The Use of Multilevel Item Response Theory Modeling in Applied Research: An Illustration

Dena A. Pastor
Center for Assessment and Research Studies
James Madison University

Embedding item response theory (IRT) models within a multilevel modeling framework has been shown by many authors to allow better estimation of the relationships between predictor variables and IRT latent traits (Adams, Wilson, & Wu, 1997). A multilevel IRT model recently proposed by Kamata (1998, 2001) yields the additional benefit of being able to accommodate data that are collected in hierarchical settings. This expansion of multilevel IRT models to three levels allows not only the dependency typically found in hierarchical data to be accommodated, but also the estimation of (a) latent traits at different levels and (b) the relationships between predictor variables and latent traits at different levels. The purpose of this article is to provide both a description and application of Kamata’s 3-level IRT model. The advantages and disadvantages of using multilevel IRT models in applied research are discussed and directions for future research are given.

The practice of embedding item response theory (IRT) models within a multilevel modeling framework originated several decades ago and since then, many different specifications have been proposed for a variety of purposes (see Kamata, 2001, for a review). Recently, certain multilevel item response theory models have received increased attention as evidenced by a 2002 American Educational Research Association symposium devoted entirely to Kamata’s (1998, 2001) multilevel reformulation of the Rasch model. This heightened interest in multilevel IRT modeling may be partly attributable to the increased training of social science doctoral students in both IRT and multilevel modeling as well as the widely available, non-IRT specific software that can be used for estimation (i.e., HLM, SAS,

Requests for reprints should be sent to Dena A. Pastor, MSC 6806, Anthony-Seeger Hall, James Madison University, Harrisonburg, VA 22807. E-mail: pastorda@jmu.edu
Regardless of the source, interest in such models has led to a growing area of research that includes the use of multilevel IRT models to explore such psychometric issues as DIF detection (Cheong, 2001; Kamata, 1998; Luppescu, 2002), test equating (Chu & Kamata, 2000), and dimensionality assessment (Beretvas & Williams, 2002).

The purpose of this article is to explore and illustrate the use of Kamata’s (1998, 2001) three-level IRT model in educational and psychological measurement and research. The advantages of this model include the ability (a) to accommodate hierarchically structured data, (b) to estimate latent traits at different levels, (c) to provide improved estimates of the relationships between predictor variables and latent traits, and (d) to investigate the relationships between latent traits and predictor variables at different levels.

The first advantage of being able to accommodate hierarchically structured data is particularly relevant in educational measurement, where the nested structure of students within classrooms (or students within schools, students within school districts) is frequently encountered. When one is interested in the relationships between variables, the use of traditional statistical models, such as regression models, to analyze such data is problematic because the assumption of independent observations is often violated (Kreft & de Leeuw, 1998). The second advantage of being able to estimate student latent traits and latent traits at higher levels (i.e., classroom, school, school district) is also beneficial in educational measurement, particularly when teacher-level or school-district-level ability estimates are desired for the purposes of accountability. In fact, it is not uncommon for state standards-based assessments to provide scores at both the student and school level (Lane & Stone, 2002).

The third advantage in using Kamata’s (1998, 2001) three-level IRT model is that it can result in improved estimation of the relationship between latent traits and predictor variables. Traditionally, exploring the relationships between predictor variables and IRT latent trait estimates has involved a two-step approach of first, estimating IRT parameters and second, using the resulting latent trait estimates as dependent variables in a regression analysis. Unfortunately, because IRT latent trait estimates often have heteroscedastic or uneven error, the assumption of independent and identically distributed errors in many statistical analyses is violated (Mislevy, 1984). As illustrated by many authors, including Adams et al. (1997), such an assumption violation often leads to an attenuation of the true relationship between the latent trait and the predictor variable (see also Kamata, 1998; Mislevy, 1984; Zwinderman, 1991). Multilevel IRT modeling is advantageous in that it can obtain more accurate estimates of the relationship between predictor variables and the latent traits by simultaneous estimation of not only the IRT item and latent trait parameters, but also the parameters that describe the effects of the predictor variables. Simultaneous estimation allows for better
estimation of the true relationship by incorporating the standard errors of the latent traits into the model (Maier, 2001).

Increased precision of the estimated effects within a group, such as a school, is also possible with multilevel IRT models that use empirical Bayes estimation. For instance, if the estimated effect within a school is considered unreliable, possibly because of small sample size, more weight is placed on the estimate of the effect in the overall sample. By estimating the effects for a particular school as a compromise between the relationship within a school and the relationship across schools, the borrowing of strength advantage of multilevel modeling can be used to obtain accurate estimates of the relationships within schools, regardless of sample size (Raudenbush & Bryk, 2002). Of course, the borrowing of strength advantage is desirable only if the schools are, indeed, related to one another, the model is specified correctly at the school-level (Raudenbush & Bryk, 2002), and the “shrinking” of the estimate for a particular school toward the overall estimate in the entire sample can be tolerated (Kreft & de Leeuw, 1998).

