JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Spring 2011, Vol. 48, No. 1, pp. 61–80

When Does Scale Anchoring Work? A Case Study

Sandip Sinharay, Shelby J. Haberman, and Yi-Hsuan Lee
Educational Testing Service

Providing information to test takers and test score users about the abilities of test takers at different score levels has been a persistent problem in educational and psychological measurement. Scale anchoring, a technique which describes what students at different points on a score scale know and can do, is a tool to provide such information. Scale anchoring for a test involves a substantial amount of work, both by the statistical analysts and test developers involved with the test. In addition, scale anchoring involves considerable use of subjective judgment, so its conclusions may be questionable. We describe statistical procedures that can be used to determine if scale anchoring is likely to be successful for a test. If these procedures indicate that scale anchoring is unlikely to be successful, then there is little reason to perform a detailed scale anchoring study. The procedures are applied to several data sets from a teachers' licensing test.

Testing companies are under constant pressure to report information in addition to the overall test score. Subscores (e.g., Sinharay, Haberman, & Puhan, 2007) can be reported in addition to the overall score. It is also possible to report, especially when subscores cannot be reported, information concerning the tasks that examinees at specified score levels are customarily able to perform. Although such information might appear to be readily supplied, in practice the task has been a problem in educational and psychological measurement (Carroll, 1993). Testing companies have been investigating solutions to this problem through the development of proficiency scaling procedures and question-difficulty research. Scale anchoring (Beaton & Allen, 1992) is a tool to provide such information concerning the relationship between tasks examinee can perform and observed test scores. It results in descriptions of what students at different points on a score scale know and can do. Scale anchoring produces *performance-level descriptors* (Perie, 2008), which describe the level of knowledge and skills required of different performance levels. For example, a scale anchoring study for the Reading section of the Test of English as a Foreign Language™ Internet-based Test (Garcia Gomez, Noah, Schedl, Wright, & Yolkut, 2007) found, among other things, that the test-takers who obtain a high score (22–30) in the section typically have an excellent command of academic vocabulary and grammatical structure. Scale anchoring has been used with a variety of assessments, including the National Assessment of Educational Progress (Beaton & Allen, 1992) and the Trends in International Mathematics and Science Study (Kelly, 2002).

Though scale anchoring appears promising, it is often problematic. Linn and Dunbar (1992) described the confusion of the public about the meaning of National Assessment of Educational Progress data related to score anchors. They concluded that the reasons for the discrepancy between the percentage of examinees who answer an

anchor item correctly and the percentage who score above the corresponding anchor point may be too subtle for mass communication. Phillips et al. (1993) described the danger of overinterpreting examinee performance at anchor points so that all examinees at a particular level are assumed to be proficient at all abilities measured at that level.

A more basic issue for scale anchoring is the question whether assessments, especially when applied to examinees with roughly comparable proficiency, support the underlying hypothesis that examinees at specific levels can be assumed to be able to perform selected tasks but not able to perform other tasks. An examination in mathematics given to students from ages 6 to 20 would likely find score points below which students were unlikely to solve elementary calculus problems and would likely find score points above which students might succeed in solving elementary calculus problems. Were this examination administered to undergraduate students at the Massachusetts Institute of Technology, this division by ability to solve elementary calculus problems would probably not be available. In typical assessments, the variability of student proficiency is relatively limited, so that the reality will resemble the latter situation.

The steps required in scale anchoring are the following:

1. Select a few dispersed points on the score scale (*anchor points*) that will be anchored.
2. Find examinees who score near each anchor point.
3. Examine each item to see if it discriminates between successive anchor points, that is, if most of the students at the higher score levels can answer it correctly and most of the students at the lower level cannot. The definition of "most" is subjective and depends on the investigator.
4. Review the items that discriminate between adjacent anchor points to find out if specific tasks or attributes that they include can be generalized to describe the level of proficiency at the anchor point. The outcome from this review is a description of what students at various scale points know and can do.

The above description shows that the first three steps of scale anchoring constitute a statistical component that identifies items that discriminate between successive points on the proficiency scale using specific item attributes (Beaton & Allen, 1992). These steps are closely related to the common process of item mapping (Zwick, Senturk, Wang, & Loomis, 2001). The fourth step involves generalizations not required in item mapping and involves a consensus component in which identified items are used by subject-area and educational experts to provide an interpretation of what groups of students at or close to the selected scale points know and can do. This component can be costly (because of the involvement of subject-area and educational experts) and can be time-consuming. In addition, the subjective judgment may not be accurate.

As Beaton and Allen (1992) noted, the scale anchoring process is not guaranteed to result in useful descriptions of the anchor points. A test that is well designed for its intended purpose may not have sufficient information available to differentiate between performance of examinees at given score levels on items with different attributes. In some cases, this failure may reflect the lack of a sufficient number of

items anchoring at given score levels. It may also be true that the items at an anchor level are too dissimilar to interpret.

Therefore, before performing an exhaustive scale anchoring study, it may be beneficial if a set of simple statistical analyses can be performed to find out if scale anchoring will provide useful information to the examinees (in other words, if scale anchoring will be successful). This paper suggests such a set of analyses including linear regression analysis and fitting of several popular item response theory models. Our suggested techniques are discussed in the next section. The techniques are applied to several data sets from a teachers' licensing test in the application section. Conclusions and recommendations are provided in the last section.

Methods

Scale Anchoring and Prediction of Item Statistics from Item Attributes

The description of scale anchoring indicates that scale anchoring can provide useful information if item attributes can predict item difficulties to an adequate degree and if item discriminations associated with these item attributes are high. Suppose the item attributes do not predict item difficulties well. Then the items discriminating between adjacent anchor points, which are of similar difficulty, will have different attributes. In that case, scale anchoring will not provide any useful information. Unless item discriminations are consistently high, it is also necessary for item attributes to predict item discrimination. Otherwise, in Step 4 of the description of scale anchoring, the required generalizations will not be feasible. Hence, the key to the techniques suggested later in this paper is an examination of how well item attributes predict item difficulties and item discriminations. What follows next is a mathematical proof of why it is necessary and sufficient for the item attributes to predict item difficulties and item discriminations for scale anchoring to provide useful information.

