of science achievement. One way to approach this would be to separately scale each subdomain with an IRT model, and in a subsequent step use student scale scores from each subdomain as outcome variables to be regressed on racial/ethnic dummy variables. With this approach, the underlying IRT model is only of interest as a way to generate the outcome variable. The actual answers we arrive at for (a) and (b) are based upon a linear regression that is essentially independent of the chosen item response model. However, when the interest of our research question is not restricted to measuring students, but includes an interest in explaining group differences among students, this two-step approach will often not be the best one to take, for reasons I will establish and illustrate in this paper.

PURPOSE

De Boeck and Wilson (2004) coined the term “explanatory item response models” to characterize the use of IRT as a tool for both measurement and explanation. In their edited textbook, examples are given of many different instances in which explanatory item response models are applied to empirical data. Unfortunately, there are relatively few examples using data from large-scale standardized achievement tests. The purpose of this paper is to (1) further introduce the concept of explanatory item response models in the context of addressing the sorts of substantive research questions typical of the field of education, and (2) highlight some of the potential advantages to taking an explanatory item response modeling approach in the context of comparing group differences.

There are four principal sections that follow. In the first section I present the broader statistical framework within which IRT models can be viewed as a special case. In the second section, I describe a dataset gathered from a sample of students who have been administered the standardized science assessment known as the Partnership for the Assessment of Standards-based Science (PASS) test . Over the next two sections of the paper this dataset is used to illustrate the distinction between using IRT models strictly for the purpose of measuring individuals, and using IRT models in the context of both measuring individuals, and explaining group differences among those individuals. In the third section, I show how research questions about group differences are typically addressed using the two-step approach of measurement followed by explanation. In the fourth section, I contrast the two-step approach to an explanatory item response modeling approach, and point out where the two approaches can lead to different conclusions about the same research questions. The paper concludes with a summary of some of the major strengths and limitation of the explanatory item response modeling approach.

BACKGROUND: EXPLANATORY ITEM RESPONSE MODELS

The presentation in this section draws upon a more extensive explication provided by Rijmen, Tuerlinckx, De Boeck & Kuppens (2003). Imagine we have some test instrument comprised of multiple-choice items that are to be scored dichotomously. Let the random variable Y_{ni} represent the item response from person *n* to item *i*. When the correct response is selected, Y_{ni} = 1, otherwise, Y_{ni} = 0. The probability of a correct response can be expressed as a function of fixed effects, $\boldsymbol{\beta}$, and random effects, $\boldsymbol{\theta}_n$, such that

$$P(Y_{ni}=1 \mid \mathbf{x}_{ni}, \mathbf{z}_{ni}, \boldsymbol{\beta}, \boldsymbol{\theta}_n) = \frac{\exp(\mathbf{x}'_{ni}\boldsymbol{\beta} + \mathbf{z}'_{ni}\boldsymbol{\theta}_n)}{1+\exp(\mathbf{x}'_{ni}\boldsymbol{\beta} + \mathbf{z}'_{ni}\boldsymbol{\theta}_n)}, \qquad (1)$$

where $\mathbf{x}'_{ni}$ is an observed *P*-dimensional covariate vector for *P* fixed effects; $\mathbf{z}'_{ni}$ is an observed *Q*-dimensional covariate vector for *Q* random effects; $\boldsymbol{\beta}$ is the *P*-dimensional parameter vector of fixed effects, and $\boldsymbol{\theta}_n$ is the *Q*-dimensional parameter vector of random effects associated with respondent *n*. In (1), all observations are assumed to be independent realizations from an exponential family distribution, conditional on the specified random effects, covariates, and fixed effects. The generalized linear form of the model is most readily apparent when the mean responses of Y_{ni} are mapped to the linear predictors with what is known as a link function. In this case, with Y_{ni} taking on a Bernoulli distribution, $E(Y_{ni} = 1 | \mathbf{x}_{ni}, \mathbf{z}_{ni}, \boldsymbol{\beta}, \boldsymbol{\theta}_n) = P(Y_{ni} = 1 | \mathbf{x}_{ni}, \mathbf{z}_{ni}, \boldsymbol{\beta}, \boldsymbol{\theta}_n) = \pi_{ni}$, so the link function is the logit function, $L(\pi_{ni}) = \ln \frac{\pi_{ni}}{1-\pi_{ni}}$. Given the logit link, (1) can be rewritten as

$$L(\pi_{ni}) = \mathbf{x}'_{ni}\boldsymbol{\beta} + \mathbf{z}'_{ni}\boldsymbol{\theta}_n. \tag{2}$$

The expression above is an example of a *generalized linear mixed model* (Breslow & Clayton, 1993; McCulloch & Searle, 2001). The random effects $\boldsymbol{\theta}_n$ are typically assumed to have a multivariate normal distribution with a mean vector of **0** and covariance matrix Σ. The fixed and random effect covariate vectors merit further discussion in the context of the data typically considered in IRT. Let the matrices **X** and **Z** correspond to the respective fixed and random effect covariate vectors stacked over items and persons. Rijmen et. al. (2003; 187) define three categories of distinct covariates:

1. *Item covariates*: A covariate is an item covariate if and only if the elements of the corresponding column of **X** (and/or **Z**) vary across items but are constant across persons.
2. *Person covariates*: A covariate is a person covariate if and only if the elements of the corresponding column of **X** (and/or **Z**) vary across persons but are constant across items.
3. *Person-by-item covariates*: A covariate is a person-by-item covariate if and only if the elements of the corresponding column of **X** (and/or **Z**) vary across both persons and items.

When item response models are used in educational research, students are typically the unit of analysis. Once person or person-by-item covariates are included to describe or explain differences among students, the model, specified within the GLMM framework becomes an *explanatory item response model* (EIRM). Note that in psychological research, often the items themselves are the unit of analysis. In this case the inclusion of item covariates (or person-by-item covariates) also leads to an EIRM.

To further clarify the framework, Figure 1 shows the GLMM d951 0 uctgure

$$L(\pi_{ni}) = \theta_n + \beta_i \,. \tag{3}$$

This is recognizable as a Rasch Model, where the variable θ_n is assumed to represent the value of some latent construct of interest for person *n*, and β_i represents item easiness. The distribution of the latent variable θ_n is typically assumed to be normal[1] and the mean is constrained to equal 0 (or alternatively the mean of β_i is constrained to equal 0) for purposes of model identification. The distribution of θ_n is often referred to as the "population" distribution. From the perspective of the GLMM framework, the Rasch Model is a model in which the clustering of item responses within respondents is a function of item-specific fixed effects and one person-specific random effect. The Rasch Model is solely a measurement model because no attempt is made to explain the differences in the characteristics of persons or items by introducing other covariates into the **X** and **Z** matrices.

Now I show how the GLMM framework can be used to move from the Rasch Model to a relatively complex EIRM. Say we are interested in specifying a multidimensional model for group differences in achievement. To account for multidimensionality, we can include more than one random effect such that

$$L(\pi_{ni}) = \mathbf{z}'_{ni}\boldsymbol{\theta}_n + \beta_i \,. \tag{4}$$

If our interest is in explaining group differences associated with the random effect vector $\boldsymbol{\theta}_n$, then we can introduce person covariates into the model such that

$$\boldsymbol{\theta}_n = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \tag{5}$$

where $\mathbf{X}^*$ is a matrix of person covariates associated with each of *Q* random effects, $\boldsymbol{\beta}$ is a matrix of fixed effect parameters for each random effect/covariate combination, and $\boldsymbol{\varepsilon}_n$

represents a vector containing Q random error terms. The expression above constitutes a population model at the person level, and has been previously described as a latent regression (Adams, Wilson & Wu, 1997). In this case, we would be specifying a *multidimensional* latent regression. Once a multidimensional latent regression has been specified, the fixed effect covariate vector, $\mathbf{x}'_{ni}$, from (2) now includes both item and person covariates, and the vector $\boldsymbol{\varepsilon}_n$ becomes a multidimensional random effect. More importantly, with the addition of person covariates, the model has changed in nature from providing a measurement of the random effects $\boldsymbol{\theta}_n$, to describing or—depending upon the nature of the variables included in $\mathbf{X}^*$—explaining group differences in $\boldsymbol{\theta}_n$. The EIRM that results from the pairing of (4) with (5) is a multilevel model, having both a within-person level and a between-person level, with the multidimensional Rasch Model characterizing the former, and a population model (i.e., latent regression) characterizing the latter.