A fourth advantage in using a multilevel IRT model is that it allows multilevel research questions to be pursued using a latent trait as the dependent variable. In the past 10 years there have been numerous applications of non-IRT-based multilevel models in educational research (i.e., Anderman & Kimweli, 1997; Lee, 2000; Lee, Loeb, & Lubeck, 1998; Raudenbush, Rowan, & Kang, 1991). In fact, Lee (2000) advocated for the use of non-IRT-based multilevel models as the quantitative method of choice for studying school effects. An example of an application of multilevel modeling in educational research is the Tennessee Value-Added Assessment System, where multilevel modeling is used to determine what effect school systems, schools, and teachers have on student learning (Sanders, Saxton, & Horn, 1997). Recently, Stone and Lane (2003) proposed using a multilevel model to explore school-level and classroom-level effects on standardized test performance over time. With Kamata’s (1998, 2001) three-level model, it is possible to explore similar multilevel research questions with the additional benefit of having a latent trait as the dependent variable.

As opposed to non-IRT-based multilevel models, multilevel IRT models are relatively new and have rarely been used in applied research. An exception is Cheong and Raudenbush (2000), who recently applied Kamata’s (1998) reformulation to study the dimensions of a child behavior scale across different demographic groups and different neighborhoods. This study will offer an additional application of multilevel IRT modeling by using a series of multilevel IRT models to examine the relationships between demographic variables and adolescent academic self-esteem (ASE). All multilevel models are specified using Kamata’s (1998, 2001) three-level formulation with Level 1 as the item-level model, Level 2 as the person-level model, and Level 3 as the site-level
model. A model building process will be used to explore the following research questions:

- Does ASE vary from person to person within a site?
- Does ASE vary from site to site?
- Does one gender have significantly different ASE than the other?
- Is there a relationship of age with adolescent ASE?
- Is the relationship of age with adolescent ASE different for males and females? (i.e., is there an interaction?)
- Do gender differences in ASE vary across sites?
- Does the relationship of age with ASE differ across sites?
- Does the interaction of age and gender vary across sites?
- Can education level of the site explain site-to-site variation in the relationships of demographic variables with ASE?

It is hoped that this application of multilevel IRT modeling will clarify for the reader the measurement applications and research questions that can be posed with the three-level IRT model. The following sections describe in detail the self-esteem measure, the sample, and the multilevel IRT models.

METHOD

Instrument

The third edition of the Culture Free Self-Esteem Inventories (CFSEI–3; Battle, 2002) is a norm-referenced, self-report instrument used to measure both the global and specific dimensions of child and adolescent self-esteem. This study uses the ASE subscale, which consists of dichotomous items that tap into a person’s “perception of their abilities, attitudes and values as they relate to intelligence, school and academic skill” (Battle, 2002). Preliminary psychometric analyses of the items supported the omission of three items on the academic subscale resulting in a final scale consisting of eight items. All items were scored such that item endorsement indicated a higher level of academic self-esteem. Although the use of eight items is satisfactory for the purposes of illustrating a model, it should be noted that for applied purposes, eight items typically would be considered too few to obtain an accurate estimation of a person’s theta in IRT (Hambleton, Swaminathan, & Rogers, 1991).

Preliminary analyses were used to ensure that the eight ASE items were appropriate for use with the Rasch model (see Pastor, 2001). The results of such analyses supported equal discrimination across items and the measurement of a single dimension by all eight items.
Sample

Data were collected between the fall of 1998 and the fall of 2000 for the norming of the CFSEI–3 from adolescents between the ages of 12 and 18 years of age. The final sample consisted of 905 respondents from 13 data collection sites located throughout the United States. Demographic information for the sample by site is listed in Table 1.

Person-Level Predictor Variables

Gender was coded such that a 0 represented a female and a 1 represented a male. Age was recoded by subtracting the number 12 from all values. Recoding age in this fashion eased interpretation of the intercept for those multilevel IRT models including the age variable. Multiplying the values of age by the values of gender created the interaction between gender and age.