Let X_{is} denote the response of Examinee s, $1 \leq s \leq n$, to Item i, $1 \leq i \leq m$. Let θ_s denote the latent proficiency parameter of Examinee s. We assume θ_s to have a standard normal distribution. Suppose that conditional on θ_s, the X_{is} are mutually independent and the probability that $X_{is} = 1$ is

$$\frac{\exp(a_i\theta_s - \beta_i)}{1 + \exp(a_i\theta_s - \beta_i)},$$

that is, the two-parameter logistic (2PL) model holds for the data. For Item i, a_i is the discrimination parameter and β_i the intercept parameter. If $a_i > 0$, then $b_i = \beta_i/a_i$ is the difficulty parameter of Item i.

Let us consider an item that anchors at $\theta_s = \omega$. Then, from the earlier description of scale anchoring, the probability of a correct response is at least p_1 for $\theta_s = \omega$ and no more than p_2, $p_2 < p_1$ for $\theta_s = \upsilon < \omega$. That means

$$a_i\omega - \beta_i \geq \log[p_1/(1 - p_1)],$$

and

$$a_i \upsilon - \beta_i \leq \log[p_2/(1-p_2)].$$

The above inequalities imply that the discrimination parameter a_i must be at least

$$\frac{\log[p_1/(1-p_1)] - \log[p_2/(1-p_2)]}{\omega - \upsilon}, \tag{1}$$

which indicates that an item with a low discrimination parameter may not anchor at all. Given $a_i > 0$, the intercept parameter β_i must be between

$$a_i \upsilon - \log[p_2/(1-p_2)]$$

and

$$a_i \omega - \log[p_1/(1-p_1)],$$

so that the difficulty parameter b_i must be between

$$\upsilon - \frac{1}{a_i} \log[p_2/(1-p_2)]$$

and

$$\omega - \frac{1}{a_i} \log[p_1/(1-p_1)].$$

Suppose the item discrimination parameter a_i is sufficiently large that Equation 1 holds. For example, if $\omega = 0.5$, $\upsilon = 0$, $p_1 = 0.6$, and $p_2 = 0.4$, then the discrimination must be at least 1.62. In addition, unless a_i is somewhat larger than 1.62, the interval for the item difficulty will be very narrow. If scale anchoring is informative for this data set, that means that this item and a few other items that anchor at $\theta_s = \omega$ possess a few specific item attributes. That phenomenon, along with the above-mentioned bounds of the parameters of an item anchoring at ω, implies that these item attributes determine the above-mentioned bounds on a_i and b_i, or, in other words, that the item attributes predict item discrimination and item difficulty. On the other hand, if the item attributes predict item discrimination and item difficulty adequately, the above-mentioned bounds will be associated with a few specific item attributes; these item attributes are then associated with $\theta_s = \omega$, which means that scale anchoring provides useful information for this data set. Thus, a necessary and sufficient condition for scale anchoring to be informative is that item attributes predict item discrimination and item difficulty adequately. In typical cases, adequacy involves item difficulty more than item discrimination, for the requirement on item discrimination involves sufficiently high discrimination while the requirement on difficulty involves falling within the proper range.

Existing Research on Prediction of Item Statistics from Item Attributes

There has been substantial research on prediction of item discrimination and item difficulty from item attributes. Linear regression models and tree-based regression models have been applied to examine prediction of item difficulty from item attributes for several tests such as Praxis™, Graduate Record Examinations®, and the Grade 8 Reading Assessment of the National Assessment of Educational Progress (e.g., Gorin & Embretson, 2006; Sheehan, Kostin, & Persky, 2006; Sheehan & Mislevy, 1994; Wainer, Sheehan, & Wang, 2000). These studies show a low to moderate amount of success in predicting item difficulty from item attributes. Sheehan and Mislevy (1994) reported that item attributes explained between 20% and 40% of the variance in item difficulty and between 4% and 14% of the variance in item discrimination for 510 pretest items from a Praxis-I test that measures mathematics, reading, and writing. Sheehan et al. (2006) reported that item attributes explained between 14% and 50% of the variance in item difficulty for the Grade 8 Reading Assessment of the National Assessment of Educational Progress. Gorin and Embretson (2006) reported that their rigorously created cognitive variables predicted between one-fourth to one-third of the variance in item difficulty of 200 disclosed Graduate Record Examinations Verbal paragraph comprehension items. The reported values were not examined by either cross-validation or by rigorous statistical analysis designed to adjust for the occasional tendency of regression methods to include more predictors than what is appropriate (see, for example, Draper & Smith, 1998, pp. 342–343, for a description of this tendency).

Methods to Determine if Scale Anchoring Will Provide Useful Information

We suggest two sets of methods: (i) fitting of linear regression models; and (ii) fitting of item response theory models. These methods will examine if the item attributes adequately predict item statistics, which, according to the earlier discussion, is equivalent to examining if scale anchoring will provide useful information. These methods assume availability of test data concerning item attributes—these are usually variables used in test development to characterize items in terms of features such as domain covered or type of tasks covered. Such variables are usually available because test developers use them to create test forms that conform to specifications.

Fitting of linear regression models. The first technique that can be used to determine if scale anchoring will provide useful information is linear regression of item statistics (item difficulty or item discrimination) on indicators of appropriate item attributes (for examples of such analyses see, e.g., Gorin & Embretson, 2006; Sheehan & Mislevy, 1994). The squared multiple correlations from these regressions will provide an idea of how well the item statistics can be predicted by the item attributes. If there are too many item attributes, then, rather than including all of them in the regression model, one should use a technique such as stepwise regression to include only the relevant attributes. In addition, a technique such as cross-validation (e.g., Neter, Kutner, Nachtsheim, & Wasserman, 1996, p. 437) should be used, especially if the sample size (which is the number of items in our analyses) is small, because, for example, stepwise regression tends to admit more predictors than is appropriate (see, for example, Draper & Smith, 1998, pp. 342–343). The Application section of this paper employs cross-validation.

