Student and School factors Associated with Aberrant Response Patterns on a Large Scale Assessment

Amin Mousavi
Ying Cui

Centre for Research in Applied Measurement and Evaluation
University of Alberta

Poster Presented at the 2014 annual meeting of American Educational Research Association
Philadelphia, Pennsylvania, USA
April 3-7
Introduction

Large-scale assessment is an important tool for evaluating student achievement and monitoring the quality of educational systems. Often important decisions regarding accountability and placement are made on the basis of test scores generated from large-scale assessments. Therefore, it is important to evaluate the validity of the inferences derived from test results, which depends critically on the measurement model used in the design, analysis and scoring of large-scale assessments. Invalid score inferences will be made if the measurement model fails to reflect accurately the real aspects of student responding processes.

Attempts to assess the fit of a student’s responses to the measurement model have led researchers to studies of "person-fit" statistics. These studies have centered on evaluating how well a student’s item response vector matches the expected response patterns derived from a given model. For example, if a student produces correct answers to the more difficult items but fails to answer the easier ones successfully, the student’s responses are considered as “unexpected”, “aberrant”, or “misfitting”. The detection of misfitting responses has been investigated by many researchers during the last few decades. The rationale for this effort is that many problems may arise from the existence of misfitting responses on a test. For example, Emons, Meijer, and Sijtsma (2002) suggested that misfitting responses may serve as an indication that the student’s response behavior may have been influenced by factors that are not intended to be measured by the test. Nering and Meijer (1998) indicated that the estimated proficiency level of examinees with misfitting responses may be invalid and misleading. Nering (1998) suggested that the existence of misfitting responses might also affect the estimation of the item parameters (e.g., item difficulties) of the items in the test. Meijer (1997) and Schmitt,
Cortina, and Whitney (1993) suggested that the validity and reliability of a test might be compromised because of the existence of misfitting responses in test data.

Many researchers have hypothesized the possible reasons for misfitting responses (e.g., Meijer, 1997; Meijer & Sijtsma, 2001; Stricker & Emmerich, 1999; Zickar & Robie, 1999). Student characteristics such as gender, ethnicity, language deficiencies, anxiety, low motivation, misconceptions, and deficiency of sub-abilities have been proposed as being possibly associated with misfit. Test-taking strategies such as guessing, cheating, "plodding" (working too carefully and consequently too slowly), the use of test-wiseness, and alternative knowledge and skills have also been discussed as possible sources of misfit. In addition, external factors such as fatigue, noise, faulty items, response alignment errors, and atypical schooling may also influence student responses.

Studies related to person fit fall into two categories. One category of studies focuses on the development of person fit statistics to quantitatively identify misfitting responses and the investigation of the power of these statistics using simulated data. In educational and psychological measurement literature, many person fit statistics have been proposed to help identify observed response patterns that are incongruent with the measurement model used in test design and analysis. For example, Meijer and Sijtsma (2001) reviewed over 40 person fit statistics. Karabatsos (2003) conducted a simulation study that compared the performances of 36 person fit statistics under different testing conditions.

The other category of person fit studies applies person fit statistics to empirical data to explore factors that may be associated with misfitting responses or to examine the impact of misfitting responses on the psychometric properties of tests. As pointed out by many researchers (e.g., Bracey & Rudner, 1992; Meijer, 2003; Meijer & Sijtsma, 2001), not enough person fit
research has been conducted using empirical data. A few studies have been conducted to examine the potential sources of person misfit with educational data from large-scale assessments and findings were inconsistent across studies. For example, Rudner, Bracey and Skaggs (1996) analyzed data from the 1990 National Assessment of Educational Progress (NAEP), and found no association between person misfit and different background variables. The authors concluded that person fit statistics had little to offer in the analysis and reporting of NAEP. In comparison, Lampriaou and Boyle (2004) evaluated the measurement accuracy using Mathematics National Curriculum test data (England) for ethnic minority students and students speaking English as a second language, and they found these students were more likely to produce misfitting responses. Petridou and Williams (2007) used the hierarchical linear modelling (HLM) to account for aberrant response patterns on a mathematical assessment and found significant proportions of person misfit attributable to person- and class-level variables. For example, Petridou and Williams (2007) found that all person-level variables being equal, students in a high-ability class are more likely to produce misfitting responses.

Although the use of person fit statistics is promising for improving measurement practice, person-fit statistics are currently studied largely by researchers and have not been used routinely in the analysis of student response data in practice. The lack of practical uses of person fit statistics is likely due to the fact that results of these statistics do not provide clear indications of how misfits occur or what types of misfitting response behavior underlie test performance. Due to the paucity of research in this area and the inconsistency of existing findings, it remains unclear whether person fit analysis is indeed useful for large-scale assessment and, if yes, what student and school factors are associated with aberrant response patterns on a large-scale assessment.
In this study, we applied one of the most studied person fit statistics, $l_z$, to data from the Trends in Mathematics and Science Study (TIMSS 2007) to further explore student and classroom factors that influence person fit measures. TIMSS was chosen in this study as it collects a wide range of contextual information from students, teachers and schools about background factors, and attitudes and beliefs about learning, which have the potential to provide insights into the possible types and sources of person misfit.