I have made two simplifying assumptions for didactic reasons. Only models for dichotomous items have been presented, and the range of item response theory models has been limited to the Rasch family of models. The extension to polytomous items involves an additional level of nesting, with item categories $j = 1, \ldots, J$ nested within items. So the GLMM expression of (2) for polytomous items becomes

$$L(\pi_{nij}) = \mathbf{x}'_{nij}\boldsymbol{\beta} + \mathbf{z}'_{nij}\boldsymbol{\theta}_n. \qquad (6)$$

The measurement framework can be extended outside the Rasch family of IRT models with the inclusion of item covariates with unknown values (i.e., discrimination parameters). By doing this we move outside the GLMM framework to a nonlinear mixed modeling framework (NLMM; Davidian & Gilitinan, 1995). A presentation of the extension to the NLMM takes us somewhat outside the scope of the illustration in this paper, so I do not include it here. The interested reader is directed to Rijmen et. al, (2003) and Molenberghs & Verbeke, (2004).

The relevant parameters for item response models and EIRMs specified within the GLMM or NLMM statistical frameworks can be estimated in a number of different ways. Most typically this will involve the use of marginal maximum likelihood (MML) estimation in conjunction with the EM algorithm (Bock & Aitken, 1981). The general idea is to specify a marginal likelihood function, formed by taking the product of the data likelihood and the population distributions for the random effects, and then take the integral over persons. Denote this marginal likelihood as L_θ. Because there is no closed form solution to L_θ, it must be approximated numerically using either Gaussian quadrature, adaptive Gaussian quadrature, or Monte Carlo integration, all within the context of the EM algorithm. A different approach to the estimation of both GLMM and NLMM models involves the use of Bayesian estimation methods (Gelman, Carlin, Stern & Rubin, 2004); in particular Markov Chain Monte Carlo estimation. There are many available programs that can be used for the purposes of parameter estimation. Some of these program exist as procedures within broader statistical software environments. For example, the procedures NLMIXED and GLLAMM in the SAS and STATA software environments provide for MML estimation of models within the GLMM and NLMM frameworks. The program WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2004) has been used for the purpose of Bayesian estimation. Of course, when more traditional IRT models are specified, more specialized programs such as BILOG (Zimowski, Muraki, Mislevy & Bock, 1995) and MULTILOG (Thissen, Chen & Bock, 2002) could also be used, and are typically faster and more convenient. The software ConQuest (Wu, Adams & Wilson, 1998) is an interesting case in that it is specialized for IRT applications, but can be used to specify all the models within the GLMM framework. For a thorough review of estimation and software issues, see Tuerlinkx et. al. (2004).

METHOD

Data Source

The context I will use to illustrate the EIRM approach comes from a sample of schools that chose to administer a large-scale standardized science assessment to their students. The assessment, known as the Partnership for the Assessment of Standards-based Science (PASS) test, was developed to support K-12 schools that were in the process of transforming their science education curricula such that it placed a greater emphasis upon scientific inquiry and hands-on activities. Among school districts involved in this kind of curricular reform, an explicit expectation was that this transformation would help students to build a deeper conceptual understandings of scientific concepts, and that this deeper conceptual understanding would translate into higher achievement on standardized tests designed to measure proficiency in science. The PASS tests were designed to assess the extent to which this is occurring. The test is available in English and Spanish at grades 5, 8 and 10, and was developed to be in alignment with both the *National Science Education Standards* (National Research

Council, 1996) and the *Benchmarks for Scientific Literacy* (American Association for the Advancement of Science, 1993).

The PASS test includes three sections that differ by item format: multiple-choice (MC) items, constructed-response items, and performance task items. In the first section of the test, students are given one of six parallel forms of 29 MC items. The parallel forms of this section of the test were designed to facilitate the horizontal and vertical equating of PASS test across grades and years. This is done through the use of a common item design, in which a subset of linking items are embedded within a given MC test form. This design is similar in nature to the matrix sampling approach taken in large-scale assessments such as the National Assessment for Educational Progress (NAEP). While no single student will answer more than 29 MC items in a single administration of the PASS test, an entire sample of test-takers will have responded to as many as 100 unique MC items. Each MC item is developed to explicitly align with one of five subdomains (i.e., content or process areas in science) as designated by the National Science Education Standards: physical science, earth science, life science, scientific inquiry, and science and technology.

The second and third sections of the PASS test consist of performance tasks and constructed response investigations. For a performance task, students are provided with hands-on equipment and asked to perform short experiments, communicate scientific information, make scientific observations, generate and record data, and analyze results based on their data. A constructed response investigation is similar in nature to a performance task, but does not involve the use of hands-on equipment and places a greater emphasis on secondary analysis and hypothesis testing. Unlike the MC section of the PASS tests, all students are administered the same performance task and constructed response investigation within a given grade and a given year. (For more information about the PASS tests, and to see some sample items across the different formats, see WestEd, 2006.)

Historically, the PASS tests have been calibrated and scaled using IRT, more specifically the Partial Credit Model (Masters, 1982). Scores and corresponding confidence intervals are reported for each grade tested at the student, school, and district levels, and also in terms of the full population of test-takers. The scores are reported in terms of both linearly transformed scale scores and the percentage of total points for each section and subdomain of the test.

Sample Characteristics

In the analyses that follows, I restrict my attention to a sample of 433 10th grade students who completed form 4 of the MC section on the PASS test in 1999. There are two reasons for this restriction. First, it limits the illustration of the EIRM approach within the context of (a) explaining group differences, and (b) doing so as a function of science subdomains. If the full PASS test and test-taking population were used, the EIRM approach would also need to account for a horizontal equating step across test forms, the use of different raters in the open-ended sections, and violations of the IRT local independence assumption in the clustering of the open-ended items around common prompts. All of these adjustments can be accommodated within the GLMM framework, but they make the illustration overly complex. Second, this restriction is intended to keep readers from making unwarranted inferences and generalizations about the PASS test. The aim here is not to evaluate the characteristics of the PASS test and test-takers, but to illustrate an application of the EIRM approach with empirical data. Form 4 was chosen because it happened to have the largest sample of test-takers.

Table 1 presents the available demographic characteristics of the PASS student sample. The students are a random sample from 66 participating schools in four districts from the states of Arizona, California and New Jersey. Participation in PASS is at the discretion of individual school districts, hence the students taking form 4 of the test were drawn from a "population" which itself constitutes a self-selected convenience sample. Among the students who reported information about their race/ethnicity and primary language spoken at home, roughly 46% are nonwhite, and 20% speak a language other than English at home. Students were also asked questions about their attitudes, effort, and academic performance as they related to science. The responses to these questions are summarized for all 10th grade test-takers in the second column of Table 2.

Insert Tables 1 and 2 about here

Research Questions

There are many interesting questions that could be posed on the basis of the variables shown in Tables 1 and 2, and the performance of students on the PASS test. One topic of clear importance among educational researchers in the United States is the achievement gap between white students and racial/ethnic minorities (c.f., Jencks &

Philips, 1998; Camara & Schmidt, 1999, Lee, 2002). With this in mind, we may wish to address the following questions with data from the PASS test:

1. To what extent are there group differences in science achievement as a function of racial/ethnic self-identification?
2. Do these group differences change after controlling for self-reported contextual variables?
3. Do the patterns that emerge from addressing questions 1 and 2 change when the outcome of interest, science achievement, is examined by subdomains?

In the next two sections I address these questions using two different approaches: one in which the GLMM framework is used solely to specify a measurement model, and another whether the GLMM framework is used to specify an EIRM. I refer to the former as the "two-step approach," and the latter as the EIRM approach.

RESULTS

Measurement Followed by Explanation: The Two-Step Approach

One way to compare group differences in achievement on the PASS test would be to take the two-step approach. In the first step, a unidimensional or multidimensional Rasch Model can be used as the basis for generating science achievement estimates for the 433 students in our sample. The estimates typically used in the context of a large-scale standardized test are the means of each student's posterior distribution, known as the expected a posteriori, or EAP[2]. In the second step, the EAP for each student on one or more dimensions can be used as an outcome variable to be regressed upon a set of racial/ethnic dummy variables[3].