Fitting of item response theory models. The second set of techniques involve fitting of several item response theory models to the data. In the operational data examples considered later, all items are right-scored. Associated with Item i are item attributes q_{ik}, $1 \leq k \leq K$, for some integer $K \geq 1$. The q_{ik} are indicator variables, with $q_{ik} = 1$ if Attribute k is present for Item i and $q_{ik} = 0$ otherwise. All models considered here are special cases of the 2PL model mentioned earlier. We also consider the Rasch model, for which the discrimination parameter a_i is assumed to be the same for all items. In the zero-parameter logistic (0PL) model, it is assumed that both the discrimination parameter a_i and the intercept parameter β_i are the same for all items. This model is included as a baseline model for comparison of other models (for example, the Rasch or 2PL models).

The 0PL, Rasch and 2PL models do not use the indicator variables q_{ik}. In several models, these indicators are employed to predict item parameters. In the linear logistic test model (LLTM; Fischer, 1973), the Rasch model is assumed, and it is assumed that the item intercept satisfies a linear model

$$\beta_i = \eta_1 q_{i1} + \eta_2 q_{i2} + \cdots + \eta_K q_{iK} = \sum_{k=1}^{K} \eta_k q_{ik} \tag{2}$$

in which η_k represents the effect of Attribute k on the intercept β_i of Item i. The LLTM reduces to the 0PL model if $K = 1$ and $q_{i1} = 1$ for all i. The LLTM is the same as the Rasch model if $K = m$ and the m by K matrix $\mathbf{Q}$ of q_{ik} has rank m. The LLTM has two generalizations to 2PL models. In the constrained 2PL model (Embretson, 1993), it is assumed that the difficulty

$$b_i = \sum_k \gamma_k q_{ik} \tag{3}$$

of Item i satisfies a linear model in which γ_k represents the effect of Attribute k and the item discrimination satisfies a linear model

$$a_i = \sum_k \tau_k q_{ik} \tag{4}$$

in which τ_k represents the effect of Attribute k. If $q_{i1} = 1$ for all i and $K = 1$, then the constrained 2PL (C2PL) model reduces to the 0PL model. In the alternative constrained 2PL (AC2PL) model, the relation given by Equation 2 is assumed to hold for some η_k and Equation 4 is assumed to hold for some τ_k. The models and the number of parameters in them are listed in Table 1.

To compare models, the information-theoretic measure *Minimum Estimated Expected Log Penalty Per Item* (see, e.g., Gilula & Haberman, 2001; Haberman, 2006), henceforth referred to as Penalty, may be employed. This is a popular measure for comparing models and is based on the logarithmic penalty function developed by

Table 1
A List of the Item Response Theory Models

Model	Number of Parameters	Mathematical Expression
0PL	2	$\mathrm{P}(X_{is}=1)=\frac{\exp(a\theta_s-\beta)}{1+\exp(a\theta_s-\beta)}$
Rasch	m+1	$\mathrm{P}(X_{is}=1)=\frac{\exp(a\theta_s-\beta_i)}{1+\exp(a\theta_s-\beta_i)}$
2PL	2m	$\mathrm{P}(X_{is}=1)=\frac{\exp(a_i\theta_s-\beta_i)}{1+\exp(a_i\theta_s-\beta_i)}$
LLTM	K+1	$\mathrm{P}(X_{is}=1)=\frac{\exp(a\theta_s-\beta_i)}{1+\exp(a\theta_s-\beta_i)}$, $\beta_i=\sum_k \eta_k q_{ik}$
C2PL	2K	$\mathrm{P}(X_{is}=1)=\frac{\exp(a_i\theta_s-\beta_i)}{1+\exp(a_i\theta_s-\beta_i)}$, $b_i=\beta_i/a_i=\sum_k \gamma_k q_{ik}$, $a_i=\sum_k \tau_k q_{ik}$
AC2PL	2K	$\mathrm{P}(X_{is}=1)=\frac{\exp(a_i\theta_s-\beta_i)}{1+\exp(a_i\theta_s-\beta_i)}$, $\beta_i=\sum_k \eta_k q_{ik}$, $a_i=\sum_k \tau_k q_{ik}$

Savage (1971). For a model, the Penalty is obtained as

$$\text{Penalty}_{\text{model}}=-\frac{\ell}{2nm},$$

where ℓ is the maximum log-likelihood under the model. For example, $\text{Penalty}_{\text{Rasch}}$ is the penalty for the Rasch model and $\text{Penalty}_{\text{2PL}}$ is the penalty under the 2PL model. Note that as the likelihood increases, the penalty decreases. Among the models under study,

$$\text{Penalty}_{\text{0PL}}\geq\text{Penalty}_{\text{LLTM}}\geq\text{Penalty}_{\text{Rasch}}\geq\text{Penalty}_{\text{2PL}},$$

$$\text{Penalty}_{\text{0PL}}\geq\text{Penalty}_{\text{LLTM}}\geq\text{Penalty}_{\text{C2PL}}\geq\text{Penalty}_{\text{2PL}},$$

$$\text{Penalty}_{\text{0PL}}\geq\text{Penalty}_{\text{LLTM}}\geq\text{Penalty}_{\text{AC2PL}}\geq\text{Penalty}_{\text{2PL}},$$

and

$$\text{Penalty}_{\text{Rasch}}\geq\text{Penalty}_{\text{2PL}},$$

because, for example, the 0PL model is a special case of the LLTM, the LLTM is a special case of the Rasch model, and the Rasch model is a special case of the 2PL model. An LLTM is most attractive if $\text{Penalty}_{\text{LLTM}}$ is close to $\text{Penalty}_{\text{AC2PL}}$, $\text{Penalty}_{\text{C2PL}}$, $\text{Penalty}_{\text{Rasch}}$, and $\text{Penalty}_{\text{2PL}}$. The constrained 2PL model is most attractive if $\text{Penalty}_{\text{C2PL}}$ is close to $\text{Penalty}_{\text{2PL}}$, and the alternative constrained 2PL model is most attractive if $\text{Penalty}_{\text{AC2PL}}$ is close to $\text{Penalty}_{\text{2PL}}$. Evaluation of closeness can be

considered in terms of relative improvement in penalty and in terms of improvements in penalty per independent parameter which are described below.