Site-Level Predictor Variables

Whereas the person-level predictor variables of age and gender were acquired directly from the adolescent, the site-level predictor variables were acquired through linking the zip code reported by the adolescent to U.S. 2000 census data available for the zip code. In the U.S. 2000 census data, the proportion of the population with various education levels in the zip code area was available. By linking the

<table>
<thead>
<tr>
<th>Site</th>
<th>n</th>
<th>Male</th>
<th>Female</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin, TX</td>
<td>68</td>
<td>43</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>24</td>
<td>29</td>
<td>13.62</td>
</tr>
<tr>
<td>Bismarck, ND</td>
<td>102</td>
<td>54</td>
<td>48</td>
<td>0</td>
<td>11</td>
<td>11</td>
<td>20</td>
<td>18</td>
<td>19</td>
<td>23</td>
<td>10.04</td>
</tr>
<tr>
<td>Brownsville, TX</td>
<td>50</td>
<td>28</td>
<td>22</td>
<td>19</td>
<td>3</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.71</td>
</tr>
<tr>
<td>Burlington, VT</td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>12.08</td>
</tr>
<tr>
<td>Hattiesburg, MS</td>
<td>239</td>
<td>107</td>
<td>132</td>
<td>17</td>
<td>44</td>
<td>65</td>
<td>28</td>
<td>51</td>
<td>21</td>
<td>13</td>
<td>9.83</td>
</tr>
<tr>
<td>Katonah, NY</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>14.95</td>
</tr>
<tr>
<td>New Lenox, IL</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>11.88</td>
</tr>
<tr>
<td>Newark, NJ</td>
<td>76</td>
<td>37</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>11</td>
<td>20</td>
<td>19</td>
<td>16</td>
<td>8.78</td>
</tr>
<tr>
<td>Overland Park, KS</td>
<td>21</td>
<td>10</td>
<td>11</td>
<td>1</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14.58</td>
</tr>
<tr>
<td>Perry, NY</td>
<td>57</td>
<td>28</td>
<td>29</td>
<td>15</td>
<td>35</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10.47</td>
</tr>
<tr>
<td>Racine, WI</td>
<td>18</td>
<td>14</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>10.37</td>
</tr>
<tr>
<td>St. George, UT</td>
<td>212</td>
<td>95</td>
<td>117</td>
<td>29</td>
<td>40</td>
<td>46</td>
<td>26</td>
<td>25</td>
<td>24</td>
<td>22</td>
<td>12.22</td>
</tr>
<tr>
<td>Tyler, TX</td>
<td>24</td>
<td>14</td>
<td>10</td>
<td>0</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>11.20</td>
</tr>
<tr>
<td>Overall</td>
<td>905</td>
<td>454</td>
<td>451</td>
<td>86</td>
<td>154</td>
<td>194</td>
<td>102</td>
<td>134</td>
<td>118</td>
<td>117</td>
<td>—</td>
</tr>
</tbody>
</table>
individual data to the census data, the average proportion of the site sample in each education level was obtained. This proportion was multiplied by the approximate number of years of education the category level represented, and these values were then summed across categories to create a variable named *Education Index* for a site. This variable represents the number of years of education for a typical person in the site. The second site-level predictor variable is the sample size of a site, which was used to determine if the site sample size had an impact on the estimate of site effects. The education index and sample size for each site are displayed in Table 1.

**Models**

*The unconditional model.* The first of the multilevel IRT models fit to the data was an unconditional model, which did not include either person-level or site-level predictor variables. The first level of the unconditional three-level IRT model, which models variation of item responses within people, was used to model the log-odds of the probability of endorsing item \(i\) \((i = 1, \ldots, 8)\) for person \(j\) \((j = 1, \ldots, 905)\) in site \(g\) \((g = 1, \ldots, 13)\).

\[
\log \left( \frac{p_{ijg}}{1 - p_{ijg}} \right) = \beta_{0jg} + \beta_{1jg} X_{1ijg} + \beta_{2jg} X_{2ijg} + \cdots + \beta_{(k-1)jg} X_{(k-1)ijg} \\
= \beta_{0jg} + \sum_{q=1}^{k-1} \beta_{qjg} X_{qijg},
\]

(1)

where \(X_{qijg}\) is the \(q\)th indicator variable for person \(j\) and will equal \(-1\) when \(q = i\) and \(0\) when \(q \neq i\). Given the constraint that \(\beta_{qjg} = 0\) (to ensure that the design matrix is full rank), the intercept for a person, \(\beta_{0jg}\), can be interpreted in two ways: (a) the effect of the \(k\)th item, the reference item, for person \(j\) or (b) an overall effect common to all items for person \(j\), which is related to the latent trait estimate of self-esteem for person \(j\) in site \(g\). The offset of the \(q\)th item from the \(k\)th item effect is represented by \(\beta_{qjg}\). The item with the lowest difficulty index (the hardest item), as indicated in a classical item analysis, was used as the reference item.