Step 1: Measuring Science Achievement

Reporting Test Results Unidimensionally

The Rasch Model could be applied to the PASS data using any number of software routines. In this case I have used the software ConQuest, but the same analysis could also have been conducted using, for example, NLMIXED in SAS. When taking a two-step approach, the emphasis in the first step is on producing estimates of student achievement. Yet there is also a great deal of psychometric information about the underlying test being produced in the form of item parameter estimates, standard errors and fit statistics. Another important piece of information that is generated in the measurement stage is known as the deviance statistic. This statistic equals $-2\log_e L$, where in this context *L* represents the marginal likelihood L_θ previously described. The deviance statistic measures the deviation between a saturated IRT model and the fitted IRT model that has been specified. The smaller the number, the better the fit of the model. Nested IRT models that is, where one model can be written to equal the other by imposing one or more parameter constraints can be statistically evaluated by comparing the difference in their deviance statistics, which will have an approximately Chi-Square sampling distribution with degrees of freedom equal to the difference in each models free parameters.

One piece of psychometric information from step 1 of a two-step approach that is almost always reported before proceeding to stage 2 is test score reliability. While the concept of reliability is fundamentally one of classical test theory, a reliability index can also be estimated for an IRT analysis as

$$R_p = 1 - \frac{\bar{\sigma}_p^2}{\text{var}(\theta)}, \tag{7}$$

where the R_p represents *marginal reliability* (Green et. al., 1984; Mislevy et. al., 1992; Adams, 2006). In Equation 7 the term $\bar{\sigma}_p^2$ represents the mean variance taken over each student's estimated posterior distribution of science achievement, while the term $\text{var}(\theta)$ represents the population variance in the distribution of science achievement. The ratio of $\bar{\sigma}_p^2$ to $\text{var}(\theta)$ quantifies the extent to which the uncertainty in a student's estimated achievement can be reduced by administering the set of MC items found on the PASS test form. The values of the marginal reliability index R_p will, like Cronbach's alpha, range between 0 and 1, and provide for a similar interpretation. After applying the unidimensional Rasch Model to the 29 MC items on the PASS test form, $\bar{\sigma}_p^2 = .137$ and $\text{var}(\theta) = .653$, so $R_p = .79$.

Reporting Test Results By Subdomain

There are two reasons why it might be sensible to report the results from the PASS test by subdomains. The first reason is that if the test is actually multidimensional as a function of these subdomains, but modeled as if it were unidimensional, then the IRT assumption of local independence will be violated. This will have an impact both on estimates of student achievement and the uncertainty surrounding these estimates. In the context of using the Rasch Model, apparent problems such as misfitting items and differential item functioning can sometimes be attributed to the presence of multidimensionality in a test assumed to be unidimensional. All this constitutes what I would classify as a purely psychometric reason for using a multidimensional item response model. A second reason for reporting PASS test scores by subdomains is to facilitate diagnostic interpretations of the results. For example, as Monfils, Dawber, Han & Henderson-Montrero (2006) point out, "since the passage of the No Child Left Behind Act of 2001 states are required to report 'diagnostic' score results for each content sub-domain/strand provided the subscores are based on sufficient information." Under this second rationale for reporting test results by subdomain, even if the underlying test construct appears to be unidimensional, there is still a psychometric reason—the need for "sufficient information"—to model subdomains multidimensionally. However, for the second rationale, the psychometric basis is secondary to other substantive concerns that may well be codified into policy (i.e., the desire to provide teachers and students with disaggregated information about test performance). In other words, there will be times when test results must be reported multidimensionally by subdomains even when there is little to no psychometric justification for the practice.

It has been previously established that subdomain scores deriving from a multidimensional item response model will be more reliable than subdomain scores deriving from a unidimensional model (c.f., Adams, Wilson & Wu, 1997; Briggs & Wilson, 2002). The concept at work here has been described as "subscore augmentation" by Wainer et. al. (2001), and has received considerable attention of late[4]. The concept can be illustrated within the present context because the results of the MC section of the PASS test are, in fact, disaggregated and reported to students, schools and districts in terms of the five subdomains of earth science, life science, physical science, scientific inquiry, and science & technology. On each of the six MC forms of the PASS tests, students respond to items that are intended to correspond directly to each of these different subdomains. Across all six forms, the average numbers of items per subdomain are 8, 6, 6, 4 and 4. On form 4 of the test considered here, the respective numbers are 7, 6, 7, 6 and 3. One way to report student-level scores for the PASS subdomains would be to produce five separate unidimensional estimates of student achievement for each subdomain. When the Rasch Model is used, this is analogous to reporting the percentage of MC items answered correctly by subdomain. The problem with this approach is that it would result in very unreliable estimates of subdomain scores. The marginal reliabilities when the five subdomains are each modeled unidimensionally are shown in Table 3, and range from a low of .241 (life science) to a high of .545 (earth science). The upshot is that with reliabilities so low it would be very difficult to rule out measurement error as a principal explanation for individual differences in student scale scores across subdomains.

Insert Table 3 about here

An attractive alternative is to augment the reliability of these subdomain scores by modeling them as part of a multidimensional Rasch Model. Because each MC item is intended to map onto only one of the five subdomains, I proceed by taking a confirmatory approach in which each subdomain is treated as a distinct dimension of science achievement[5]. This adjustment to the measurement model is made quite easily within the GLMM framework. Instead of a single random effect as in Equation 3, we specify the vector $\boldsymbol{\theta}_n$ in Equation 4 to include five random effects, with the covariate vector $\mathbf{z}'_{ni}$ specified to consist for any given item, of a single 1 and four 0s. The random effects $\boldsymbol{\theta}_n$ (i.e., dimensions) are assumed to have a multivariate normal distribution with a mean vector $\boldsymbol{\mu}$ and a variance-covariance matrix Σ. Analogous to the unidimensional Rasch Model, for identification purposes either $\boldsymbol{\mu}$ is constrained to 0, or the mean of the item difficulties associated with each of the five dimensions is constrained to equal 0. For the PASS data I apply the latter identification approach, and again use the software ConQuest to estimate the relevant item and person parameters.

In Table 3 the marginal reliabilities of student subdomain scores for the multidimensional Rasch model can be compared to the reliabilities from the five separate unidimensional models of the subdomains. The augmentation of marginal reliabilities through use of the multidimensional model is evident: reliability estimates now range from a low of .645 (science & technology) to a high of .749 (physical science). The reliability of each subdomain is augmented within the multidimensional model because individual student achievement for each science subdomain is estimated as a weighted average between a student's observed performance on items associated with the subdomain, and the population means for all subdomains. The values of the weights in the weighted average are a function of the original reliability of each subdomain, and the correlation between the subdomains. The subdomain reliabilities are augmented because one subset of items effectively "borrows strength" from the present information available in remaining items, and prior information available from the underlying population distributions. The optimal way to augment the reliability for any given subdomain through the use of the multidimensional approach is to borrow strength from other subdomains that are 1) themselves more reliable than the target subdomain, and 2) strongly correlated with the target subdomain. This is why in the context here the reliability augmentation is strongest for the life science subdomain: it is strongly correlated (r > .85) with three more reliable subdomains (earth science, physical science and scientific inquiry). By contrast, the augmentation effect is weaker (though still quite substantial) for the earth science and scientific inquiry subdomains because these subdomains tend to be more reliable than the other subdomains from which they are borrowing strength.

Step 2: Describing and Explaining Group Differences in Science Achievement

Linear Regressions with Unidimensional EAP Estimates

After applying the Rasch Model, we are left with 433 EAP estimates as measures of science achievement. To explore the extent to which there are group differences in these measures as a function of race/ethnicity, the EAP values are regressed on the dummy variables BLACK, HISPANIC, ASIAN and OTHER (with the dummy variable WHITE as the reference group). There were 13 students who provided no information about their race/ethnicity, so they have been excluded from the model, decreasing the sample size from 433 to 420. Two sets of estimated coefficients from this linear regression are shown in the columns under Model 1 in Table 4. In the first set (Model 1a), the coefficients are provided in logit units; in the second set (Model 1b), the coefficients are provided in effect size units, after dividing the coefficients in the first set by the observed standard deviation (SD) of the student EAPs that were estimated in step 1. From this it is clear that there are sizeable and statistically significant gaps in the mean performance of black and Hispanic students relative to the white student reference group: black and Hispanic students score .87 and .75 SDs lower on the PASS MC section, respectively. On the other hand, the gap between Asian and white students is negligible and not statistically significant.