Method

Data

Data from TIMSS 2007 were used in the current study. TIMSS is an international large-scale assessment designed to inform educational policy and practice by assessing students at the fourth and eighth grades every four years to provide information about student educational context and academic achievements on Mathematics and Science. Fifty-nine countries and eight benchmarking entities participated in TIMSS 2007. A sample of students from each country and benchmarking entity was chosen according to a two-stage sampling design. The data used in the present study consisted of Canadian grade eight student responses to Mathematics items ($n=11660$, girls=5846, boys=5814). The data were retrieved from TIMSS website.

TIMSS 2007 Grade Eight Mathematics includes both dichotomous and polytomous items. For dichotomous items, the three-parameter item response theory (IRT) model was used for calibrating the parameters of multiple-choice items, and the two-parameter model for constructed-response items. A generalized partial credit model (Muraki, 1992) was used with polytomous constructed-response items. The item parameters for each model were estimated independently of the parameters of other models using the Parscale software. Regarding the student IRT proficiency scores, TIMSS produced five plausible values for each student (see von...
Davier, Gonzalez & Mislevy, 2009, for theoretical justification of the use of plausible values),
instead of generating a single point estimate of $\theta$. Each plausible value was randomly drawn
from the corresponding estimated ability distribution for the observed response pattern.

**Evaluating Person Fit**

For each student’s responses, the $l_2$ statistic developed by Drasgow, Levine, and
Williams (1985) was used to indicate the degree of person misfit. $l_2$ is one of the most studied
person fit statistics in the literature. It is a standardized version of the $l_0$ index (Levine & Rubin,
1979), which is simply the log-likelihood of an observed item response pattern calculated based
on an IRT model. $l_0$ is given by

$$l_{0i} = \ln \left\{ \prod_{j=1}^{J} P_j(\theta_i)^{x_{ij}} \left[ 1 - P_j(\theta_i) \right]^{1-x_{ij}} \right\}, \quad (1)$$

where $x_{ij}$ is the binary (0, 1) response to item $j$ ($j = 1, 2, \ldots, J$) by student $i$, $\theta_i$ is the latent trait
of student $i$, and $P_j(\theta_i)$ is the probability of correctly answering item $j$ computed based on an
IRT model. A low value of $l_{0i}$ suggests that the probability of obtaining the response pattern
produced by student $i$ is small given the hypothesized IRT model and therefore the response
pattern can be considered as a misfit of the IRT model. As pointed out by Drasgow et al (1985),
an undesirable property of $l_0$ lies in the fact that $l_0$ is conditional on the latent trait, $\theta$, which
suggests that the classification of an observed response vector as fitting or misfitting is
influenced by $\theta$. Therefore, the standardized normally distributed statistic $l_z$ is written as

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}}, \quad (2)$$

where
\[ E(l_0) = \sum_{j=1}^{J} \{ P_j(\theta) \ln[P_j(\theta)] + [1 - P_j(\theta)] \ln[1 - P_j(\theta)] \}, \]  

(3)

and

\[ Var(l_0) = \sum_{j=1}^{J} P_j(\theta)[1 - P_j(\theta)] \left[ \ln \frac{P_j(\theta)}{1 - P_j(\theta)} \right]^2. \]  

(4)

Large negative values of \( l_z \) suggest potential misfit. One can dichotomize the \( l_z \) values into two categories (fitting or misfitting) by the use a critical value from the standardized normal distribution. It should be noted, however, the asymptotic distribution of \( l_z \) is not standard normal anymore when the estimated ability parameters (instead of the true parameters) are used (Molenaar & Hoijtink, 1990; Nering, 1995, 1997; Reise, 1995). In addition, the above equations can be used to calculate \( l_z \) for student responses to dichotomous items only. Drasgow et al. (1985) also introduced a general polytomous model for computing \( l_z \), which was employed in this study given that TIMSS 2007 included polytomous items as well.

**Variable Specification for the HLM Analysis**

Given the multilevel structure of the TIMSS data, HLM was used to examine the potential factors at different levels that were associated with person misfit. With students being nested within classes, two levels – students and classes – were considered in the analysis. As a result, the class, as an important organizational unit, was explicitly represented in the analysis. In total, the data consisted of 11660 students nested within 584 classes. An advantage of an HLM analysis is the clear distinction made between student- and class-level variables. Although the outcome variable for the HLM analysis was the person fit statistic at the student level, the predictors included both student- and class-level measures. HLM facilitates the examination of
relationships occurring at each level (i.e., students and classes) and across levels (i.e., whether predictors at one level interact with those at the other level), and assesses the amount of variation accounted for at each level (Snijder & Bosker, 2011; Klinger, Rogers, Anderson, Poth, & Calman, 2006).