The second level or person-level of the unconditional model was used to model variation of person self-esteem within sites.

\[
\begin{align*}
\beta_{0jg} &= \gamma_{00g} + u_{0jg} \\
\beta_{1jg} &= \gamma_{10g} \\
\vdots \\
\beta_{(k-1)jg} &= \gamma_{(k-1)0g}
\end{align*}
\]

(2)
The item effects \((\beta_{1jg}, \ldots, \beta_{(k-1)jg})\) were specified as fixed across persons, so that the second level of the unconditional model was modeling only variation in \(\beta_{0jg}\) among adolescents within sites. It was assumed that the distribution of \(u_{0jg}\) (the ASE estimates for persons) was \(N(0, \tau_\beta)\) where \(\tau_\beta\) represents the variation of ASE among adolescents within sites. It was also assumed that the variation of ASE within a site was the same across sites, such that all sites had variation equal to \(\tau_\beta\).

The third level of the model was used to model variation among sites in ASE using the parameters estimated for each site \((\nu_{00g}, \nu_{10g}, \ldots, \nu_{(k-1)0g})\) in level 2 as outcome variables. In the following specification of the model, the item effects were specified as fixed across sites whereas the latent trait effects were allowed to vary at random across sites.

\[
\begin{cases}
\nu_{00g} = \pi_{00} + r_{00g} \\
\nu_{10g} = \pi_{10} \\
\nu_{20g} = \pi_{20} \\
\vdots \\
\nu_{(k-1)0g} = \pi_{(k-1)0}
\end{cases}
\]  

Variation in \(r_{00g}\) (the ASE estimates for sites) was assumed to be distributed as \(N(0, \tau_\gamma)\) with \(\tau_\gamma\) representing the variation among sites in ASE. Again, the item effects were fixed across sites (as specified in level 3) and fixed across persons (as specified in level 2), such that all sites and all persons received the same item effect estimates.

Kamata (1998, 2001) illustrated how the combined model can be written with the dependent variable as the probability of person \(j\) in site \(g\) endorsing item \(i\):

\[
p_{ijg} = \frac{1}{1 + \exp\left(-[(r_{00g} + u_{0jg}) - (\pi_{000} - \pi_{00})]\right)}. \tag{4}
\]

When the model is written in this format, it emphasizes the fact that the Rasch model is incorporated into the estimation procedures for the entire model.

There are several reasons why the unconditional model was estimated prior to multilevel IRT models incorporating person-level or site-level predictor variables. By allowing ASE latent trait estimates to vary randomly across persons within a site and randomly across sites, it was determined whether there existed variation at the person-level or site-level that might be explained by the addition of predictor variables in the model. If there was little to no variation in ASE among persons, for example, then adding person-level predictor variables in the model would offer no additional information. To determine if there was significant variation among persons within sites and among sites in ASE to warrant the inclusion of
predictor variables, the following null hypotheses were tested using an alpha level of 0.05: $H_0: \tau_\beta = 0$ and $H_0: \tau_\gamma = 0$.

**Residual plots.** Because previous literature suggests a possible nonlinear relationship between age and ASE, each person’s empirical Bayes latent trait estimate from the unconditional model, $u_{ojg}$, was plotted against his or her age value by site to determine the nature of the relationship between age and ASE. The empirical Bayes latent trait estimate for the person, $u_{ojg}$, was then added to the empirical Bayes latent trait estimate for the site, $r_{oog}$, to put the latent trait estimates for all persons in the sample on the same scale. The sum of these two values was then plotted against age to explore the relationships of age with ASE in the overall sample. Regression lines for both males and females using linear and quadratic terms for age were imposed on the plots to determine how the relationship of age with ASE should be represented.

**Age and gender effects model.** Once the unconditional model was fit to the data and given that there was significant variation within and between sites in ASE, the model building approach proceeded by explaining within site variation first and between site variation second as advocated by Raudenbush and Bryk (2002). To determine if ASE variation across persons was associated with the age of the person or his or her gender, a model including the effects of age and gender was fit to the data. The interaction between age and gender was also included to determine if the relationship of ASE with age differed for males and females. The level 1 model remained the same as in equation 1 and the second level model was specified as follows:

$$
\begin{align*}
\beta_{ijg} &= \gamma_{00g} + \gamma_{01g}(\text{gender})_{jg} + \gamma_{02g}(\text{age})_{jg} \\
&\quad + \gamma_{03g}(\text{gender} \times \text{age})_{jg} + u_{ojg} \\
&\quad + \gamma_{10g} \\
&\quad + \ldots \\
&\quad + \gamma_{(k-1)0g} \\
\end{align*}
$$

Because of the way in which the gender and age variables were coded, the parameter $\gamma_{00g}$ is interpreted as the level of ASE for a 12-year-old female in a site $g$, $\gamma_{01g}$ is the gender effect in site $g$ controlling for all other variables in the model, $\gamma_{02g}$ is the age effect for site $g$ controlling for all other variables in the model, and $\gamma_{03g}$ is the interaction between gender and age in site $g$ controlling for all other variables in the model. The residual, $u_{ojg}$, or latent trait estimate of ASE for person $j$ in site $g$ is now his or her level of ASE controlling for the effects of gender, age, and the gender by age interaction.

