Listening. Learning. Leading.®

Research Report
ETS RR–11-02

Statistical Procedures to Evaluate Quality of Scale Anchoring

Shelby J. Haberman

Sandip Sinharay

Yi-Hsuan Lee

January 2011

Statistical Procedures to Evaluate Quality of Scale Anchoring

Shelby J. Haberman, Sandip Sinharay, and Yi-Hsuan Lee
ETS, Princeton, New Jersey

January 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Matthias von Davier

Technical Reviewers: Isaac Bejar and Andreas Oranje

Abstract

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 (Carroll, 1993). Scale anchoring (Beaton & Allen, 1992), a technique that 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 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. This paper describes 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 teacher licensing test.

Key words: augmented subscore, distinctness, mean-squared error, proportional reduction in mean-squared error, reliability

Acknowledgments

We are grateful to Daniel Eignor, Matthias von Davier, Isaac Bejar, and Andreas Oranje for their helpful advice, to Kevin Larkin and Titus Teodorescu for their help with the data, to Susan Embretson for sending us a computer program, and to Ruth Greenwood and Kim Fryer for editorial help.

Testing companies are under constant pressure to produce information in addition to the overall test score. Subscores (e.g., Sinharay, Haberman, & Puhan, 2007) are one potential source of such additional information. Another source is information for test takers and test score users concerning the type of tasks examinees at specified score levels are typically able to perform. Although such information might appear to be readily supplied, in practice the task has been a persistent 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), which results in descriptions of what students at different points on a score scale know and can do, is a tool to provide such information concerning the relationship between tasks examinee can perform and observed test scores. For example, a scale anchoring study for the TOEFL iBT™ Reading section (Garcia Gomez, Noah, Schedl, Wright, & Yolkut, 2007) found, among other things, that the test-takers who obtain a high score (22-30) in TOEFL iBT Reading typically have a very good command of academic vocabulary and grammatical structure. Scale anchoring has been used with a variety of assessments, including the National Assessment of Educational Progress (NAEP; Beaton & Allen, 1992) and the Trends in International Mathematics and Science Study (TIMSS; Kelley, 2002). The procedure of scale anchoring produces *performance-level descriptors* or PLDs (Perie, 2008), which describe the level of knowledge and skills required of different performance levels.

The U. S. Government's No Child Left Behind (NCLB) Act of 2001 demands, among other things, that students should receive diagnostic reports that allow teachers to address their specific academic needs; scale anchoring could be used in such a diagnostic report. Some researchers (e.g., Sinharay & Haberman, 2008) recommended consideration of scale anchoring for tests that are under pressure to report additional information, but do not have high-quality subscores.

Nonetheless, scale anchoring is not without problems. Linn and Dunbar (1992) described the confusion of the general public about the meaning of NAEP 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 potential 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.

The steps required in scale anchoring are the following:

1. Select a few carefully 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 (greater than 50%) of the students at the higher score levels can answer it correctly and most (less than 50%) of the students at the lower level cannot.

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. What students at various scale points know and can do can be summarized this way.

The above description shows that scale anchoring involves a statistical component (the first three steps) 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. The fourth step involves generalizations not required in item mapping. Scale anchoring 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 consensus component can be costly (because of the involvement of subject-area and educational experts) and can be quite time-consuming. In addition, the subjective judgment involved may not be reliable.

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 simply reflect 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 a scale anchoring will provide useful results. This paper suggests such a set of analyses—they include simple regression analysis and fitting of several popular item response theory (IRT) models. The next section discusses our suggested set of techniques and describes why they are appropriate. The techniques are applied to several data sets from a teacher licensing test in the application section. Conclusions and recommendations are provided in the last section.

1 Methods to Predict Success of Scale Anchoring

The description of scale anchoring given in the previou section indicates that scale anchoring can only succeed (which means that it can provide useful information to the examinees) if, for each pair of successive anchor points (which correspond to a small range of ability or difficulty level of items, for example, a range of proportion correct of 0.60 to 0.75), there are items with specific attributes that most students at the lower point cannot answer but most students at the higher point can, that is, the items are highly discriminating at specific levels of difficulty. Thus scale anchoring can only succeed if item attributes can predict item difficulties to an adequate degree and if item discriminations associated with these item attributes are high. If item attributes do not predict item difficulties well, then the items discriminating between adjacent anchor points will not be readily interpreted in terms of item attributes. Unless item discriminations are consistently high, it is also necessary for item attributes to predict item discrimination. In Step 4 of

the description of scale anchoring, the required generalizations will not be feasible. Hence, the key to our suggested techniques is an examination of how well item attributes predict item difficulties and item discriminations. A mathematical proof of why it is necessary and sufficient for the item attributes to predict item difficulties and item discriminations for the success of scale anchoring is given towards the end of this section.