**Outcome variable.** The degree of person fit, as indicated by the \( l_z \) value, was used as the outcome variable for the HLM analysis. For each student, five \( l_z \) values were calculated, each with a different plausible \( \theta \) value. Using the Canadian sample, the average \( l_z \) values ranged from -37.26 to 3.05, with mean of -3.20 and standard deviation of 3.49. And 63.2% of students were found to have average \( l_z \) values below the cut point of -1.65 under the null distribution at the critical level of .05. The overall low values of \( l_z \) indicated that a large proportion of student responses did not fit expectations under the models used in the data analysis of the test. One possible explanation is related to the low stakes nature of the test. The results for the TIMSS test are reported at a national level and no results are given to students or schools. That is, test performance has no direct impact on students. As a result, students are unlikely to be motivated to perform at their optimal level, which leads to the misfit of their responses.

Our HLM analyses took into account the five plausible values by repeating the analysis five times using the \( l_z \) values calculated from each set of plausible values as the outcome variable. Final parameter estimates are achieved by averaging the estimates over the five replications. The standard error of the average parameter estimate is calculated by combining the average of the corresponding sampling errors associated with each of the five replications and the imputation variance among the five parameter estimates (HLM 6).

**Student-Level Predictors.** The information on student level predictors was collected in the student questionnaire completed by each student, including gender, frequency of use of the
language of the test at home, time students spend on homework in mathematics, valuing mathematics, self-confidence in learning mathematics, and positive affect toward mathematics. All predictors except gender were composite variables derived by TIMSS on the basis of student responses to a series of statements from the student questionnaire. Time students spend on homework in mathematics was based on the frequency of homework students are given and the amount of time they spend on that homework. Valuing mathematics was based on students’ responses to four statements about mathematics: 1) I think learning mathematics will help me in my daily life; 2) I need mathematics to learn other school subjects; 3) I need to do well in mathematics to get into the university of my choice; and 4) I would like to do well in mathematics to get the job I want. Self-confidence in learning mathematics was based on students’ responses to four statements about mathematics: 1) I usually do well in mathematics; 2) I learn things quickly in mathematics; 3) Mathematics is more difficult for me than for many of my classmates; and 4) Mathematics is not one of my strengths. Finally, positive affect toward mathematics was based on students’ responses to three statements about mathematics: 1) I enjoy learning mathematics; 2) Mathematics is boring; and 3) I like mathematics. The polarity of some statements was reversed so that all statements were of the same polarity

According to the TIMSS technical report, prior to creating these composite variables, principal component analysis had been conducted to explore the dimensionality of each underlying scale, and the scale reliability had been assessed using Cronbach’s alpha. Results suggested that each scale was predominantly unidimensional and the reliability was acceptable (all above .8). Given this, the scores for the composite variables were computed by averaging the numerical values associated with each response option. Larger scores on time students spend on homework in mathematics, valuing mathematics, self-confidence in learning and positive affect
toward mathematics indicated the student spent more time on homework assignments, viewed mathematics more valuable and showed more self-confidence and positive affect in the subject area, respectively. All the student level predictors were significantly correlated with $l_z$. Of test variables, the correlations of student gender, frequency of use of the language of the test at home, and Time students spend on homework in mathematics with $l_z$ were weak (< .100). The correlation of valuing mathematics with $l_z$ was somewhat larger (-.141), and the strongest correlations were for self-confidence in learning mathematics and positive affect toward mathematics (-.483 and -.253, respectively).

Class Level Predictors. Class level predictors were obtained from the teacher’s questionnaire. The classroom teacher of each sampled class responded to the questionnaire. Selected class level variables included class size, teacher's age, gender, teacher's experience in teaching, teacher's job satisfaction, teacher's understanding of curriculum goals, and teacher's expectation of student's success. Teacher's experience in teaching was measured by the number of years each teacher had been teaching. Teacher's job satisfaction, teacher's understanding of curriculum goals, and teacher's expectation of student's success were measured by asking each teacher to rate on a five-point Likert scale from 1 (very high) to 5 (very low). These three variables were re-coded such that higher values were associated with more positive outcomes. Of all the class level predictors, the correlations of teacher's age, gender, and teacher's understanding of curriculum goals with the class mean of $l_z$ were weak (< .100). The correlations for class size, teacher's experience in teaching and teacher's job satisfaction were somewhat larger (-.123, -.125, -.159), and the strongest correlation was for teacher's expectation of student's success (-.221).
HLM Model Building Strategies

The two-level HLM analysis was used to determine the association of student and class level variables with the degree of person fit as reflected by the $l_2$ statistic. All the variables were grand mean centered except for the dichotomous variable gender. Our analysis was completed using the restricted maximum likelihood estimation with the HLM 6 computer program (Raudenbush, Bryk, & Congdon, 2004). In our HLM analysis, the null model with no predictor variables was first examined to obtain the initial partitioning of the total variability into two components corresponding to the two levels of the analyses (i.e., students and classes).