Insert Table 3 about here

To put the magnitude of these gaps in a broader perspective, the achievement gaps found for black and Hispanic 17 year olds who took the long-term NAEP science test in 1999 relative to their white counterparts were 1.19 and .68 SD units (Campbell, Hombo & Mazeo, 2000). The next question is whether some portion of these gaps can be explained by differences in the contextual variables gathered from students taking the PASS test.

For the PASS subsample, there are small positive correlations between estimated student achievement on the MC section of the test and 1) the amount of time students spend on their science homework (*r* = .193), and 2) whether students report higher grades in their science courses relative to other courses (*r* = .214). There are small negative correlations between estimated science achievement and whether a student reports that English is *not* the primary language spoken at home ($r = -.174$). To the extent that the values for these contextual variables differ among students in different racial/ethnic groups, they may serve to explain some portion of the observed gaps. In fact, when the variables in Table 2 are crosstabulated by racial/ethnic group (not shown here), we find that relative to white test-takers,

- black test-takers were just as likely to report doing one or more hours of homework per week, and were equally likely to report that they got better grades in science than in other courses.
- Hispanic test-takers were less likely to report doing one or more hours of homework per week, but equally likely to report that they got better grades in science than in other courses. Hispanic students were much more likely to report that English was not the primary language spoken at home—this was true of almost half of the Hispanic test-takers in the sample.
- Asian test-takers were more likely to report doing one or more hours of homework per week, more likely to report that they got better grades in science than in other courses, and more likely to report that English was not the primary language spoken at home. As with Hispanic students, the latter was true of roughly half of all Asian test-takers in the sample.

From this information we can hypothesize that the addition of these contextual variables to the linear regression model should affect the size of the gaps for Hispanic and Asian students, but not for black students. Model 2 in Table 4 presents the results from adding contextual variables to the linear regression model. The variables added to the model are academic performance (GRADES), time spent on science homework (HOMEWORK), the interaction of speaking a language other than English at home with being either Hispanic or Asian (HISP_ESL, ASIAN_ESL), and a newly created variable, interest in science (INTEREST). The definitions of the variables GRADES, HW, HISP_ESL, and ASIAN_ESL can be discerned from their descriptions in Table 2. For GRADES, the lowest value ("grades in science are lower compared to other subjects") takes a values of 0, while the highest value("grades in science are lower compared to other subjects ") takes a value of 2. For HOMEWORK, the lowest value ("no homework") takes a values of 0, while the highest value("more than 2 hours of homework per week") takes a value of 3. The variable INTEREST was created by taking the sum of the variables CONCEPTS, FUTURE, and PARENTS as defined in Table 2. For each of the latter variables the lowest category ('no") takes a value of 0 and the highest category ("yes") takes a value of 2. Hence the variable INTEREST has a range from 0 to 6 with a mean of 2.9 and an SD of 1.5. The variable is a crude measure of a student's interest in science: higher scores would typically represent students who responded with a "yes" or "sometimes" when asked if they found science useful outside of school, thought science would be useful to them in the future, and talked to their parents about what they did in science class. Using Cronbach's alpha, the estimated reliability of this three item measure is about .66. External variables reflecting student and family socioeconomic status along with academic transcripts of student grades would be other ideal variables to include in the regression model, but these were not collected as part of the PASS test administration.

As predicted, the inclusion of these contextual variables has almost no effect on the size of the black-white gap (which actually increases slightly), reduces the size of the Hispanic-white gap by about .11 SD units, and reverses the sign on the Asian-white gap (though it is still not statistically significant). The latter effect is largely attributable to the inclusion of the interaction variable ASIAN_ESL. For Asian test-takers who do not speak English at home, the size of the achievement gap is predicted to increase on average by .87 SD units. Interestingly, the inclusion of the variable HISP_ESL does not appear to have the same sort of impact on the Hispanic achievement gap. The newly included variables GRADES and HOMEWORK each have a positive partial association with performance on the PASS test. Those students who report that their grades in science are higher than their grades in other classes are predicted to score about .70 SDs higher on the test relative to those students who report that their grades in science are lower than their grades in other classes. Likewise, the model predicts a .63 SD advantage for students who report that they do more than 2 hours of homework a week relative to students who report doing no homework per week. The proxy variable for student interest in science, INTEREST, has no significant partial association with performance on the PASS test.

So far it would seem there are at least two conclusions that can be drawn from the measurement of PASS test performance and subsequent linear regressions.

1. There are sizable gaps in the performance of black and Hispanic students relative to their white counterparts.

2. The size of the gap for Asian and Hispanic students is reduced when contextual variables are added to the linear regression model, but stays about the same for blacks students.

Next we turn to our third research question, and examine whether the pattern of results that were found above remains the same when the PASS test results are reported by subdomain.

Linear Regressions with Multidimensional EAP Estimates

In parallel with the unidimensional two-step approach, student level EAPs for each of the five science subdomains can be used as outcome variables in subsequent linear regressions to both describe and explain the size of racial/ethnic differences in estimated science achievement. The results from these regressions are provided under the columns in Table 5. There are two distinct columns for each subdomain; one column reflects a regression model in which student subdomain EAP scores are racial/ethnic dummy variables (e.g., ES 1), and another column that adds contextual variables to potentially explain the differences in racial/ethnic achievement (e.g., ES 2).

Insert Table 5 about here

Relative to the regressions in which science achievement was measured unidimensionally, the results in Table 5 tell the same story with regard to the magnitude of racial/ethnic gaps and the extent to which they change when contextual variables are included in the model. So it would appear that while examining PASS results by subdomain does produce the expected augmentation effect on the marginal reliability of each PASS subdomain, it provides for no new information about group differences. This would seem to raise the question of whether, using a two-step approach, it was worthwhile to model the PASS test by subdomain at the individual student level in the first step.

The multidimensional and unidimensional Rasch Models are nested models, and can be evaluated statistically by parameterizing the multidimensional model such that if the variances associated with additional random effects are 0, it reduces to the unidimensional model (c.f., Rijmen & Briggs, 2004, 252). The relative fit of the two models can be compared using a likelihood ratio (LR) test-statistic, computed as the difference in deviance between the two models, which in this case is $14245.3 - 14182.5 = 62.86$. It has been shown by Verbeke & Molenberghs (1997) that the asymptotic null distribution of this LR test-statistic will be an equally weighted mixture of two Chi-Square distributions, in this case with *df* = 1 and *df* = 5. We find that p < .001, which supports the conclusion that the multidimensional Rash Model has a better goodness of fit than the unidimensional Rasch Model.

There is, however, another established method for evaluating test dimensionality. That method is the nonparametric procedure known as DIMTEST (Stout, 1987; Nandakumar & Stout, 1993; Stout , Froelich & Gao, 2001). When the procedure DIMTEST was applied to MC form 4 the PASS test, only weak evidence could be found to reject the null hypothesis of essential unidimensionality. Using a variety of confirmatory mappings of items to the PASS subdomains, in no case was *p* < .20. The latter finding presents us with something of a paradox. The aim in reporting individual student scores by subdomain is to provide more diagnostic information reliably. To provide scores that are reliable through subscore augmentation, the more strongly correlated the dimensions, the better. But if the dimensions are too strongly correlated, then for all intents and purposes at the student level they are essentially unidimensional. Hence the reporting of separate scores for students according to each subdomain will be largely redundant, even when such reporting is required by law, as is the case with NCLB. One way to avoid this paradox is to report subdomain scores multidimensionally only at the group level using an EIRM approach, as I describe next.