The third level of the model did not contain any site-level predictor variables and was specified as follows:
where the effects of the person-level predictor variables of Level 2 were allowed to vary randomly over sites. Of main interest for this study was whether or not the relationships between person-level predictor variables and ASE varied across sites; therefore, the effects of all person-level predictor variables were initially allowed to vary at random across sites. The parameters $\pi_{000}$, $\pi_{010}$, $\pi_{020}$, and $\pi_{030}$, have similar interpretation as $\gamma_{00g}$, $\gamma_{01g}$, $\gamma_{02g}$, and $\gamma_{03g}$ given previously, except the interpretation now applies to the overall sample, not just the site. The parameters $r_{00g}$, $r_{01g}$, $r_{02g}$, and $r_{03g}$ are offsets of the site from the overall fixed effect.

The assumption was made that the random effects have the following distribution:

$$
\begin{bmatrix}
    r_{00g} \\
    r_{01g} \\
    r_{02g} \\
    r_{03g}
\end{bmatrix}
= N
\begin{bmatrix}
    0 \\
    0 \\
    \vdots \\
    0
\end{bmatrix}
\begin{bmatrix}
    \text{var}(r_{00g}) & \text{cov}(r_{00g}, r_{01g}) & \text{var}(r_{01g}) \\
    \text{cov}(r_{00g}, r_{02g}) & \text{cov}(r_{01g}, r_{02g}) & \text{var}(r_{02g}) \\
    \vdots & \vdots & \vdots \\
    \text{cov}(r_{00g}, r_{03g}) & \text{cov}(r_{01g}, r_{03g}) & \text{cov}(r_{02g}, r_{03g}) & \text{var}(r_{03g})
\end{bmatrix}
$$

Of particular interest was the significance of the variation in the random effects, $\text{var}(r_{00g})$ through $\text{var}(r_{03g})$, and the significance of the fixed effects, $\pi_{000}$ through $\pi_{030}$. If the variation for a random effect was not significant, further models were fit to the data specifying the effect of the variable as fixed across sites. Variables were dropped from the model only if two conditions were satisfied: The fixed effect for the variable was not significant and the variation of the random effect was not significant. If this situation occurred, the variable was deleted and the more parsimonious model with fewer predictor variables was fit to the data. This is the model building process recommended by Raudenbush and Bryk (2002).

Models with person-level and site-level predictor variables. Given that there was significant variation in the random effects of the previous models, site-level predictor variables were entered into the third level of the model in an attempt to account for such variation. The two site-level variables (the site sample size and education index) were selected because they were available, not necessarily because there is strong theory supporting their use. Therefore, inclusion of
such variables in the model was exploratory in nature. As such, t-to-enter statistics (Raudenbush & Bryk, 2002) were used as justification to include such variables in the third level of the model.

**Fit Indexes**

The goal of the model building process just described is to retain the most parsimonious model that can best describe the data. More complex models will always fit the data better than more simplistic models. The fit indexes of nested models can be compared to determine if a more complex model yields a significantly better fit than that of a more simplistic model. Typically the deviance statistic, which is a function of the likelihood, is used to compare the fit of nested models. Unfortunately, Snijders and Bosker (2000) warned against the use of the deviance statistic when multilevel logistic models are estimated with penalized quasi-likelihood (PQL) estimation. As such, to determine which model best fits the data in this study, the within-site error variance of the model with predictor variables, \( \tau_\beta \), was compared to that of the unconditional model to determine whether including person-level predictor variables reduced the within-site error variance. The same comparison was done with the between-site error variance, \( \tau_\gamma \).

**Software and Estimation**

The hierarchical generalized linear models (HGLM) component of the software program HLM 5 (Raudenbush, Bryk, & Congdon, 1999) was used to perform the analyses. For further information on how to use the HLM software to estimate multilevel IRT models, see Kamata (2002).

The estimation procedure used in the HGLM component of HLM 5 is PQL estimation (Breslow & Clayton, 1993), one of the most utilized estimation procedures for hierarchical generalized linear models (Guo & Zhao, 2000). The details of PQL estimation or how PQL compares to other estimation procedures is beyond the scope of this article; interested readers should consult Breslow and Clayton (1993), Guo and Zhao (2000), or Snijders and Bosker (2000) for further information.