If Model 1 is a special case of Model 2, but the models are not equivalent, then the improvement in penalty per independent parameter from Model 1 to Model 2 is

$$\text{Improvement}_{1-2} = \frac{\text{Penalty}_1 - \text{Penalty}_2}{\text{Number of parameters in Model 2} - \text{Number of parameters in Model 1}}.$$

Larger values of Improvement_{1-2}, which indicate that the penalty is much less (or, in other words, likelihood is much more) for Model 2 than Model 1, indicate that Model 2 fits the data considerably better than does Model 1. So, for example, if one is interested to know whether the 2PL model does substantially better than the Rasch model, one can compute Improvement_{1-2}, where the Rasch and 2PL models play the role of Models 1 and 2, respectively, in the above formula; a large value of Improvement_{1-2} would favor the 2PL model over the Rasch model.

In some applications, one might have a situation such that Model 1 is a special case of Model 2, Model 2 is a special case of Model 3, and the goal is to compare the performance of Model 2 and Model 3, where Model 1 provides a baseline for comparison of Model 2 and Model 3. An example would be an evaluation of the LLTM compared to the Rasch model where the 0PL model provides a baseline for comparison. In this case, one may examine Improvement_{1-2} and Improvement_{2-3}; somewhat larger value of Improvement_{1-2} than Improvement_{2-3} would indicate that the Model 2 performs almost as well as Model 3 in terms of reduction of penalty from Model 1 and would favor Model 2. One may also examine the relative improvement

$$\text{RI}_{1-2-3} = \frac{\text{Penalty}_1 - \text{Penalty}_2}{\text{Penalty}_1 - \text{Penalty}_3}$$

to know how Model 2 compares to Model 3. The closer the value of RI_{1-2-3} to 1, the better is the performance of Model 2 compared to Model 3. It is certainly desired that RI_{1-2-3} be somewhat larger than

$$\frac{\text{Number of parameters in Model 2} - \text{Number of parameters in Model 1}}{\text{Number of parameters in Model 3} - \text{Number of parameters in Model 1}},$$

which can be shown to be equivalent to the requirement that the gain per independent parameter from Model 1 to Model 2 is somewhat larger than is the gain per independent parameter from Model 2 to Model 3.

For example, consider an evaluation of the LLTM in comparison to the Rasch model where the 0PL model provides a baseline for comparison. Assume that $0 < K < m$. For the LLTM to be favored, it is desirable that

$$\text{Improvement}_{\text{0PL}-\text{LLTM}} = \frac{\text{Penalty}_{\text{0PL}} - \text{Penalty}_{\text{LLTM}}}{K - 1}$$

be somewhat larger than is

$$\text{Improvement}_{\text{LLTM}-\text{Rasch}} = \frac{\text{Penalty}_{\text{LLTM}} - \text{Penalty}_{\text{Rasch}}}{m - K},$$

which would indicate that the LLTM can account for a substantial part of the difference between the 0PL model and the Rasch model. For the LLTM to be favored, it is also desirable that

$$\text{RI}_{\text{0PL}-\text{LLTM}-\text{Rasch}} = \frac{\text{Penalty}_{\text{0PL}} - \text{Penalty}_{\text{LLTM}}}{\text{Penalty}_{\text{0PL}} - \text{Penalty}_{\text{Rasch}}}$$

be close to 1 and somewhat larger than $(K - 1)/(m - 1)$. Results favorable to LLTM suggest some ability to predict item difficulty by use of item attributes. Similar arguments can be applied to the constrained 2PL model or the alternate constrained 2PL model. In the case of the former, it is desirable that

$$\text{Improvement}_{\text{0PL}-\text{C2PL}} = \frac{\text{Penalty}_{\text{0PL}} - \text{Penalty}_{\text{C2PL}}}{2(K - 1)}$$

be somewhat larger than

$$\text{Improvement}_{\text{C2PL}-\text{2PL}} = \frac{\text{Penalty}_{\text{C2PL}} - \text{Penalty}_{\text{2PL}}}{2(m - K)}.$$

It is also desirable that

$$\text{RI}_{\text{0PL}-\text{C2PL}-\text{2PL}} = \frac{\text{Penalty}_{\text{0PL}} - \text{Penalty}_{\text{C2PL}}}{\text{Penalty}_{\text{0PL}} - \text{Penalty}_{\text{2PL}}}$$

be close to 1 and somewhat larger than $(K - 1)/(m - 1)$. Favorable results suggest some ability to predict item difficulty and item discrimination from item attributes.

In principle, it is possible to apply chi-square tests to compare models. Let Model 1 be a special case of Model 2, and let Models 1 and 2 not be equivalent. If Model 1 holds, then the likelihood-ratio chi-square statistic $L_{12}^2 = 2nm(\text{Penalty}_1 - \text{Penalty}_2)$ has an approximate chi-square distribution with degrees of freedom equal to

$$\text{Number of parameters in Model 2} - \text{Number of parameters in Model 1}.$$

In large samples, L_{12}^2 will be quite large even if the deviation of Model 1 from the data is small, so that this approach is not very helpful in practice. In all cases in this report, L_{12}^2 is highly significant—so they are not discussed henceforth. The Penalty statistic, which is used in this paper, does not have any such limitations and performs well even for large samples.