Explanatory Measurement: The EIRM Approach

Note that in the two-step approach the measurement model only plays a role in the first step of the analysis, in which an outcome variable has been generated for use in the second stage. By the second step, we are formally addressing each of our three research questions using a linear regression, with the usual assumptions implied by such a model. If the latter is our principal aim, then it might be easy to sweep much of what has occurred during the first step under the rug. By making the two-steps concurrent within the GLMM framework, the analyst has the opportunity to tie what is learned about group differences directly back to the underlying measurement model. All of the three research questions addressed above using the two-step approach can also be addressed in a single stage by taking an EIRM approach. The key advantage of this approach is that it provides a framework for both the psychometric and statistical analyses that may be of interest. In the context of answering research questions about group differences, the fundamental distinction between specifying an EIRM relative to the traditional item response model used in the first step of the two-step approach is the incorporation of person covariates directly into the measurement model. Because differences in achievement are parameterized directly at the group (i.e., population) level, we are able to get disattenuated measures of group differences that will differ from the ones found under the two-step approach. To answer questions 1 and 2, a unidimensional Rasch Model with two different sets of person level covariates is specified. To answer question 3, a multidimensional Rasch Model with two different sets of person level covariates is specified. In what follows I contrast the results from answering research questions 1-3 using the two-step approach with the results from using an EIRM approach.

Insert Tables 6 and 7 about here

The results presented in Tables 6 and 7 parallel those presented in Tables 4 and 5. A key distinction is that the outcome variables reflected in Tables 6 and 7 are not estimated EAPs, but individual item responses. The contextual variables used in the regressions remain the same, and the coefficients are expressed as before both in logit units, and in SD units. To accomplish the latter, the unadjusted coefficients are divided by the population SD from an unconditional form of the item response model. So for example, when a unidimensional Rasch Model is specified with no person covariates (i.e., the unconditional item response model), the population SD, $\sqrt{\text{var}(\theta)}$, = .81. It follows that when expressed as an effect size, the regression coefficient for the variable

BLACK for Model 1 in Table 6 will be $-.77/.81=-.95$. The same sort of approach is taken to compute effect sizes for the regression coefficient corresponding to the multidimensional Rasch Model.

How do the results from the EIRM approach, presented in Table 6 and 7, compare to those from the two-step approach?

1. The raw regression coefficients from the EIRM latent regressions tend to be larger in absolute value than those from linear regressions. The reason for this is that in the two-step approach, the regression coefficients represent conditional EAPs for subgroups of students, while in the EIRM approach the coefficients represent the conditional means of the supgroup populations. In the two-step approach, before they are regressed on racial/ethnic dummy variables, the EAPs have already been shrunken toward the overall population mean. The larger the measurement error associated with student MC responses, the more that individual EAP estimates of science achievement shrink to the overall population mean. Hence, when a two-step approach is taken to represent and analyze racial/ethnic group differences, the regression coefficients are effectively attenuated by measurement error.
2. Because the use of EAPs shrinks student scale scores towards the population mean, it follows that the SD of these EAPs will be smaller than the SD of the population as a whole. This can be seen when comparing model results across Tables 4 and 6: the SD of unidimensional EAPs is .70 logits (Table 4), while the population SD is .81 logits (Table 6). When these SDs are used to compare the regression coefficients in effect size units, the magnitude of racial/ethnic gaps for

black and Hispanic students in the EIRM approach is a about 1/10 of an SD higher.

3. The inclusion of contextual variables in the regressions in either approach has the same *relative* impact on racial/ethnic gaps estimates when science achievement is modeled unidimensionally. The size of the gap stays the same for black students, decreases by about .11 to .12 SDs for Hispanic students, and reverses for Asian students. This is good news, because it means that in a relative sense, if we were interested in the effect of some planned intervention intended to reduce racial/ethnic gaps, both the two-step and the EIRM approach would be expected to yield the same answer.
4. There are important substantive differences between the regression coefficients expressed in SD units for the two-step and EIRM approach when science achievement is modeled multidimensionally. Under the two-step approach the patterns in estimated racial/ethnic gaps showed no change across the five subdomains of science proficiency. Under the EIRM approach with no contextual variables, the respective size of the gaps for black and Hispanic students varies by subdomain relative to the unidimensional composite, from a high of 1.29 SDs (black students on the physical science subdomain) to a low of .44 SDs (Hispanic students on the science and technology subdomain).
5. There are also differences in the extent to which gaps are reduced when contextual variables are included in the two approaches. Under the EIRM approach, the Hispanic gap decreases by .28 SDs for the earth science subdomain, and .20 for scientific inquiry subdomain, stays about the same for the physical

> science and science and technology subdomains, and actually increases by .14 SDs for the life science subdomain. This is in contrast to the results from the linear regression approach, where the inclusion of contextual variables consistently decreased the Hispanic gap by about .12 to .14 SDs for each subdomain.

The difference in answers about group differences under the two approaches is a function of the precision with which student achievement is being measured. When test scores are highly reliable, there will be little difference between the two-step and EIRM approaches, as the effect of attenuation due to measurement error is negligible. However, when test scores have low to moderate reliability—and this is likely to be the case when scores are reported in terms of subdomains defined with a small number of items—the effect of attenuation may lead to substantive differences in interpretation. In such scenarios taking the EIRM approach is clearly preferable when an evaluation of group differences is of interest.

Another useful aspect of the EIRM approach as implemented here is that it not only produces information about salient group differences in science achievement, but can also place these differences in a criterion-referenced context. That is, when a Rasch-family model has been specified, the EIRM approach allows for the creation of a variable map that relates group differences to test items together along a unidimensional continuum (or multidimensional continua) of measurement. This is illustrated by the variable map in Figure 2, for the scenario in Table 6 when the PASS test is modeled unidimensionally and includes a latent regression with racial/ethnic dummy variables (Model 1). In this map, the difficulty of items and achievement of respondents are both expressed in terms of logits, and horizontal lines are used to demarcate the location of racial/ethnic population averages. Differences in achievement can be now be interpreted not just in terms of the logit scale, but in also in terms of the probability of responding correctly to specific subsets of items on the test. In this case, we can see that white and Asian students have at least a 50% probability of responding correctly to items 4, 7, 10, 12 and 20, while Hispanic and Black students have a significantly lower probability of answering these items correctly. On the other hand, items at the low end of the map (i.e., 1-3, 5, 9, 15, 23-24) are ones that are likely to be answered correctly by white, Asian, black and Hispanic students alike. This sort of item-specific information might prove quite useful diagnostically, depending upon how well the content of these items can be aligned with the science curricula from which students learn.

Insert Figure 2 about here

To recap, the general story in this analysis has been that black and Hispanic students do not do as well on the PASS test items as their white and Asian counterparts. This general story is captured equally well under both the two-step and EIRM approaches. However, the devil is often in the details, and in this, the EIRM approach has some clear advantages. Like the two-step approach, the EIRM approach posits a measurement model to account for within-student differences it item responses. But unlike the two-step approach, in the EIRM approach the key research questions of interest—all at the group level—can be appropriately parameterized in terms of between-student (i.e., group) differences within the same model. This allows for more nuance in our specific conclusions about racial/ethnic achievement gaps. One of these nuances was illustrated through the use of a variable map in Figure 2. As another example, the results under the EIRM approach shown in Table 7 indicate that gaps for Hispanic students are smallest in the process-oriented subdomains of scientific inquiry and science and technology. This might constitute some encouraging news at schools where the science curriculum has been transformed towards a greater emphasis in these areas.

DISCUSSION

In this paper I have presented and illustrated the concept of an explanatory item response model. Both item response models and explanatory item response models have been shown to be special cases within the broader statistical frameworks of generalized linear mixed models and nonlinear mixed models. These frameworks help clarify the inherent multilevel structure of item response models, and offer considerable flexibility in specifying research questions at the appropriate level of interest. Using data from a sample of 10th grade students taking a standardized science test, I have shown that when research questions are posed in terms of group differences, the answers from taking an EIRM approach will differ from those that result from taking a two-step approach, because the latter will be attenuated by measurement error.

What has been illustrated is just one example of an EIRM. Depending upon our research questions and the variables that had been collected, other types of EIRMs could have been specified:

- Item covariates could be included to explain differences in item difficulty. For example, perhaps items with charts and graphs are more difficult for students to answer than items with just text. This would be an example of a linear logistic test model (c.f., Janssen, Schepers & Peres, 2004).
- Person-by-item covariates could be included to determine if item difficulty changes as a function of group differences. This would be an example of a model used to analyze Differential Item Functioning (DIF; c.f., Meulders & Xie, 2004). In addition, person or item covariates could be added in a subsequent attempt to explain DIF. This would constitute the item side analog to the illustration provided in this paper.
- A third level, school, could be added to the model, along with school-level covariates. This would be an example of a three level model with latent regression (c.f., Van den Noortgate & Paek, 2004).