**RESULTS**

**Unconditional model.** The partial results for the unconditional multilevel IRT model are given in Table 2. The item effect estimates are displayed in Figure 1, which will be discussed. There was statistically significant within-site, \( \tau_{\mu_1} \), and
between-site, $\tau_{g11}$, variation in ASE. Although the between-site variation was statistically different than zero, it was small in comparison to the within-site variation. The between-site variance accounted for 11% of the between- and within-site variance combined.

The fact that the fixed effect for the intercept, $\pi_{000}$, was significantly different than zero is not necessarily meaningful information. This effect, which also represents the estimate of ASE in the overall sample, or the effect of item 31 in the overall sample, is on the scale of the items used in the analysis. Whether or not it is different than zero then is arbitrary. To gain a better understanding of how the items relate to one another as estimated by this model, the fixed effect, $\pi_{000}$, of the reference item (item 31) was subtracted from the effect for each item, $\pi_{q00}$. The resulting values are in the scale of logit and therefore do not range from –3 to +3 as is typical with IRT difficulty parameters. The values are plotted in Figure 1. As one would expect, the reference item—selected for having the lowest classical difficulty index—required higher levels of ASE for endorsement than all other items.

To better understand how the sites relate to one another in terms of ASE, each of the empirical Bayes residuals, $r_{00g}$, which represent the offset of each site from the overall ASE estimate, were added to the overall ASE estimate, $\pi_{000}$, and plotted in Figure 2. The resulting values are the ASE estimate for each site.

**Age and gender effects model.** The overall and site residual plots together justified the use of age as a linear, rather than quadratic, effect in the following analyses.

The results for the final age and gender effects model are shown in Table 3. The final model yielded a significant fixed effect for gender ($\pi_{010} = –0.545$), indicating that controlling for age, male ASE and female ASE are significantly different from one another in the overall sample. The negative coefficient for this effect indicates
that female ASE is higher than male ASE. Because the effect for gender was fixed, this interpretation applies to all sites. Controlling for gender, there was also a significant fixed effect for age, with ASE significantly decreasing as the age of the adolescents in the overall sample increased. The significant random variation associated with this effect indicated that such an interpretation of the age effect did not apply to all sites.
To facilitate interpretation of the model, the fixed effects were used to obtain the probability of endorsing the reference item. The resulting values were plotted by age for males and females in Figure 3. This graph illustrates the relationship between age and ASE for males and females in the overall sample with a bold line. To illustrate the same relationships by site, the fixed effect estimates and empirical Bayes residuals ($r_{00g}$ and $r_{02g}$) were used to obtain the probability of endorsing the reference item for each site. These values are represented by the 13 nonbold lines in Figure 3. As can be seen from the bold line or the overall effect in the graph, the probability of endorsing a typical ASE item is 13.5% higher for females than males, at all ages. For both males and females, the probability of endorsing a typical ASE item decreases by 5% for each 1-year increase in age.
The 13 regression lines in Figure 3, show clearly what is meant by random intercepts and slopes. The probability of endorsing a typical ASE item at age 12 varies across sites, as illustrated by the 13 intercepts in Figure 3. The relationship between age and the probability of endorsing a typical ASE item also differs by site, with some sites having stronger negative relationships with age than other sites. The graphs indicate that the variation in slopes does not seem to be as dramatic as the variation in intercepts, which would be expected because the variance of the intercepts was estimated to be $\tau_{\gamma 11} = 0.234$ and the variance of the slopes was estimated to be $\tau_{\gamma 22} = 0.011$. All site slopes or age effects were negative, indicating that the probability of endorsing a typical ASE item decreases with age in all sites. Some sites, however, had almost null relationships with age and others had much stronger negative relationships with age. Including the predictor variables of age and gender reduced the person-to-person variation in ASE by only 8% when compared to the unconditional model.

Because the $t$-to-enter statistics for this model indicated that the site-level variable education index may have a relationship with the age effect estimates of the sites ($t = 1.303$), a subsequent model included education index as a predictor of the random age effect. The fixed effect for education index was significantly different from zero and positive ($\pi_{021} = 0.037, p = 0.015$), meaning that as the education level for a site increased, the relationship of age to ASE in a site became less negative, or closer to null. For illustration, the fixed effect estimates $\pi_{020}$ and $\pi_{021}$ were used to plot the relationship between the site age effect and the education index in Figure 4. Including education index as a predictor decreased the