The techniques that will be suggested here assume availability of test data concerning item attributes along with 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. A common problem will be that the test design is likely not to be optimal for the purpose of inferences concerning item attributes. This issue will receive further attention in the concluding section.

The first technique that can be used is simple linear regression of item statistics (item difficulty or item discrimination) on indicators of appropriate item attributes (see Sheehan & Mislevy, 1994, for examples of such analyses). The squared multiple correlations from these regressions will provide an idea of how well the item statistics can be predicted by the item attributes.

The second set of techniques involves fitting of several item-response theory (IRT) models to the data. In the operational data examples considered later, all items are right-scored and n examinees respond to m items. 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. The response of Examinee s, $1 \leq s \leq n$, to Item i is X_{is}, and the latent proficiency parameter of Examinee s is a random variable θ_s with a standard normal distribution. Conditional on θ_s, the X_{is} are mutually independent and the probability that $X_{is} = 1$ is p_{is}. The logit of p_{is} is $\lambda_{is} = \log\left[\frac{p_{is}}{1-p_{is}}\right]$, so that

$$p_{is} = \frac{\exp(\lambda_{is})}{1 + \exp(\lambda_{is})}.$$

All models considered are special cases of the two-parameter logistic (2PL) model M_2 in

which

$$\lambda_{is} = a_i\theta_s - \beta_i \tag{1}$$

for the discrimination a_i and the intercept β_i of Item i. If $a_i > 0$, then $b_i = \beta_i/a_i$ is the difficulty of Item i. In the one-parameter logistic (1PL) model M_1, also known as the Rasch model, it is assumed that the discrimination parameter a_i is the same for all items. In the zero-parameter logistic (0PL) model M_0, it is assumed that both the discrimination parameter a_i and the intercept parameter β_i are the same for all items. In the independence model M_I (Haberman, 2006), it is assumed that the item discrimination parameter a_i is 0 for all items, so that the X_{is} are mutually independent and p_{is} does not depend on θ_s.

Models M_0, M_1, M_2, and M_I 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) M_L, the Rasch model M_1 is assumed, and it is assumed that the item intercept satisfies a linear model

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

in which η_k represents the effect of Attribute k on the intercept β_i of Item i. Model M_L reduces to the 0PL model M_0 if $K = 1$ and $q_{i1} = 1$ for all Items i. Model M_L is the same as the Rasch model M_1 if $K = m$ and the m by K matrix $\mathbf{Q}$ of q_{ik} has rank m. The linear logistic test model has two generalizations to 2PL models. In the constrained 2PL model (Embretson, 1993) M_C, 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 model reduces to the 0PL model M_0. In the alternative constrained 2PL model M_A, (2) is assumed to hold for some η_k and (4) is assumed to hold for some τ_k.

To compare models, the information-theoretic measure *minimum estimated expected log penalty per item* (MEELPPI; see, e.g., Gilula & Haberman, 2001; Haberman, 2006) may be employed. For Model M_x, the MEELPPI is obtained as $\hat{H}_x = -\ell_x/(2nm)$, where ℓ_x is the maximum log-likelihood under the model. For example, $\hat{H}_1$ is the MEELPPI for Model M_1 and $\hat{H}_A$ is the MEELPPI under Model M_A. Among the models under study, $\hat{H}_0 \geq \hat{H}_L \geq \hat{H}_1 \geq \hat{H}_2$, $\hat{H}_0 \geq \hat{H}_L \geq \hat{H}_C \geq \hat{H}_2$, $\hat{H}_0 \geq \hat{H}_L \geq \hat{H}_A \geq \hat{H}_2$, and $\hat{H}_I \geq \hat{H}_1 \geq \hat{H}_2$ because, for example, M_0 is a special case of M_L and M_1 is a special case of M_2. An LLTM model is most attractive if $\hat{H}_L$ is close to $\hat{H}_A$, $\hat{H}_C$, $\hat{H}_1$, and $\hat{H}_2$. The constrained 2PL model is most attractive if $\hat{H}_C$ is close to $\hat{H}_2$, and the alternate constrained 2PL model is most attractive if $\hat{H}_A$ is close to $\hat{H}_2$. Evaluation of closeness can be considered in terms of relative reduction of MEELPPI and in terms of reductions of MEELPPI per independent parameter. Let M_x have d_x independent parameters, so that $d_0 = 2$, $d_I = m$, $d_1 = m + 1$, $d_2 = 2m$, $d_L = 1 + K$, and $d_C = d_A = 2K$. If Model M_x implies Model M_y but the models are not equivalent, then the improvement in MEELPPI per independent parameter is

$$\nu_{xy} = \frac{\hat{H}_x - \hat{H}_y}{d_y - d_x}.$$

Larger values of ν_{xy} are favorable for Model M_y. If Model M_x implies Model M_y, and Model M_y implies Model M_z and the models are not equivalent, then one may examine the relative improvement

$$R^2_{xyz} = \frac{\hat{H}_x - \hat{H}_y}{\hat{H}_x - \hat{H}_z}$$

to know how M_y compares to Model M_z, where M_x provides a baseline for comparison of M_y and M_z. Values of R^2_{xyz} near 1 are desirable. It is certainly desired that R^2_{xyz} be somewhat larger than $(d_y - d_x)/(d_z - d_x)$, so that the gain per independent parameter from Model M_x to Model M_y is somewhat larger than is the gain per independent parameter from Model M_y to Model M_z.