These are not new models, but they are all examples of models that could be specified within the general GLMM framework captured by the expression (1) or (2), and estimated within a single statistical environment (i.e., SAS or Stata).

There are some limitations to the EIRM approach. First, it can be quite challenging to properly specify and estimate item response models, even without adding an explanatory component. One could only reasonably expect researchers with suitable training in psychometrics to specify, estimate and interpret an EIRM. Second, in many cases the data is simply not available at the item level. In most cases where something like the two-step approach is used to answer research questions about group differences with a large-scale standardized test, the researcher has not been given access to student item responses. Finally, the EIRM approach stops short of introducing latent variables as person, item, or person-by-item covariates. This is a key distinguishing feature between EIRMs and structural equation models.

In educational research it is seldom the case that our questions hinge solely upon the quantification of student (or item) level measures for some latent construct. Rather, we inevitably wish to examine and explain group level differences among these measures. When this is the case, taking the EIRM approach illustrated here and described in greater detail in De Boeck and Wilson (2004) may be advantageous. When this is the case and the EIRM approach is not taken, it is important for researchers to be aware of how the results of their investigations might be affected.

References

Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education,* 7, 255-278.

Ackerman, T., Gierl, M. J., and Walker, C. M. (2003) Using multidimensional IRT to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, Fall 2003, 37-53.

Adam, R. (2006) Reliability as a measurement design effect. Paper presented at the 2006 Bi-annual meeting of the Institute for Objective Measurement Workshop, Berkeley, CA.

Adams, R., Wilson, M., and Wu, M. (1997). Multilevel item response models: an approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22(1): 46-75.

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

American Association for the Advancement of Science. (1993). *Benchmarks for scientific literacy*. New York: Oxford University Press.

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

Briggs, D. C. & Wilson, M. (2003) An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4(1), 87-100.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443-459.

Camara, W. J. and Schmidt, A. E. (1999). Group differences in standardized testing and social stratification. New York, The College Board**:** 1-18.

Campbell, J. R., Hombo, C. M., Mazzeo, J. (2000). NAEP 1999 trends in academic progress: three decades of student performance. Washington, DC, U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Davidian, M. and Gilitinan, D. M. (1995) *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.

De Boeck, P. and Wilson, M., eds. (2004) *Explanatory item response models: a generalized linear and nonlinear approach*. New York: Springer.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis, 2nd Edition*. London: Chapman & Hall/CRC Press.

Green, B., Bock, R., Humphreys. L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Jannsen, R., Schepers. J., Peres, D. (2004) Models with item and item group predictors. In: *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, P. De Boeck and M. Wilson, eds., New York, Springer: 189-212.

Jencks, C. & Phillips, M. (1998). *The black-white test score gap*. Washington, D.C., Brookings Institution Press.

Kamata, A. (2001) Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38(1), 79-93.

Lee. J. (2002) Racial and ethnic achievement gap trends: reversing the progress toward equity? *Educational Researcher*, 31(1), 3-12.

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

McCulloch, C.E. & Searle, S.R. (2001) *Generalized, linear, and mixed models*. New York: Wiley.

Meulder, M., & Xie, Y. (2004) Person-by-item predictors. In: *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, P. De Boeck and M. Wilson, eds., New York, Springer: 214-40.

Mislevy, R. J., Beaton, A. E., Kaplan., B., & Sheehan, K. M. (1992) Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29, 133-161.

Molenberghs, G. & Verbeke, G. (2004) An introduction to (generalized (non)linear mixed models in: *Explanatory item response models: a generalized linear and nonlinear approach*, P. De Boeck and M. Wilson, eds. New York, Springer: 111-153.

Monfils, L. Dawber, T., Han, N, & Henderson-Montrero, D. (2006) Supporting reform efforts through diagnostic subscore reports: implications for schools. Paper presented at the 2006 annual meeting of the National Council for Measurement in Education. San Francisco, CA.

Nandakumar, R & Stout, W. (1993) Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics,* 18, 41-68.

National Research Council. (1996). *National Science Education Standards*. Washington, DC, National Academy Press**:** 262.

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

Rijmen, F. & Briggs, D. C. (2004) Multiple Person Dimensions and Latent Item Predictors. In: *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, P. De Boeck and M. Wilson, eds., New York, Springer: 247-65.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8(2), 185-205.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling*. London: Chapman & Hall/CRC.

Smith, J. P. (1995). Racial and ethnic differences in wealth in the Health and Retirement Study. *Journal of Human Resources*, Supplement.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2004). WinBUGS 1.4. http://www.mrc-bsu.cam.ac.uk/bugs

Stout, W. A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617.

Stout, W., Froelich, A. G., & Gao, F. (2001) Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), Essays on item response theory (357-376). New York: Springer.

Thissen, D., Chen, W.-H, & Bock, R. D. (2002). *MULTILOG 7*. Lincolnwood, IL: Scientific Software.

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noorgate, W., Meulders, M, and De Boeck, P. (2004) Estimation and Software. In: *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, P. De Boeck and M. Wilson, eds., Springer.

U.S. Department of Education, National Center for Education Statistics (2000). Digest of Education Statistics, 1999. Washington, DC.

Van den Noortgate, W. and Paek, I. (2004) Person regression models. In: *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, P. De Boeck and M. Wilson, eds., New York, Springer: 167-85.

Verhlest, N. & Verstralen, H. (2001). An IRT model for multiple raters. In A. Boomsma, M van Duijn, and T. Snijders (Eds.), *Essays on Item Response Theory*, 88-108. New York: Springer-Verlag.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B. III, Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (2001). Augmented Scores – "Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, W.-C., Wilson, M., & Adams, R. (1997). Rasch models for multidimensionality between items and within items. In G. Englehard, Wilson, Mark (Ed.), *Objective Measurement* (Vol. 4,): Ablex Publishing.

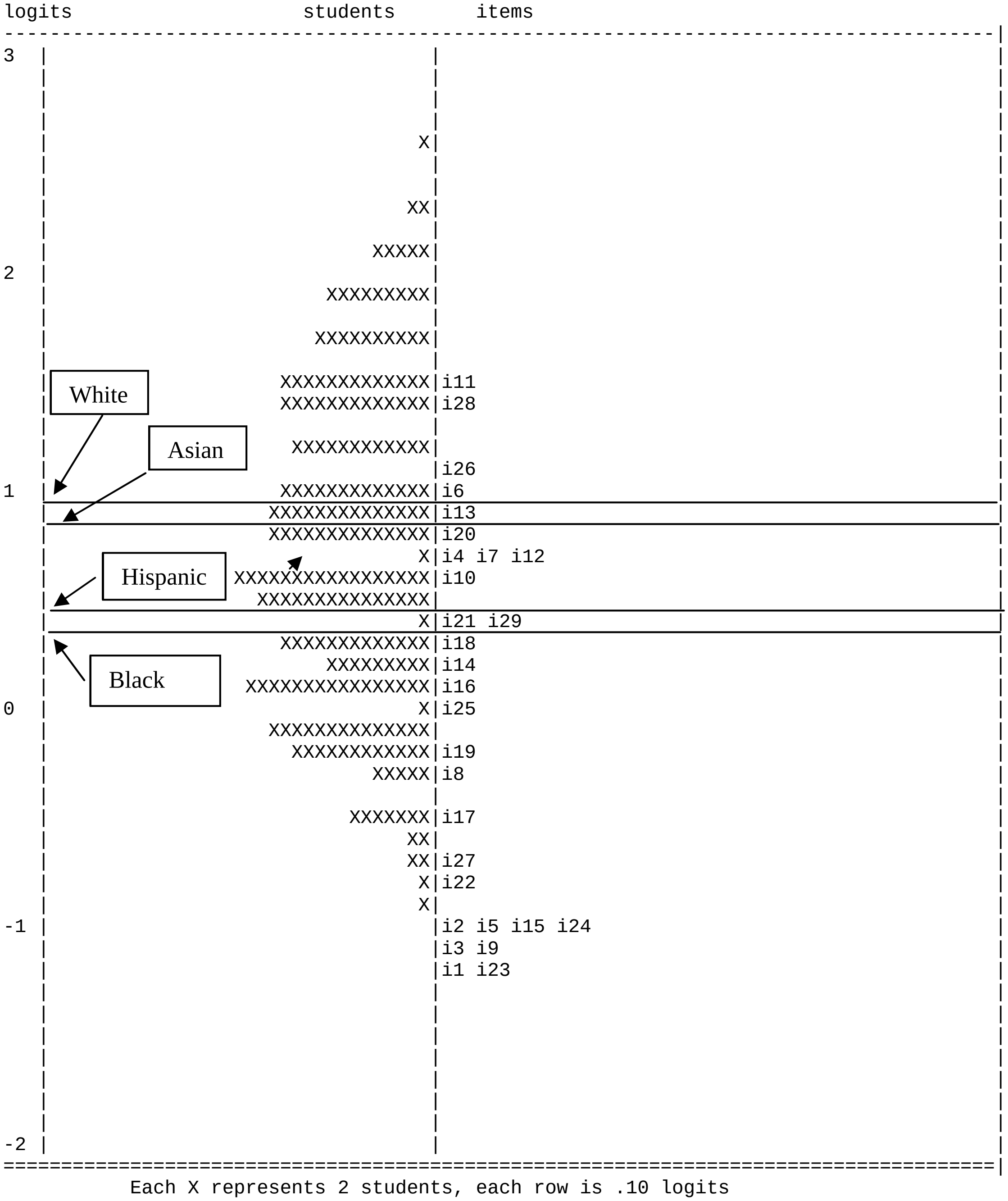
WestEd (2006). Partnership for the Assessment of Standards-based Science: FAQs and Sample Assessment Items. San Francisco, CA. http://www.wested.org/cs/we/view/rs/612 Retrieved from website June 27, 2006.