---

**TABLE 3**

Results for the ASE Age and Gender Effects Model

<table>
<thead>
<tr>
<th>Estimates of Fixed Effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>df</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_{000}$ Intercept</td>
<td>0.421</td>
<td>0.222</td>
<td>12</td>
<td>1.896</td>
<td>.082</td>
</tr>
<tr>
<td>$\pi_{010}$ Gender</td>
<td>-0.545</td>
<td>0.110</td>
<td>902</td>
<td>-4.983</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>$\pi_{020}$ Age</td>
<td>-0.182</td>
<td>0.052</td>
<td>12</td>
<td>-3.495</td>
<td>.005</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimates of Random Effects</th>
<th>Reliability</th>
<th>Variance</th>
<th>SE</th>
<th>df</th>
<th>$\chi^2$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau_{p11}$ Intercept</td>
<td>0.692</td>
<td>1.825</td>
<td>0.124</td>
<td>878</td>
<td>2463.15</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Level 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau_{\gamma 11}$ Intercept</td>
<td>0.337</td>
<td>0.234</td>
<td>0.177</td>
<td>12</td>
<td>26.24</td>
<td>.010</td>
</tr>
<tr>
<td>$\tau_{\gamma 22}$ Age</td>
<td>0.262</td>
<td>0.011</td>
<td>0.012</td>
<td>12</td>
<td>30.80</td>
<td>.002</td>
</tr>
</tbody>
</table>

*Note.* ASE = academic self-esteem.
FIGURE 3  ASE age effect for females and males by site. Overall effect is indicated by bold line in each graph. ASE = academic self-esteem.
variance of the age effects ($\tau_{\gamma^2} = 0.006$, $p = .015$) across sites by 45% when compared to the previous model.

It should also be noted that the item fixed effects remained fairly constant across models, as would be expected. Kamata (1998) noted that including predictor variables in the multilevel IRT model should not alter the item effect estimates. This study lends support to such a conclusion. As can be seen in Table 4, there was little difference between the fixed effect estimates for the items

![FIGURE 4](image)

**FIGURE 4** Relationship of site age effect with education index.

<table>
<thead>
<tr>
<th>Item Name</th>
<th>Unconditional Model</th>
<th>Age and Gender Effects Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 13</td>
<td>−1.121</td>
<td>−1.127</td>
</tr>
<tr>
<td>Item 19</td>
<td>−1.540</td>
<td>−1.547</td>
</tr>
<tr>
<td>Item 50</td>
<td>−0.708</td>
<td>−0.711</td>
</tr>
<tr>
<td>Item 54</td>
<td>−1.214</td>
<td>−1.220</td>
</tr>
<tr>
<td>Item 66</td>
<td>−1.288</td>
<td>−1.294</td>
</tr>
<tr>
<td>Item 67</td>
<td>−1.508</td>
<td>−1.514</td>
</tr>
<tr>
<td>Item 68</td>
<td>−1.530</td>
<td>−1.535</td>
</tr>
</tbody>
</table>
DISCUSSION

Using a model-building process, this study illustrated how a three-level IRT model could be used to obtain item parameters for the overall sample, as well as latent traits estimates for both persons and sites. The model was also able to partition the total variance of the latent trait into person- and site-level components, indicating that although most of the variance in ASE was between persons, a substantial amount existed between sites (11%). The models were used further to examine the relationships between person-level predictor variables and ASE, allowing such relationships to vary across sites. Although the differences between males and females in ASE did not vary from site to site, the average ASE at age 12 and the relationship of age with ASE did vary. This finding encourages the use of multilevel modeling in the study of ASE. If multilevel modeling had not been used, the conclusion would have been that ASE decreases as age increases in all sites, when in fact some sites indicate no relationship between ASE and age. Use of a multilevel model also allowed the effects of the variables of the model to vary randomly across sites, and an interesting relationship was found between a site’s ASE age effect and the site’s education level. As the average education of the persons within a site increases, the age effect for a site goes from being negative to null. This implies that ASE declines with age more sharply for adolescents in sites with low education levels and is more constant across age in sites with high education levels. The reason behind this relationship can only be speculated and is a cause for future research.