Steps Involved in Analysis of a Data Set

Our recommended approach to determine if scale anchoring is likely to provide useful information for a test, which involves examining if the item attributes predict item statistics successfully, involves the following steps:

- Perform a linear regression of item difficulty on the indicators of item attributes and a linear regression of item discrimination on the indicators of item attributes. Use stepwise regression and cross-validation as necessary. High values of squared multiple correlations from these regressions will imply that scale anchoring is likely to provide useful information for the test.
- Compare the LLTM to the Rasch model where the 0PL model provides a baseline for comparison and compare the constrained 2PL model (and the alternative constrained 2PL model) to the 2PL model where the 0PL model provides a baseline for comparison. Good performance by the LLTM and the constrained 2PL model (and the alternative constrained 2PL model) implies that scale anchoring is likely to provide useful information for the test.

Application

Data from a Scale Anchoring Study

A scale anchoring study was recently performed using four forms of a teachers' licensing test in mathematics. The lowest and highest possible scaled scores for the test are 150 and 190. The four anchor levels considered were: 150 to 168, 169 to 173, 174 to 178, and 179 to 190. The score 169 is the least passing score among the states that use the test, 178 is the largest passing score, and 173 and 174 lie approximately midway between the lowest and highest passing scores.

For the anchor level i, $i = 2, 3, 4$, an item anchored if:

- At least 65% of examinees scoring in the range defined by the anchor level i answered the item correctly.
- At most 50% of examinees scoring in the range defined by the anchor level $i - 1$ answered the item correctly.

Because the above criteria led to few items being anchored, items that meet a less stringent set of criteria were also identified. The criteria to identify items that "almost anchored" were the following:

- At most 60% of examinees scoring in the range defined by the anchor level $i - 1$ answered the item correctly.
- The difference between the percent of examinees in the range defined by anchor level i that answered the item correctly and the percent of examinees in the range defined by anchor level $i - 1$ that answered the item correctly is at least 15%.

To further supplement the pool of items, those that met only the criterion of at least 65% of the students answered correctly (regardless of the performance of examinees at the next lower level) were identified. The three categories of items, shown

Table 2
Number of Items that Anchored

Anchor Level	Anchored	Almost Anchored	Met the 65% Criterion	Total
2	7	8	14	29
3	2	6	13	21
4	25	22	10	57
Total	34	36	37	107

Note. The total number of items in the four forms is 160.

in Table 2, ensure that there were enough items available to inform the descriptions of examinee achievement at the anchor levels.

The next step was the consensus component where the subject-area experts (that is, the test developers) reviewed the items that anchored and tried to interpret the results.

The outcomes of the scale anchoring procedure were statements such as that the examinees in anchor level 2 can (i) order positive integers, (ii) follow simple directions (2 steps or fewer), etc. The participants of the consensus component of the study found the component to be tedious and they often struggled to come up with a meaningful list of skills at any anchor level.

Results

Test developers classify each item in the test into one of two classifications based on item type (pure or real) and one of five classifications based on item content (algebra, data analysis and probability, geometry, measurement, numbers and operations). These classifications, along with several other classifications, are used by the test developers to assemble test forms that conform to specifications. We had the item type and item content classifications available for all items in Forms 1 to 4. In addition, for only one of the four test forms (referred to as Form 1), we obtained a table that shows a list of 63 attributes (for example, one attribute is whether the item has a stimulus such as a table/figure or not) and the attributes (out of these 63) that apply to each item ($q_{ik}s$)—the content experts created this table during the scale anchoring procedure.

Results from Fitting of Linear Regression Models

We fitted the 2PL model to data from Forms 1 to 4. Then, for each form, we used a linear regression model to predict the 40 estimated item difficulty parameters $\hat{b}_i$ and the estimated item discrimination parameters $\hat{a}_i$ from indicators of the item type and item content classifications. To avoid linear dependence of indicator variables, only five of the seven indicator variables plus a constant predictor can be employed. The regression model performed poorly. The F statistics provided no indication that any relationship between the dependent variables and the indicator variables existed. The squared multiple correlation coefficient R^2 ranged between 0.05 and 0.16 for the model predicting estimated item difficulty for the four forms, and between 0.03 and 0.30 for the model predicting estimated item discrimination. Similar results are

obtained if, instead of the estimated difficulty and discrimination parameters, item proportions correct and item R-biserial correlations are used as the response variables in the regressions. Note that, if item type and item content classification have no effect on the dependent variable and if the dependent variable is normally distributed, then the R^2 statistic has a beta distribution with parameters 5/2 and 17 (e.g., Seber, 1977, p. 423), and hence has a mean of $5/39 = 0.13$ and a standard deviation of

$$\left[\frac{(5/2)[(39-5)/2]}{(39/2)^2(1+39/2)}\right]^{1/2} = 0.07,$$

the probability is 0.95 that R^2 is no greater than 0.27, and the probability is 0.99 that R^2 is no greater than 0.35 (Rao, 1973, ch. 3). Thus no evidence exists that the item type and item content classifications are useful in predicting the four item statistics—estimated item difficulty, estimated item discrimination, proportion correct, and R-biserial correlation. This conclusion reflects two considerations. An R^2 of 0.3 or less does not indicate much ability to predict an item attribute. In addition, in view of the eight R^2 statistics examined, the fact that the largest is about 0.30 provides no clear evidence that any relationship at all exists between item difficulty and item discrimination on the one hand and the item content and item type attributes on the other hand.

For Form 1, we performed a stepwise linear regression (Draper & Smith, 1998, ch. 15) to predict the estimated item difficulty parameters and the estimated item discrimination parameters from the indicators of the 63 item attributes. The trivial indicator function with value one for all items was always included. Variables were added one by one to the model only if the F statistic for a variable was significant at the 0.15 level (the default value in SAS® for stepwise linear regression). The same criterion was used for removal of variables. At first glance, the results might appear more promising than for the regressions on item type and item content classifications. The algorithm picked six nontrivial attributes out of the possible 63 in predicting the estimated item difficulty parameters, and the resulting R^2 statistic was 0.43. In the case of item discrimination parameters, eight nontrivial item attributes were chosen, and the resulting R^2 was 0.64. The number of nontrivial item attributes that were included both in the final model for item discrimination and the final model for item difficulty was only one.