Next, student and class level predictors were included in the HLM analysis to investigate their impact on person fit. As discussed by Hox (2010), the HLM analysis could be very complex even with a modest number of predictor variables. Therefore, it is not generally recommended to estimate the complete model given the computational and interpretational complexities. Normally, researchers use an exploratory approach to find the most parsimonious model that fits the observed data. In the HLM literature, there exist two types of model building strategies: 1) the step up approach outlined by Raudenbush and Bryk (2002) and Snijders and Bosker (2011); and 2) the top down approach suggested by Verbeke and Molenberghs (2000). To identify significant predictors of person fit at different levels and to examine the robustness of the results with different model building strategies, we employed both step up and top down approaches. Steps of the two approaches were given below.

The main idea of the step up approach was to build up the HLM model from the lowest level (i.e., the student level in this study). That is, student level predictors were selected first, followed by class level predictors. Using this approach, the HLM analysis was completed in two sequential steps outlined below:
1. Select student level predictors: all the student level predictors were added to the model to determine which student variables significantly predicted person fit with the assumption of all the effects to be constant across classes (fixed slopes); the insignificant student level predictors were removed; for the significant student level predictors, their effects were then allowed to be different across classes (random slopes), which was modeled as a function of the mean effect across all classes and an error term; the significance of the random slopes were examined by testing the significance of the variance associated with the error term; insignificant random slopes were then fixed.

2. Select class level predictors: all class level predictors were added to the model from step 1 to determine the impact of class level variables on the person fit after accounting for the student level variables; for student level variables with random slopes, class level variables were examined as an attempt to show why the relationship between student level variables and person fit differ across classes; insignificant predictors were removed from the model.

For the top down model building strategy, the analysis started with a model that included all the potential predictors at both levels, followed by the removal of insignificant effects. The steps of the top down approach we used to build the model were given by:

1. A two-level HLM model was analyzed with all student level predictors and all class level predictors; student level predictors were assumed to be random; class level predictors were used to explain both the between class variation in person fit (random intercepts) and the between class variation in the effects of student level predictors
(random slopes); for each random slope, the variance of the error term was tested for significance; insignificant error terms were removed from the model.

2. The model from step 1 was further reduced by removing insignificant level 2 predictors, followed by the removal of insignificant level 1 predictors.

To detect the significance of a predictor variable (a fixed effect), a t test was used. It should be noted that the t distribution under the null hypothesis is not exact due to the complexity associated with two-level data. However, according to Maas and Hox (2004), the t test of fixed effects using restricted maximum likelihood tends to have reliable type I error rates when the number of groups at level 2 is 30 or larger. For random effects, there are at least two ways to test the significance of the variance of the error term: the chi-square test developed by Raudenbush and Bryk(1992) and the likelihood ratio test based on comparing deviances of nested models. The deviance of an HLM model is defined as the natural logarithm of the likelihood of the model multiplied by -2. Although the likelihood ratio test is the preferred approach to testing the significance of random effects, a simulation study conducted by LaHuis and Ferguson(2009) did not find large differences between the two tests, especially when sample sizes are large at both levels. Given that the deviance values are not reported in HLM output when plausible values are used in the analysis, we used the chi-square test instead. Since our analysis was an iterative process with modifications made to the model during testing, model equations were presented in next section, along with the predictors that were entered (or removed) at each step and the modifications that were made based on the testing results.

Results

This section presented results from the HLM analysis. The null model without any predictor variables was summarized first, followed by the detailed description of findings using
the step up model building strategy. Given that the final model obtained using the top down strategy was identical to the one with the step up strategy, the detailed stepwise results for the top down strategy were not reported but available from authors upon request.