Although the complexity of the multilevel IRT model allows researchers to better model reality and to easily pursue a variety of interesting research questions, it is not without the requirement that several assumptions be satisfied. For instance, the use of a multilevel IRT model requires satisfying the assumptions of both the IRT model and the multilevel model. The models in this study were based on the Rasch IRT model, which assumes equal discrimination across items. It was thought that this assumption was satisfied in this study (see Pastor, 2001), but other measures may require use of a 2-PL or polytomous IRT model, in which case the model would have to be respecified. Moreover, IRT models assume that there is a single dimension underlying the responses to all items and that the dimension accounts for the relationship between the responses for a given person (Hambleton et al., 1991). A single dimension underlying the item responses was assumed in this study (see Pastor, 2001); however, tests or scales with multiple dimensions are quite common and can be easily accommodated by Kamata’s multilevel IRT model (Kamata, 1998) using the HLM software.
In addition to satisfying the assumptions of IRT and the particular IRT model chosen, multilevel IRT models require the assumptions that the latent trait is a random parameter, that the latent trait is normally distributed within and between sites, and that the variance of the latent trait is the same across sites. It is possible that nonparametric extensions of the model could be developed for situations in which these assumptions cannot be met. Maier (2001) recommended the use of fully Bayesian estimation procedures for situations in which the assumption of normality and a minimum allowable sample size are of concern. In fact, a fully Bayesian approach utilizing Markov chain Monte Carlo (MCMC) techniques has been recommended and used in the estimation of multilevel IRT models by several authors (Johnson, 2002; Maier, 2001; Patz & Junker, 1999; Patz & Junker, 2000). Unfortunately, estimation using MCMC techniques can be time-consuming and often requires the use of specialized software. For this reason, future researchers may want to pursue how robust the three-level IRT model is to assumption violations, as estimated using the HLM software.

The HLM software used in this study utilized PQL for estimation of the random effects. Snijders and Bosker (2000) are critical of such an estimation technique, arguing that the algorithms for PQL are not very stable; whether or not algorithms converge may be dependent upon the data set, the complexity of the model, and the starting values. According to Snijders and Bosker (2000), the algorithms have the two additional disadvantages of (a) providing only a crude approximation of the deviance statistic and (b) estimating the variance parameters for the random effect with high mean squared error. Use of the Laplace approximation (Raudenbush, Yang, & Yosef, 2000) has been suggested to overcome these problems with PQL, although its implementation did not aid convergence of preliminary models in this study nor did it produce results substantially different than those reported. Of course, many of the convergence problems encountered in this study could be attributed not to the estimation technique, but in trying to fit a model too complex for the available data. Future researchers may want to investigate how large a sample is needed to ensure accurate estimation at the various levels for both unidimensional and multidimensional three-level IRT models.

In regard to the sample used in this study, the sample sizes in some sites and the sparsity of ages within some sites may seem too low for proper estimation of the model. However, the reader should be reminded of the prior discussion regarding the borrowing of strength advantage in multilevel modeling. If a site has a low sample size, then the individual estimate for that site would be “shrunk” toward the effect in the overall sample. Ideally, of course, if would be desirable to have a large number of sites, a large number of participants per site, and a wide range of ages represented within each site to ensure that no one site was unduly influencing the results.

In applying these models, it was also found that the choice of reference item in the multilevel IRT model does have an impact on the results. The estimates of the
fixed effects for person-level and site-level predictor variables, the estimates of the random effects, and the rank order and distance between item parameters are not affected by which item is chosen as the reference item. What is affected is the interpretation of the log-odds for a particular respondent. For instance, the results for the fixed effects between gender and age for ASE are plotted twice in Figure 5. In the top part of the figure, the easiest item was chosen as the reference item. In the bottom part of the figure, the hardest item was chosen as the reference item. As can be seen from the figure, the relationship is the same between gender, age, and ASE when using the easiest item as the reference item compared to the hardest item, but the scale is not. For instance, the log-odds of 12-year-old female endorsement is different when the reference item is item 33 than when the reference item is item 35. Therefore, conclusions regarding the parameters of the model will not differ regardless of which item is used as the reference item. However, when using the parameter estimates to make conclusions regarding the probability or log-odds of a particular respondent endorsing an item, the probability or log-odds will vary depending upon the reference item chosen. Researchers using these models need to be aware of this difference when interpreting and plotting their results.

The possible applications of Kamata’s (1998, 2001) multilevel IRT model as well as other multilevel IRT models extend far beyond how Kamata’s model was used in the current study. For instance, item effects can be allowed to vary randomly across persons or sites, thereby allowing investigation of the IRT item parameter invariance assumption. Although few, there are studies that have used the model for such a purpose (Beretvas & Williams, 2002; Cheong, 2001; Kamata, 1998; Luppescu, 2002).

Three-level IRT models have a promising future in applied research. It is hoped that research investigating the various applications of the model is supplemented

![Figure 5](image.png)

**FIGURE 5** Comparison of ASE age effect by gender in overall sample using the easiest and hardest items as reference items. ASE = academic self-esteem.
with simulation studies investigating the model’s behavior under various conditions (i.e., various sample sizes, estimation techniques). Although many applications have been proposed for the three-level IRT model, only the results from simulation studies can identify the applications for which the use of the model is best suited.

REFERENCES