Wu, M. L., Adams, Raymond J., Wilson, Mark R. (1998). ACER Conquest. Melbourne, Australia, Australian Council for Educational Research.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (1995). *BILOG-MG: Multiple-Group Item Analysis and Test Scoring*. Chicago, IL: Scientific Software.

Obs	n	i	Y_{ni}	$x_{ni(1)}$	$x_{ni(2)}$	$x_{ni(3)}$	$x_{ni(4)}$	$x_{ni(5)}$	$z_{ni(1)}$
1	1	1	1	1	0	0	0	0	1
2	1	2	1	0	1	0	0	0	1
3	1	3	0	0	0	1	0	0	1
4	1	4	1	0	0	0	1	0	1
5	1	5	0	0	0	0	0	1	1
6	2	1	1	1	0	0	0	0	1
7	2	2	0	0	1	0	0	0	1
8	2	3	0	0	0	1	0	0	1
9	2	4	1	0	0	0	1	0	1
10	2	5	1	0	0	0	0	1	1

Figure 1. Data Structure for Rasch Model Specification using GLMM Framework



```
logits                     students       items
---------------------------------------------------------------------------------------|
3   |                                |                                                 |
    |                                |                                                 |
    |                                |                                                 |
    |                                |                                                 |
    |                               X|                                                 |
    |                                |                                                 |
    |                                |                                                 |
    |                              XX|                                                 |
    |                                |                                                 |
    |                           XXXXX|                                                 |
2   |                                |                                                 |
    |                       XXXXXXXXX|                                                 |
    |                                |                                                 |
    |                      XXXXXXXXXX|                                                 |
    |                                |                                                 |
    |                   XXXXXXXXXXXXX|i11                                              |
    |                   XXXXXXXXXXXXX|i28                                              |
    |                                |                                                 |
    |                    XXXXXXXXXXXX|                                                 |
    |                                |i26                                              |
1   |                   XXXXXXXXXXXXX|i6                                               |
    |                  XXXXXXXXXXXXXX|i13                                              |
    |                  XXXXXXXXXXXXXX|i20                                              |
    |                               X|i4 i7 i12                                        |
    |               XXXXXXXXXXXXXXXXX|i10                                              |
    |                 XXXXXXXXXXXXXXX|                                                 |
    |                               X|i21 i29                                          |
    |                   XXXXXXXXXXXXX|i18                                              |
    |                       XXXXXXXXX|i14                                              |
    |                XXXXXXXXXXXXXXXX|i16                                              |
0   |                               X|i25                                              |
    |                  XXXXXXXXXXXXXX|                                                 |
    |                    XXXXXXXXXXXX|i19                                              |
    |                           XXXXX|i8                                               |
    |                                |                                                 |
    |                         XXXXXXX|i17                                              |
    |                              XX|                                                 |
    |                              XX|i27                                              |
    |                               X|i22                                              |
    |                               X|                                                 |
-1  |                                |i2 i5 i15 i24                                    |
    |                                |i3 i9                                            |
    |                                |i1 i23                                           |
    |                                |                                                 |
    |                                |                                                 |
    |                                |                                                 |
    |                                |                                                 |
    |                                |                                                 |
    |                                |                                                 |
    |                                |                                                 |
-2  |                                |                                                 |
=======================================================================================|
           Each X represents 2 students, each row is .10 logits
```

Figure 2. Variable Map of PASS Multiple Choice Section, Form 4.

Demographic Characteristics	Percent
Gender	
Male	43.4
Female	53.3
Missing	3.2
Race/Ethnicity	
African American	9.7
Asian American	9.9
Hispanic	20.3
White	53.8
Other	2.5
Missing	3.7
English is primary language spoken at home	
Yes	78.5
No	19.9
Missing	1.6
Sample Size	433

Table 1. Demographic Characteristics of 10th Grade PASS Sample

Contextual Variables from PASS Data	Percent
Compared to other subjects my grades in science are (GRADES)	
Lower	14.8
The same	60.7
Higher	21.5
MISSING	3.0
Compared to other subjects, I like science (SCI_ATT)	
Less	25.6
The same	44.1
More	27.9
MISSING	2.3
How much time do you usually spend outside of school doing science homework each week? (HOMEWORK)	
None	17.3
Less than 1 hour	42.7
Between 1 and 2 hours	28.4
More than 2 hours	9.5
MISSING	2.1
Are there things that you learn in science that are useful to you when you’re not in school? (USEFUL)	
Yes	20.8
Sometimes	61.9
Never	14.3
MISSING	3.0
Do you think that knowing and understanding science will be useful when you grow up? (FUTURE)	
Yes	46.0
Sometimes	42.5
Never	9.0
MISSING	2.5
Do you talk to your parent(s) or guardian about what you do in science class? (PARENTS)	
Yes	13.2
Sometimes	43.4
Never	40.0
MISSING	3.5
Sample Size	433

Table 2. Contextual Variables 10th Grade PASS Sample

Dimension	ES	LS	PS	SI
LS	0.848			
PS	0.926	0.918		
SI	0.827	0.908	0.931	
ST	0.892	0.684	0.859	0.743

Marginal Reliability	ES	LS	PS	SI	ST
Unidimensional	0.545	0.241	0.434	0.517	0.319
Multidimensional	0.723	0.673	0.749	0.708	0.645

Note: ES = Earth Science, LS = Life Science, PS = Physical Science
SI =Scientific Inquiry, ST = Science & Technology

Table 3. Dimensional Correlations and Marginal Reliabilities of Science Subdomains

	Model 1		Model 2	
	logit units	SD units	logit units	SD units
Intercept	0.90**	1.29	0.41**	0.58
Black	-0.61**	-0.87	-0.60**	-0.86
Hispanic	-0.52**	-0.75	-0.44**	-0.64
Asian	-0.10	-0.14	0.12	0.17
Other	-0.30	-0.43	-0.38	-0.54
ESL_Hisp			-0.02	-0.04
ESL_Asian			-0.61**	-0.87
Grades			0.24**	0.35
Homework			0.15**	0.21
Interest			0.01	0.02
SD of Outcome Variable	0.70		0.70	
Sample Size	420		411	
Variance Explained	0.12		0.19	