The Null Model

In our HLM analysis, the null model was first evaluated. This model is valuable as it provides a picture of how the total variability is partitioned among the levels of analysis. The null model can be written as:

Level-1: \( Y_{ij} = \beta_{0j} + e_{ij} \)
Level-2: \( \beta_{0j} = \gamma_{00} + u_{0j} \)

where \( Y_{ij} \) is the \( l_z \) value of student \( i \) in class \( j \), \( \beta_{0j} \) is the level-1 intercept which can be considered as the mean \( l_z \) value of class \( j \), \( \gamma_{00} \) is the level-2 intercept for \( \beta_{0j} \) which is essentially the grand mean of \( l_z \) values across all students and all classes, and \( e_{ij} \) and \( u_{0j} \) are level-1 and level-2 error terms respectively. In level 1, the \( l_z \) value of student \( i \) in class \( j \) is modeled as a function of the mean \( l_z \) value of class \( j \) and the deviation of the individual \( l_z \) from the class mean. And in level 2, the mean \( l_z \) value of class \( j \) is modeled as a function of the grand mean and the deviation of the class mean from the grand mean. Table 1 shows the estimates for the null model. The estimate for level-2 intercept (i.e., the overall average of \( l_z \) value) was -3.237. As we discussed before, the overall low values of \( l_z \) was likely because the low stakes nature of the test. Of the total variability of \( l_z \) values, about 70% was found to exist between students and 30% occurred between classrooms.
The Step up Strategy

Step 1. A two level model with only student level predictors was tested. All student variables were fixed at first. Time students spend on homework in mathematics (TSHOMEW), self-confidence in learning mathematics (SELF), and positive affect toward mathematics (PAFFECT) were found to be significant predictors of person fit and therefore the rest of student variables were removed. Next, we considered the effects of the three significant student level predictors to be random. Given that the variances of the error terms for TSHOMEW and PAFFECT were found to be insignificant, the effect of these variables were then fixed (i.e., the effect was assumed to be constant across classes). Therefore, only the effect of student SELF was found to vary across classes. The final model in step 1 is presented below:

Level-1: \( Y_{ij} = \beta_{0j} + \beta_{1j}(TSHOMEW) + \beta_{2j}(SELF) + \beta_{3j}(PAFFECT) + e_{ij} \)

Level-2: \( \beta_{0j} = \gamma_{00} + u_{0j} \)
\[
\begin{align*}
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} + u_{2j} \\
\beta_{3j} &= \gamma_{30} 
\end{align*}
\]

Table 2 shows the estimates for the above model. With respect to the definition of \( l_z \) statistics, negative values indicate further deviant from the normal responses. The positive coefficients of "TSHOMEW" and "PAFFECT" suggests that students who spend more time on doing homework in mathematics and have more positive affect toward mathematics are more likely to respond normally. On the other hand, the negative coefficient SELF implies that more self-confidence in learning mathematics tends to be associated with higher levels of misfit. In addition, the significant variance of the level 2 error term for SELF indicates that its effect on person fit is different among classes. In total, this model accounts for 18.8\% of variation of \( l_z \) values compared to the null model.
Step 2. We further added class-level variables to the final model from step 1 to identify significant predictors of person fit at the class level. First, we included all these class level variables as predictors for the level-1 intercept (i.e., class mean of $l_z$ values) and the coefficient for “SELF” (i.e., the within class effect of “SELF” on person misfit). Next, we excluded insignificant variables from the model. We found that the variable class size was associated with a substantial amount of missing data (about 20%) and therefore was excluded from analysis. Of the rest of class level variables, only teacher’s expectation of student’s success (“TE”) contributed significantly to the prediction of the level-1 intercept and variation of “SELF” across classes. The full model containing both student- and class-level variables is presented below.

Level-1: $Y_{ij} = \beta_{0j} + \beta_{1j}(TSHOMEW) + \beta_{2j}(SELF) + \beta_{3j}(PAFFECT) + \epsilon_{ij}$
Level-2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(TE) + u_{0j}$
   $\beta_{1j} = \gamma_{10}$
   $\beta_{2j} = \gamma_{20} + \gamma_{21}(TE) + u_{2j}$
   $\beta_{3j} = \gamma_{30}$

The estimates for the full model using the step up strategy are shown in Table 3. The effect of teacher’s expectation of student’s success on the class mean of $l_z$ was found to be positive, suggesting that more expectations of student’s success from teachers tend to be associated with higher levels of misfits in student responses. Given that the effect of students’ self-confidence in learning mathematics on $l_z$ was negative at level 1, the negative coefficients of teacher’s expectation on this effect at level 2 indicated that the higher the teacher’s expectation, the stronger the effect of student’s self-confidence on person fit. In total, the full model accounts for 19.81% of variation of $l_z$ values compared to the null model.

**Discussion**

Measurement models play fundamental roles in the design, analysis and scoring of educational and psychological assessment. To validate the use of a measurement model in
analyzing student response data, it is critical to evaluate the consistency of each student’s response pattern and model expectations. Person fit analysis evaluates measurement inaccuracy at the individual level by assessing the degree to which a student’s observed item responses match the patterns of normative responses expected by the measurement models. Evaluation of person fit is important as the test scores of students with misfitting responses might fail to provide a useful and valid measure of their proficiencies. This could lead to faulty test score interpretations and actions that are unwarranted from the perspective of fair and accurate testing.