High R^2 values in regression are often a deceptive artifact of the fact that the stepwise regression procedure, when applied with a level of α, has an actual level that is much larger than α and tends to admit more predictors than is appropriate; see, for example, Draper and Smith (1998, pp. 342–343) for further discussion on this issue. To examine this issue, a cross-validation procedure (e.g., Neter et al., 1996, p. 437) was employed in which a series of stepwise regressions were employed in which one item (Item i) was removed. The regression without Item i was then used to obtain a prediction $\tilde{Y}_i$ of the value Y_i of the dependent variable for Item i, where Y_i is either

$\hat{a}_i$ or $\hat{b}_i$. The estimated mean squared error was given by

$$\tilde{\sigma}_e^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \tilde{Y}_i)^2.$$

This mean-squared error was then compared to the estimated mean squared error obtained from the same cross-validation procedure by prediction of Y_i by the arithmetic mean $\bar{Y}_i$ of the observations Y_j, $j \neq i$. This mean squared error is

$$\tilde{\sigma}_t^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y}_i)^2 = \frac{m}{m-1}s^2,$$

where s is the sample standard deviation of Y_i, $1 \leq i \leq m$ (for an application of cross-validation in educational testing see, e.g., Haberman & Sinharay, 2010). The proportional reduction of mean squared error from use of the stepwise regression rather than a constant predictor is then

$$\text{PRMSE} = 1 - \tilde{\sigma}_e^2/\tilde{\sigma}_t^2.$$

For Form 1, the variance estimates were $s^2 = 1.09$, $\tilde{\sigma}_t^2 = 1.12$, $\tilde{\sigma}_e^2 = 5.21$ for the difficulty parameter, and $s^2 = 0.13$, $\tilde{\sigma}_t^2 = 0.14$, $\tilde{\sigma}_e^2 = 0.43$ for the discrimination parameter. As a result, the observed values of PRMSE were -3.70 for item difficulty and -2.16 for item discrimination, so that the results of the stepwise regression could reasonably be regarded as worse than useless—even the mean of the other observations predicts an observation better than the predictors chosen by a stepwise regression on an average. An alternative approach to stepwise regression can be adopted with a much stricter criterion for entry and removal of variables based on the Bonferroni correction (e.g., Miller, 1991). To ensure that the probability is no greater than 0.15 that a variable will be entered at all if the dependent variable is independent of the independent variables and the dependent variable has a normal distribution, one requires a significance level of $0.15/63 = 0.00238$ (Draper & Smith, 1998, p. 142). When a level of 0.00238 was used, no indicators of item attributes were entered at all for either item discrimination or item difficulty. Note that were tree regression, which has been applied to predict item statistics from item attributes in the abovementioned references, applied in this example and were the Bonferroni correction used, it is also true that no variables would be entered, so that no tree construction would occur. Criteria for tree branching comparable to those for stepwise regression would encounter the same problems of cross-validation found with stepwise regression.

Results from Fitting of Item Response Theory Models

We fitted the 0PL model, the Rasch model, the 2PL model, the LLTM, the constrained 2PL model and the alternative constrained 2PL model to Forms 1 to 4. Results for the alternative constrained 2PL model are essentially the same as that for

Table 3
Penalty for the Different Models for Form 1

Model	Number of Parameters	Penalty	Cor($\hat{b}$, $p+$)	Cor($\hat{b}$, $\hat{b}$)	Cor($\hat{a}$, R_{bis})	Cor($\hat{a}$, $\hat{a}_2$)
0PL	2	0.633				
LLTM-IT&IC	7	0.627	−0.26	0.26		
C2PL-IT&IC	12	0.626	−0.25	0.26	0.17	0.04
LLTM-Stepwise	15	0.585	−0.73	0.70		
C2PL-Stepwise	28	0.575	−0.50	0.49	0.35	0.33
Rasch	41	0.537	−0.98			
2PL	80	0.531	−0.99		0.92	

Note. "IT&IC" refers to a model (LLTM or C2PL) based on the item type and item content classifications.

the constrained 2PL model, so that they are not reported. In the case of the LLTM and the constrained 2PL model, a model based on the six linearly independent item content and item type indicators was employed for all four forms. In addition, for Form 1, the LLTM and the constrained 2PL model were applied with 14 indicator variables. One indicator was 1 for all items, and the other indicator variables were those used in the final model from either the stepwise regression for item difficulty or the stepwise regression for item discrimination.

Table 3 shows the values of penalty for Form 1. Each row corresponds to a model. The table shows, for each model, the following quantities:

- The number of parameters.
- Penalty.
- The correlation between the proportion correct $p+$ and the estimated difficulty from the model. This is denoted as Cor($\hat{b}$, $p+$) in the table.
- (for only the LLTM and constrained 2PL model) The correlation between the estimated difficulty from the model and the estimated difficulty from the corresponding unrestricted model (which is the Rasch model for the LLTM and the 2PL model for the constrained 2PL model). The correlation is denoted as Cor($\hat{b}$, $\hat{b}$).
- The correlation between the item R-biserial coefficient R_{bis} and the estimated discrimination from the model. This is denoted as Cor($\hat{a}$, R_{bis}).
- (for only the constrained 2PL model) The correlation between the estimated discrimination from the model and the estimated discrimination from the 2PL model. This is denoted as Cor($\hat{a}$, $\hat{a}_2$).

Table 3 has two rows each for the LLTM and the constrained 2PL model—one for the fit using the item type and item content classifications and another for the fit using a stepwise regression on all the available predictors. The models are ordered in the decreasing order of penalty in the table.