For example, consider an evaluation of the linear logistic test model (Model M_L). Consider a comparison to the Rasch model (Model M_1) where the 0PL model (Model M_0) provides a baseline for comparison. Assume that $0 < K < m$. It is desirable that

$$\nu_{0L} = (K - 1)^{-1}(\hat{H}_0 - \hat{H}_L)$$

be somewhat larger than is

$$\nu_{L1} = (m - K)^{-1}(\hat{H}_L - \hat{H}_1).$$

It is also desirable that

$$R^2_{0L1} = \frac{\hat{H}_0 - \hat{H}_L}{\hat{H}_0 - \hat{H}_1}$$

be somewhat larger than $(K-1)/(m-1)$. Favorable results suggest some ability to predict item intercept by use of item attributes. In the Rasch case, the ability to predict item intercept is equivalent to the ability to predict item difficulty.

Similar arguments can be applied to the constrained 2PL model M_C or the alternate constrained 2PL model M_A. In the case of M_C, it is important to examine

$$\nu_{0C} = [2(K-1)]^{-1}(\hat{H}_0 - \hat{H}_C)$$

be somewhat larger than

$$\nu_{C2} = [2(m-K)]^{-1}(\hat{H}_C - \hat{H}_2).$$

It is also desirable that

$$R^2_{0C2} = \frac{\hat{H}_0 - \hat{H}_C}{\hat{H}_0 - \hat{H}_2}$$

be 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 M_x imply Model M_y, and let Models M_x and M_y not be equivalent. If Model M_x holds, then the likelihood-ratio chi-square statistic $L^2_{xy} = 2nm(\hat{H}_x - \hat{H}_y)$ has an approximate chi-square distribution on $d_y - d_x$ degrees of freedom. In large samples, L^2_{xy} will be quite large even if the deviation of Model x from the data is small, so that this approach is not very helpful in practice. In all cases in this report, L^2_{xy} is highly significant.

To discuss the relationship of model parameters in the 2PL model, 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 for $\theta_s = \omega$ and no more than $q < p$ for $\theta_s = \upsilon < \omega$. That means

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

and

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

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

$$\frac{\log[p/(1-p)] - \log[q/(1-q)]}{\omega - \upsilon}, \tag{5}$$

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

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

and

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

so that the difficulty parameter b_i must be between

$$\upsilon - a_i^{-1}\log[q/(1-q)]$$

and

$$\omega - a_i^{-1}\log[p/(1-p)].$$

Suppose that the 2PL model is a reasonable approximation to the data and the item discrimination parameter a_i is sufficiently large that (5) holds. For example, if $\omega = 0.5$, $\upsilon = 0$, $p = 0.6$, and $q = 0.4$, then the discrimination must be at least 1.6. In addition, unless a_i is somewhat larger than 1.6, 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 in turn 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 is informative 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.

When scores are obtained in terms of number of items correct, one may still compare distribution functions of θ_s and the number of items correct in order to obtain the appropriate analysis in terms of the θ_s parameter.

There has been a substantial research on prediction of item discrimination and item difficulty from item attributes. Simple 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™, GRE®, and NAEP reading (e.g., Sheehan & Mislevy, 1994; Sheehan, Kostin, & Persky, 2006; Wainer, Sheehan, & Wang, 1998). These studies show low to moderate amount of success in predicting item difficulty from item attributes. For example, 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 NAEP Reading. Nonetheless, it should be emphasized that the reported values were not examined by either cross-validation or by rigorous statistical analysis designed to adjust for the effects of selection bias. In this report, in addition to the IRT analysis, conventional regression analysis is performed to predict basic item statistics.

2 Application

Data From a Scale Anchoring Study

A scale anchoring study was recently performed using four forms of a teacher licensing test in mathematics. The least and largest 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 least and largest 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 percentage of examinees in the range defined by anchor level i that answered the item correctly and the percentage 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 in Table 1, 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 outcome of the scale anchoring procedure were statements such as that the examinees in Group 2 can (a) order positive integers, (b) follow simple directions (two steps or fewer), and so on. The participants of the consensus component of the study found the component to be quite tedious and they often struggled to come up with a meaningful list of skills at any anchor level.

Table 1

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.