The current study applied the person fit statistic, $l_z$, to the TIMSS 2007 data to explore student and classroom factors associated with person fit on the test. Overall, there is a very poor fit between TIMSS student responses and the IRT expectations, which makes one question the validity of the inferences that have been drawn from the test results. Although international large scale assessments such as TIMSS do not report student level test results, one might suspect the aggregation of inaccurate individual student scores could lead to severe bias at the national level.

We used the HLM approach to take into account the multilevel structure of test data. Our results suggest that 30% the total variability of $l_z$ values is attributable to the class level, which echoes the previous finding from Petridou and Williams (2007) that person misfit is a two-level phenomenon and the classroom tends to make a significant contribution to the misfitting responses. In our HLM analysis, we used two model building strategies for identifying significant student and class level predictors for person fit. Results were unchanged when different strategies were used, which strengthened our findings. Three student-level variables and one class-level variable were found as significant predictors of the degree of person fit as indicated by the values of $l_z$. Student-level variables included time students spend on homework in mathematics, positive affect toward mathematics and self-confidence in learning Mathematics.
The positive effect of the first two student-level variables on the values of $l_z$ is understandable as the more time a student spend on homework in mathematics and the more positive affect a student has toward mathematics, one would predict the student is more likely to respond expectedly and consistently. However, the negative effect of self-confidence in learning mathematics seems surprising. Our hypothesis is that the negative relationship might be due to overconfidence that leads to careless mistakes or due to creative responding behaviors that the student has answered easy items incorrectly for the reason of interpreting these items in a unique, creative manner (Meijer, 1996). An example of misfitting item responses produced by a student with high self confidence in learning mathematics is presented in Table 4. Based on their difficulty parameter estimates, items have been reordered from easiest to most difficult. The average plausible $\theta$ value for the student was 2.08, indicating a moderately high ability. Inspection of his/her item responses revealed that the student was able to answer most of the difficult items (16 out of the last 20 items) but failed to correctly answer most of the easy items (14 out of the first 20 items). A possible reason is that the student reinterpreted easy items because they were “too easy to be true”, which led to incorrect answers. In addition, the full HLM model indicates the effect of self-confidence in learning mathematics tends to vary across classes.

The class-level variable, teacher’s expectation of student’s success, was found to significantly predict the class mean $l_z$ values as well as the strength of the effect of self-confidence in learning mathematics on person fit across classes. One would think that higher expectations of student’s success from teachers could result in the fact that students put more efforts to do well on tests thereby producing more consistent and interpretable responses. However, the results suggest the opposite. It is unclear why the relationship of teacher’s
expectation of student’s success and the mean \( t_2 \) of the class is negative. In addition, the negative effect of self-confidence in learning mathematics on person fit tends to be stronger as teacher’s expectation of student’s success becomes higher. A tentative explanation is that when teachers set high expectations of students’ success, high ability students are more likely to reinterpret easy items and therefore the effect of creative responding that leads to incorrect responses to easy items tends to be strengthened.

Limitations

The current investigation was based on secondary data analyses with existing large scale data. Therefore, not all relevant variables that could potentially lead to person misfit were available for our analysis. In addition, all variables considered in the study were measured (versus manipulated) which limits our ability to make causal inferences. For example, the impact of self-confidence in learning mathematics on misfitting responses is open to alternative explanations such as reverse causality and unmeasured “third variables”. However, experimental designs that allow causal interpretation of these relationships is not feasible with the specific variables we chose to analyze in this paper due to ethical and practical considerations.

In addition, some student and class level variables are self-report measures. Although self-report measures are common as they are relatively easy to obtain and are often the only feasible method to measure certain constructs, many researchers have questioned the construct validity of self-report measures. Both theory and research indicate that psychological, sociological, linguistic, experiential and contextual factors could lead to self-report response bias (e.g. Harrison, McLaughlin & Coalter, 1996; Lanyon & Goodstein, 1997). For example, people may have the tendency to answer questions in a socially desirable way, and this tendency could potentially influence how students and teachers responded to the TIMSS questionnaires. If this is
the case, our findings would be contaminated by the presence of self-report bias. Therefore, the results of this paper must be interpreted cautiously.