Interpretation of Table 3 is straightforward, except for the models based on item attributes from stepwise regression. The 2PL model is a bit more successful than is the Rasch model, but the difference is small. The $RI_{\text{0PL-Rasch-2PL}}$ statistic is 0.95, so that the preponderance of the improvement in penalty from the 0PL to the 2PL

model is obtained from the transition from the 0PL to the Rasch model. This result and the observed differences in penalty are relatively common in educational tests (e.g., Haberman, 2006). Note that

$$\begin{aligned}&\text{Number of parameters in the 2PL model} - \text{Number of parameters in the Rasch model}\\ &= \text{Number of parameters in the Rasch model} - \text{Number of parameters in the 0PL model}\\ &= 39,\end{aligned}$$

so that the improvement in penalty per independent parameter is

$$\text{Improvement}_{\text{0PL}-\text{Rasch}} = 0.0025$$

for the comparison of the 0PL and Rasch models and

$$\text{Improvement}_{\text{Rasch}-\text{2PL}} = 0.0001$$

for the comparison of the Rasch and 2PL models.

The LLTM based on the item content and item type attributes is relatively unsuccessful. The $\text{RI}_{\text{0PL-LLTM-Rasch}}$ statistic is only 0.06, so that relatively little of the improvement from the 0PL to the Rasch model is explained by the LLTM. In addition,

$$\text{Improvement}_{\text{0PL}-\text{LLTM}} = 0.0012$$

is a somewhat smaller improvement of penalty per independent parameter for the comparison of the 0PL model and LLTM than the corresponding value

$$\text{Improvement}_{\text{LLTM}-\text{Rasch}} = 0.0027$$

from comparison of the LLTM to the Rasch model. Similar comments apply to the constrained 2PL model based on the item content and item type classifications.

The LLTM based on the item attributes from stepwise regression is not successful, but it appears more successful than the LLTM based on the item content and item type attributes. The $\text{RI}_{\text{0PL-LLTM-Rasch}}$ statistic is 0.50,

$$\text{Improvement}_{\text{0PL}-\text{LLTM}} = 0.0037,$$

and

$$\text{Improvement}_{\text{LLTM}-\text{Rasch}} = 0.0019,$$

so that the LLTM is substantially less effective than the full Rasch model, but it does reduce penalty per independent parameter compared to the 0PL model somewhat better than in the case of the LLTM based on item content and item type. Results for the constrained 2PL case are somewhat similar.

Nonetheless, as mentioned earlier, stepwise regression has a tendency to include more predictors than is appropriate. To check this issue, 20 additional LLTMs were considered in which one item attribute indicator was 1 for each item and 13 item attribute indicators were selected at random from the 63 available indicators for item attributes. The additional restriction was imposed that the number of independent parameters be 14. There are about 10^{13} LLTMs that meet these requirements—that is because there can be about 10^{13} different combinations of 13 attribute indicators out of a total of 63. For each combination of 13 nontrivial indicators, the penalty was computed along with the R^2 statistics for prediction of item difficulty from the indicator variables. The sample mean of the penalty statistics was 0.610, and the sample standard deviation was 0.0063. The smallest penalty observed from the 20 additional models was 0.597, and the corresponding value of $RI_{0\text{PL}-\text{LLTM}-\text{Rasch}}$ was 0.38, so that the results of stepwise regression appears to be a bit better than those typically derived by a random use of a comparable number of indicator variables for item attributes. However the observed penalty for the LLTM based on the stepwise regression, which is possibly one of the best LLTMs out of a total of about 10^{13} LLTMs, is 0.585 (from Table 3). If we had fitted the 10^{13} LLTMs using all possible combinations of 13 randomly chosen attribute indicators, then the penalty of many of them would have been 0.585 or lower.[1] Thus, it is quite plausible that the value of 0.50 of the statistic $RI_{0\text{PL-LLTM-Rasch}}$ based on stepwise regression merely reflects the tendency of the stepwise regression procedure to admit more predictors than is appropriate rather than adequate prediction of item statistics from item attributes. In addition, from Table 3, Cor($\hat{b}$, $p+$) is -0.73 for the LLTM, which means that the estimated difficulty parameters from the LLTM explain about 53% of the variation in the item proportions correct ($p+$), which is way below the 96–98% range for the Rasch and 2PL models. Similar remarks also apply to the constrained 2PL model—this model does not indicate adequate prediction of the item statistics from item attributes either—note the unimpressive values of -0.50 of Cor($\hat{b}$, $p+$) and 0.35 of Cor($\hat{a}$, R_{bis}) for this model.

Table 4 provides an analysis for Form 2 that is quite comparable to the analysis for Form 1, except that the 63 item attributes were not available. The results for Forms 3 and 4 are similar to those for Form 2 and are not shown here. The penalty for the LLTM and the constrained 2PL model are much larger than that for the Rasch and

Table 4
Penalty for the Different Models for Form 2

Model	Number of Parameters	Penalty	Cor($\hat{b}$, $p+$)	Cor($\hat{b}$,$\hat{b}$)	Cor($\hat{a}$, R_{bis})	Cor($\hat{a}$,$\hat{a}_2$)
0PL	2	0.628				
LLTM-IT&IC	7	0.616	−0.34	0.33		
TC2PL-IT&IC	12	0.615	−0.34	0.36	0.03	0.00
Rasch	41	0.518	−0.98			
2PL	80	0.514	−0.98		0.86	

Note. "IT&IC" refers to a model (LLTM or C2PL) based on the item type and item content classifications.

2PL models, respectively. In addition, Cor($\hat{b}$, $p+$) is rather low for the LLTM and the constrained 2PL model and Cor($\hat{a}$, R_{bis}) is rather low for the constrained 2PL model. These results, like the regression results above, show that item type and item content classifications are poor predictors of either item difficulty or item discrimination.

It is reasonable to conclude that the available item attributes for the four forms provide no basis for scale anchoring. It is no wonder then that the consensus component of the scale anchoring process for the first form was found tedious by the participants.

Conclusions

This paper describes a set of simple statistical and psychometric techniques that can be used to examine if a scale anchoring study will come up with useful information. The techniques involve fitting of linear regression and item response theory models to examine whether appropriate item attributes can predict item difficulty and item discrimination. The application of the techniques to four forms of a teachers' licensing examination shows that the item attributes do not predict the item difficulty and item discrimination adequately for these data. Two basic criteria support this claim. The first, based on linear regression, involves the relatively low values of R^2 achieved by prediction of item parameters by item characteristics. The second, based on use of LLTMs and constrained 2PL models, involves the relatively weak performance of item response theory models in which item difficulty and/or item discrimination are predicted by item characteristics. Thus scale anchoring is not expected to provide much useful information to the examinees for this series of examinations.