Note: ** = *p* < .01

Table 4. Linear Regressions from Two-step Approach: Unidimensional

Logit Units	ES 1	ES 2	LS 1	LS 2	PS 1	PS 2	SI 1	SI 2	ST 1	ST 2
Intercept	1.29**	0.56**	0.43**	0.12	0.91**	0.43**	1.29**	0.65**	0.53**	-0.14**
Black	-0.77**	-0.76**	-0.42**	-0.42**	-0.58**	-0.57**	-0.84**	-0.84**	-0.63**	-0.62**
Hispanic	-0.67**	-0.53**	-0.36**	-0.30**	-0.48**	-0.39**	-0.66**	-0.53**	-0.53**	-0.42**
Asian	-0.09	0.20	-0.07	0.09	-0.08	0.14	-0.14	0.20	-0.05	0.19
Other	-0.42	-0.53	-0.22	-0.27	-0.31	-0.38	-0.45	-0.56	-0.34	-0.43
ESL_Hisp		-0.10		-0.04		-0.05		-0.09		-0.05
ESL_Asian		-0.81**		-0.46**		-0.62**		-0.95**		-0.63**
Grades		0.35**		0.16**		0.23**		0.31**		0.31**
Homework		0.21**		0.10*		0.15**		0.21**		0.17**
Interest		0.03		0.00		0.01		0.01		0.04
Variance Explained	0.110	0.193	0.120	0.197	0.113	0.197	0.109	0.191	0.096	0.175
SD of Outcome Variable	0.926		0.481		0.661		0.953		0.793	

SD Units	ES 1	ES 2	LS 1	LS 2	PS 1	PS 2	SI 1	SI 2	ST 1	ST 2
Intercept	1.39	0.60	0.89	0.24	1.38	0.64	1.36	0.69	0.67	-0.17
Black	-0.83	-0.82	-0.87	-0.87	-0.87	-0.86	-0.88	-0.88	-0.80	-0.78
Hispanic	-0.72	-0.57	-0.75	-0.62	-0.72	-0.59	-0.70	-0.56	-0.66	-0.52
Asian	-0.10	0.22	-0.15	0.19	-0.12	0.21	-0.15	0.21	-0.06	0.23
Other	-0.46	-0.57	-0.45	-0.56	-0.46	-0.58	-0.47	-0.59	-0.43	-0.54
ESL_Hisp		-0.10		-0.08		-0.07		-0.09		-0.07
ESL_Asian		-0.87		-0.97		-0.93		-1.00		-0.80
Grades		0.37		0.32		0.35		0.33		0.39
Homework		0.22		0.22		0.23		0.22		0.22
Interest		0.03		0.01		0.02		0.01		0.05

Note:* = p < .05, ** = p < .01 Subdomains of PASS Test are ES = Earth Science, LS = Life Science, PS = Physical Science, SI =Scientific Inquiry, ST = Science & Technology

Table 5. Linear Regressions from Two-step Approach: Multidimensional

	Model 1		Model 2	
	logit units	SD units	logit units	SD units
Intercept	0.96**	1.19	0.35**	0.44
Black	-0.77**	-0.95	-0.77**	-0.95
Hispanic	-0.67**	-0.83	-0.58**	-0.71
Asian	-0.115	-0.14	0.176	0.22
Other	-0.373	-0.46	-0.48**	-0.59
ESL_Hisp			-0.025	-0.03
ESL_Asian			-0.82**	-1.02
Grades			0.30**	0.37
Homework			0.20**	0.25
Interest			0.01	0.01
Population SD (Unconditional)	0.81		0.81	
Sample Size	433		433	
Number of Parameters	34		39	
Deviance	13756.28		13577.40	

Note: ** = *p* < .01

Table 6. Latent Regressions from EIRM Approach: Unidimensional

Logit units	ES 1	ES 2	LS 1	LS 2	PS 1	PS 2	SI 1	SI 2	ST 1	ST 2
Intercept	1.37**	0.26	0.50**	0.53**	1.01**	0.26*	1.33**	0.71**	0.49**	-0.66**
Black	-0.93**	-0.88**	-0.38**	-0.37**	-0.97**	-0.10*	-1.04**	-1.10**	-0.56**	-0.53**
Hispanic	-0.93**	-0.62**	-0.61**	-0.70**	-0.71**	-0.68**	-0.67**	-0.45**	-0.43**	-0.39**
Asian	-0.08	0.40	-0.25	-0.07	-0.04	0.13	-0.30	0.41	0.08	0.31
Other	-0.67	-0.85	0.02	-0.01	-0.50	-0.63	-0.66	-0.82	-0.08	-0.18
ESL_Hisp		-0.33		0.14		0.20		-0.29		0.20
ESL_Asian		-1.12**		-0.38*		-0.74**		-1.65**		-0.42**
Grades		0.50**		0.05		0.32**		0.33**		0.50**
Homework		0.30**		0.04		0.29**		0.27**		0.15*
Interest		0.06		-0.04		0.01		-0.03		0.15**

Population SD (unconditional)	1.07	1.07	0.58	0.58	0.75	0.75	1.12	1.12	0.98	0.98

SD units	ES 1	ES 2	LS 1	LS 2	PS 1	PS 2	SI 1	SI 2	ST 1	ST 2
Intercept	1.28	0.24	0.86	0.91	1.34	0.34	1.18	0.64	0.50	-0.68
Black	-0.86	-0.82	-0.66	-0.65	-1.29	-1.32	-0.93	-0.98	-0.57	-0.55
Hispanic	-0.86	-0.58	-1.06	-1.20	-0.94	-0.90	-0.60	-0.40	-0.44	-0.40
Asian	-0.08	0.37	-0.44	-0.13	-0.06	0.17	-0.27	0.37	0.08	0.32
Other	-0.63	-0.79	0.04	-0.01	-0.66	-0.83	-0.59	-0.73	-0.08	-0.19
ESL_Hisp		-0.31		0.24		0.26		-0.26		0.21
ESL_Asian		-1.04		-0.66		-0.98		-1.47		-0.42
Grades		0.46		0.09		0.43		0.29		0.51
Homework		0.28		0.06		0.39		0.24		0.15
Interest		0.05		-0.07		0.01		-0.03		0.15

Note:* = p < .05, ** = p < .01 Subdomains of PASS Test are ES = Earth Science, LS = Life Science,
PS = Physical Science, SI =Scientific Inquiry, ST = Science & Technology

Table 7. Latent Regressions from EIRM Approach: Multidimensional

[1] Other distributions could also be chosen. For example, Adams, Wilson & Wang (1997) provide an example where a step distribution is specified.

[2] It would also be possible to generate maximum likelihood (ML) or maximum a posteriori (MAP) estimates of ability. While all three estimates are usually very strongly correlated, ML estimates will typically differ from EAP and MAP estimates at extreme values because EAP and MAP estimates are shrunken toward the mean of the population distribution. As a consequence of this, the disadvantage of EAP estimates is that they are biased. On the other hand, the advantage of the EAP estimate relative to ML estimates it minimizes mean square error. There are primarily computational reasons for preferring an EAP estimate to a MAP estimate. For more on this issue, see Embretson & Reise, 2000, pp. 159-179 and Thissen & Olrlando, 2001, pp. 98-114.

[3] There are five dummy variables representing the self-reported racial/ethnic categories Asian-American (Asian), African-American (black), Hispanic, white and other (where other = American Indian/Alaskan, Pacific Islander and Filipino). These categorizations are in no way scientific, but are used by convention. It is clearly debatable whether students are appropriately classified into these groups, particularly for students who are from mixed racial/ethnic backgrounds. Further, there is potentially great variation within groups. For example, a student labeled Hispanic may be of Spanish, Brazilian, Mexican or Puerto Rican descent. These types of issues are outside the scope of this paper, but raise clear caveats about how racial/ethnic achievement gaps should be interpreted, regardless of how they are measured or explained.

[4] For example, two symposia were devoted to the topic at the 2006 annual meeting of the National Council on Measurement in Education in San Francisco, CA.

[5] This constitutes an example of what Wang, Wilson & Adams (1997) have termed between-item multidimensionality, and what Ackerman, Gierl & Walker (2003) have termed "simple structure." For details on multidimensional item response models, c.f., Rijmen & Briggs, 2004; Ackerman, Gierl & Walker, 2003; Ackerman, 1994; Reckase & McKinley, 1991.