**Conclusion**

Person-fit statistics have been criticized for not providing clear indications of how and why misfits occur. We agree that misfitting responses identified based purely on person fit statistics should never be considered as sufficient evidence for invalidating individual student test results. Follow up analysis must be conducted to scrutinize the possible reasons of misfit. Results from the current study suggest that examining student and class level variables associated with person fit may have the potential to shed lights on why misfits have occurred. Furthermore, information about students’ response processes, such as students’ verbal reports, eye tracking information, and reaction time (American Educational Research Association, National Council for Measurement in Education, & American Psychological Association, 1999), can be collected. This type of information provides relatively detailed pictures of how students actually solve items on tests, which might help understand the causes for misfits. In this way, the results from person-fit statistics can be interpreted substantially and meaningfully and, more importantly, we can improve the validity of test score inferences.
References


TABLE 1
Estimates of model parameters for the null model

<table>
<thead>
<tr>
<th>Fixed part</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_{00}$ (INTERCEPT)</td>
<td>-3.237*</td>
</tr>
</tbody>
</table>

Random part

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = Var (e_{ij})$</td>
<td>10.210</td>
</tr>
<tr>
<td>$\tau_0 = Var(u_{0j})$</td>
<td>4.429*</td>
</tr>
<tr>
<td>$\hat{\rho}$</td>
<td>0.300</td>
</tr>
</tbody>
</table>

Notes: * $(p \leq 0.05)$; $\sigma^2$ is the variance of level-1 residues; $\tau_0$ is the variance of level-2 residues; and $\hat{\rho}$ is the estimate of intraclass correlation.
<table>
<thead>
<tr>
<th>Fixed part</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_{00}$ (INTERCEPT)</td>
<td>-3.162*</td>
</tr>
<tr>
<td>$\gamma_{10}$ (TSHOMEW)</td>
<td>0.135*</td>
</tr>
<tr>
<td>$\gamma_{20}$ (SELF)</td>
<td>-0.417*</td>
</tr>
<tr>
<td>$\gamma_{30}$ (PAFFECT)</td>
<td>0.038*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random part</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>$\tau_0 = Var(u_{0j})$</td>
</tr>
<tr>
<td>$\tau_1 = Var(u_{2j})$</td>
</tr>
<tr>
<td>$\tau_{01} = Cov(u_{0j}, u_{2j})$</td>
</tr>
</tbody>
</table>

Notes: * ($p \leq 0.05$), $\gamma$ is regression coefficient, $\sigma^2$ is level-1 variance component, $\tau$ is level-2 variance component
**TABLE 3**

*Estimates of model parameters for step 3*

<table>
<thead>
<tr>
<th>Fixed part</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_{00}$ (INTERCEPT)</td>
<td>-3.156*</td>
</tr>
<tr>
<td>$\gamma_{01}$ (TE)</td>
<td>-0.465*</td>
</tr>
<tr>
<td>$\gamma_{10}$ (TSHOMEW)</td>
<td>0.136*</td>
</tr>
<tr>
<td>$\gamma_{20}$ (SELF)</td>
<td>-0.415*</td>
</tr>
<tr>
<td>$\gamma_{21}$ (TE)</td>
<td>-0.067*</td>
</tr>
<tr>
<td>$\gamma_{30}$ (PAFFECT)</td>
<td>0.037*</td>
</tr>
</tbody>
</table>

**Random part**

| $\sigma^2$ | 8.187 |
| $\tau_0 = Var(u_{0j})$ | 3.078* |
| $\tau_1 = Var(u_{2j})$ | 0.058* |
| $\tau_{01} = Cov(u_{0j}, u_{2j})$ | 0.416 |

Notes: * $\hat{\gamma}$ is regression coefficient, $\sigma^2$ is level-1 variance component, $\tau$ is level-2 variance component
TABLE 4

An example of misfitting responses from a student with high self confidence in mathematics