The discouraging results for the example considered do not necessarily imply that the same results will always be observed, but they certainly indicate that success in scale anchoring is far from guaranteed. Presumably the adequacy of the list of item attributes possessed by the items is a key to the set of the techniques suggested. Such a list can be found in the test blueprint used by the test developers to build test forms, or such a list can be produced from scale anchoring of another form of the same test or a similar test. It is possible that our suggested techniques performed with a set of available attributes show that a scale anchoring study will fail to provide useful information, but, later, in a scale anchoring study, the content experts come up with a different list of item attributes to describe the anchor levels. However, in our opinion, this situation will mostly occur for tests in which the test construction process is not very rigorous, so that test forms are created without careful attention to item attributes. Note that if a testing program intends to perform scale anchoring and report performance-level descriptors, several researchers such as Bejar, Braun, and Tannenbaum (2007) have argued that the descriptors should be written early in the test development process and be used in developing test blueprints and item specifications. For example, performance-level descriptors may be written as a part of an evidence-centered design (Almond, Steinberg, & Mislevy, 2002; Mislevy, Steinberg, & Almond, 2003) of the test. If that is done, while there is no guarantee, the test will have a higher chance to result in meaningful performance-level descriptors. The methodology suggested in this paper can be used in the initial stages of the

construction of such a test, probably after a trial administration and before an operational administration. Attempts to report performance-level descriptors from a test which was not built to do so usually will not result in useful information.

Numerical quantities computed from features of the texts of the test items have been used to predict item difficulty. See, for example, Freedle and Kostin (1999) and Koslin, Zeno, and Koslin (1987). However, such quantities were not available to us.

Any attempt at scale anchoring is likely to be more successful if examinees are relatively diverse in terms of the measured proficiency. In such a case, it is more likely that certain types of items will exist that will consistently be difficult for portions of the examinee population or will consistently be easy for other groups of examinees.

A further issue is the importance of sample size. Statistical procedures are far more likely to lead to satisfactory results with larger collections of items. Longer tests are thus more attractive targets. In addition, it is reasonable to consider multiple forms, although such a study has to ensure that the item difficulty and item discrimination parameters of the different forms are comparable to each other.

Note

[1]That is because, assuming that the penalty of the LLTMs involving randomly chosen attribute indicators follow a normal distribution with mean 0.610 and standard deviation of 0.0063, the probability of observing a value of less than 0.585 is 0.00004, and, out of 10^{13} draws from such a normal distribution, a large number are expected to be below 0.585.

Acknowledgments

The authors thank Brian Clauser, Isaac Bejar, Andreas Oranje, Skip Livingston, Titus Teodorescu, Susan Embretson, and the three anonymous reviewers for their advice. The authors gratefully acknowledge the help of Ruth Greenwood and Kim Fryer with proofreading. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four process architecture. *Journal of Technology, Learning, and Assessment*, *1*, 1–64.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191–204.

Bejar, I. I., Braun, H., & Tannenbaum, R. (2007). A prospective, predictive and progressive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1–30). Maple Grove, MN: Jam Press.

Carroll, J. B. (1993). Test theory and the behavioral scaling of test performance. In N. Fredericksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (p. 297–322). Hillsdale, NJ: Lawrence Erlbaum.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York, NY: John Wiley.

Embretson, S. E. (1993). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Freedle, R., & Kostin, I. (1999). Does text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*, 2–32.

Garcia Gomez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, *24*, 417–444.

Gilula, Z., & Haberman, S. J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*, 129–187.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*, 394–411.

Haberman, S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL alternative* (ETS Research Rep. No. RR-06-14). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Sinharay, S. (2010). The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics, 35*, 586–602.

Kelly, D. (2002). Application of the scale anchoring method to interpret the TIMSS achievement scales. In D. F. Robitaille & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 375–390). London: Kluwer Academic Publishers.

Koslin, B., Zeno, S., & Koslin, S. (1987). *The DRP: An effectiveness measure in reading*. New York, NY: College Entrance Examination Board.

Linn, R. L., & Dunbar, S. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*, 177–194.

Miller, R. G. J. (1991). *Simultaneous statistical inference*. New York, NY: Springer-Verlag.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models*. Homewood, IL: Irwin.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, *27*, 15–29.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales* (NCES 93421). Washington, DC: National Center for Education Statistics, US Department of Education.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.

Seber, G. A. F. (1977). *Linear regression analysis*. New York, NY: John Wiley.

Sheehan, K., Kostin, I., & Persky, H. (2006, April). *Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying performances on the NAEP Grade 8 Reading assessment*. Paper presented at the meeting of the National Council of Measurement in Education, San Francisco, CA.

Sheehan, K., & Mislevy, R. J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills* (ETS Research Report No. RR-94-14). Princeton, NJ: Educational Testing Service.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*, 21–28.

Wainer, H., Sheehan, K., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping on the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*, 15–25.

Authors

SANDIP SINHARAY is a Principal Research Scientist, Educational Testing Service, MS 12T, Rosedale Road, Princeton, NJ 08541; ssinharay@ets.org. His primary research interests include item response theory, equating, diagnostic score reporting, Bayesian methods, and application of statistics to education.

SHELBY J. HABERMAN is a Distinguished Presidential Appointee, Educational Testing Service, MS 12T, Rosedale Road, Princeton, NJ 08541; shaberman@ets.org. His principal research interests are analysis of qualitative data, asymptotic approximations, and applications of statistics to educational measurement.

YI-HSUAN LEE is a Research Scientist, Educational Testing Service, MS 03T, Rosedale Road, Princeton, NJ 08541; ylee@ets.org. Her principal research interests are equating, analysis of response times, jackknife methods, test security, and educational statistics.