<table>
<thead>
<tr>
<th>item</th>
<th>difficulty</th>
<th>response</th>
<th>P(x=1)</th>
<th>residual</th>
<th>item</th>
<th>difficulty</th>
<th>response</th>
<th>P(x=1)</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.631</td>
<td>0</td>
<td>0.997</td>
<td>-0.997</td>
<td>24</td>
<td>0.566</td>
<td>0</td>
<td>0.9433</td>
<td>-0.943</td>
</tr>
<tr>
<td>2</td>
<td>-0.449</td>
<td>0</td>
<td>0.986</td>
<td>-0.986</td>
<td>25</td>
<td>0.627</td>
<td>0</td>
<td>0.9866</td>
<td>-0.987</td>
</tr>
<tr>
<td>3</td>
<td>-0.437</td>
<td>0</td>
<td>0.985</td>
<td>-0.985</td>
<td>26</td>
<td>0.634</td>
<td>1</td>
<td>0.9797</td>
<td>0.020</td>
</tr>
<tr>
<td>4</td>
<td>-0.352</td>
<td>0</td>
<td>0.997</td>
<td>-0.997</td>
<td>27</td>
<td>0.678</td>
<td>1</td>
<td>0.8019</td>
<td>0.198</td>
</tr>
<tr>
<td>5</td>
<td>-0.321</td>
<td>0</td>
<td>0.886</td>
<td>-0.886</td>
<td>28</td>
<td>0.710</td>
<td>1</td>
<td>0.9251</td>
<td>0.075</td>
</tr>
<tr>
<td>6</td>
<td>-0.319</td>
<td>1</td>
<td>0.987</td>
<td>0.013</td>
<td>29</td>
<td>0.711</td>
<td>1</td>
<td>0.9634</td>
<td>0.037</td>
</tr>
<tr>
<td>7</td>
<td>-0.296</td>
<td>1</td>
<td>0.935</td>
<td>0.066</td>
<td>30</td>
<td>0.746</td>
<td>1</td>
<td>0.9383</td>
<td>0.062</td>
</tr>
<tr>
<td>8</td>
<td>-0.259</td>
<td>0</td>
<td>0.994</td>
<td>-0.994</td>
<td>31</td>
<td>0.782</td>
<td>1</td>
<td>0.8536</td>
<td>0.146</td>
</tr>
<tr>
<td>9</td>
<td>-0.251</td>
<td>0</td>
<td>0.983</td>
<td>-0.983</td>
<td>32</td>
<td>0.803</td>
<td>1</td>
<td>0.9544</td>
<td>0.046</td>
</tr>
<tr>
<td>10</td>
<td>-0.002</td>
<td>1</td>
<td>1.000</td>
<td>0.000</td>
<td>33</td>
<td>0.815</td>
<td>1</td>
<td>0.964</td>
<td>0.036</td>
</tr>
<tr>
<td>11</td>
<td>0.060</td>
<td>1</td>
<td>0.965</td>
<td>0.035</td>
<td>34</td>
<td>0.855</td>
<td>0</td>
<td>0.9521</td>
<td>-0.952</td>
</tr>
<tr>
<td>12</td>
<td>0.092</td>
<td>0</td>
<td>0.927</td>
<td>-0.927</td>
<td>35</td>
<td>0.879</td>
<td>1</td>
<td>0.9405</td>
<td>0.060</td>
</tr>
<tr>
<td>13</td>
<td>0.132</td>
<td>0</td>
<td>0.972</td>
<td>-0.972</td>
<td>36</td>
<td>0.899</td>
<td>1</td>
<td>0.87</td>
<td>0.130</td>
</tr>
<tr>
<td>14</td>
<td>0.200</td>
<td>0</td>
<td>0.991</td>
<td>-0.991</td>
<td>37</td>
<td>0.949</td>
<td>1</td>
<td>0.7859</td>
<td>0.214</td>
</tr>
<tr>
<td>15</td>
<td>0.212</td>
<td>0</td>
<td>0.981</td>
<td>-0.981</td>
<td>38</td>
<td>1.039</td>
<td>1</td>
<td>0.9324</td>
<td>0.068</td>
</tr>
<tr>
<td>16</td>
<td>0.247</td>
<td>1</td>
<td>1.000</td>
<td>0.000</td>
<td>39</td>
<td>1.058</td>
<td>1</td>
<td>0.9611</td>
<td>0.039</td>
</tr>
<tr>
<td>17</td>
<td>0.277</td>
<td>0</td>
<td>0.971</td>
<td>-0.971</td>
<td>40</td>
<td>1.114</td>
<td>1</td>
<td>0.8228</td>
<td>0.177</td>
</tr>
<tr>
<td>18</td>
<td>0.311</td>
<td>0</td>
<td>0.974</td>
<td>-0.974</td>
<td>41</td>
<td>1.12</td>
<td>1</td>
<td>0.9489</td>
<td>0.051</td>
</tr>
<tr>
<td>19</td>
<td>0.332</td>
<td>0</td>
<td>0.931</td>
<td>-0.931</td>
<td>42</td>
<td>1.178</td>
<td>1</td>
<td>0.9621</td>
<td>0.038</td>
</tr>
<tr>
<td>20</td>
<td>0.363</td>
<td>1</td>
<td>0.971</td>
<td>0.029</td>
<td>43</td>
<td>1.195</td>
<td>0</td>
<td>0.9123</td>
<td>-0.912</td>
</tr>
<tr>
<td>21</td>
<td>0.384</td>
<td>1</td>
<td>0.989</td>
<td>0.011</td>
<td>44</td>
<td>1.209</td>
<td>1</td>
<td>0.824</td>
<td>0.176</td>
</tr>
<tr>
<td>22</td>
<td>0.396</td>
<td>1</td>
<td>0.947</td>
<td>0.053</td>
<td>45</td>
<td>1.258</td>
<td>1</td>
<td>0.9978</td>
<td>0.002</td>
</tr>
<tr>
<td>23</td>
<td>0.501</td>
<td>1</td>
<td>0.984</td>
<td>0.016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Student’s ability = 2.085