Results

Test developers classify each item in the test into one of two classifications (referred to as IT) based on item type (pure or real) and one of five classifications (IC) 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 IT and IC 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—the content experts created this table during the scale anchoring procedure.

Results from the fitting of simple regression models. We fitted the 2PL model to data from Forms 1 to 4. Then, for each form, we used a simple 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 IT and IC 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 quite 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, 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 IT and IC classification have no effect on the dependent variable and if the dependent variable is normally distributed, then the R^2 statistic 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, chapter 3). Thus no evidence exists that the IT and IC 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 IC and IT attributes on the other hand.

For Form 1, we performed a stepwise linear regression (Draper & Smith, 1998, chapter 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 1 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 version 9.2 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 IT and IC classification. 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 slope parameters, eight nontrivial item attributes were chosen, and the resulting R^2 was 0.64. Only one nontrivial item attribute was included in both the final model for item discrimination and the final model

for item difficulty. Nevertheless, cross-validation shows that the apparently high R^2 values are 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). To examine this issue, a cross-validation procedure was employed in which a series of stepwise regressions were employed in which one 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

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

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

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

where s is the sample standard deviation of Y_i, $1 \leq i \leq m$ (Haberman & Sinharay, 2008). The proportional reduction of mean-squared error from use of the stepwise regression rather than a constant predictor is then

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

The observed values of $\tilde{R}^2$ 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 much worse than useless. An alternative approach to stepwise regression can be adopted with a much stricter criterion for entry and removal of variables based on the Bonferroni inequality. 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 independence 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 applied in this example and were the Bonferroni approach 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 the fitting of IRT models. We fitted Models M_I, M_0, M_1, M_2, M_L, M_C, and M_A to Forms 1 to 4. Results for M_A are essentially the same as for M_C, so that they are not reported. In the case of the LLTM (M_L) and the constrained 2PL model (M_C), a model based on the six linearly independent IC and IT indicators was employed for all four forms. In addition, for Form 1, models M_L and M_C were applied with 14 indicator functions. One indicator was 1 for all items, and the other indicator functions were those used in the final model from either the stepwise regression for item difficulty or the stepwise regression for item discrimination.

Table 2 shows the values of MEELPPI for Form 1. Each row corresponds to a model. The table shows, for each model, the following quantities:

- The number of parameters
- MEELPPI
- The correlation between the proportion correct $p+$ and the estimated difficulty from the model (denoted as $\mathrm{Cor}(\hat{b}, p+)$ 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 $\mathrm{Cor}(\hat{b}, \hat{b})$.
- The correlation between the item R-biserial coefficient R_{bis} and the estimated discrimination from the model ($\mathrm{Cor}(\hat{a}, R_{bis})$)
- (For only the constrained 2PL model.) The correlation between the estimated discrimination from the model and the estimated discrimination from the 2PL model ($\mathrm{Cor}(\hat{a}, \hat{a}_2)$)

Table 2

Minimum Estimated Expected Log Penalty per Item (MEELPPI) for the Different Models for Form 1

Model	Number of parameters	MEELPPI	Cor($\hat{b}, p+$)	Cor($\hat{b}, \hat{b}$)	Cor($\hat{a}, R_{bis}$)	Cor($\hat{a}, \hat{a}_2$)
0PL	2	0.6328				
Independence	40	0.5927				
Rasch	41	0.5366	-0.98			
LLTM-IT&IC	7	0.6267	-0.26	0.26		
LLTM-Stepwise	15	0.5851	-0.73	0.70		
2PL	80	0.5313	-0.99		0.92	
2PL-C-IT&IC	12	0.6256	-0.25	0.26	0.17	0.04
2PL-C-Stepwise	28	0.5746	-0.50	0.49	0.35	0.33

It should be noted that regression results on item intercepts and item difficulties are comparable, so that the results for item intercepts and for the alternative constrained 2PL model are not reported.

Interpretation of Table 2 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 R^2_{012} statistic is 0.95, so that the preponderance of the improvement in MEELPPI 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 MEELPPI are relatively common in educational tests (Haberman, 2005, 2007). Note that $d_2 - d_1 = d_1 - d_0 = 39$, so that the improvement in MEELPPI per independent parameter is $\nu_{01} = 0.0025$ for the comparison of the 0PL and Rasch models and $\nu_{12} = 0.0001$ for the comparison of the Rasch and 2PL models. The LLTM based on the IC and IT attributes is relatively unsuccessful. The R^2_{0L1} 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, $\nu_{0L} = 0.0012$ is a somewhat smaller improvement of MEELPPI per independent parameter for the comparison of the 0PL model and LLTM than the corresponding value $\nu_{L1} = 0.0027$ from comparison of the LLTM to the Rasch model. Similar comments apply to the constrained 2PL model based on the IC and IT classifications. A notable feature is

that the LLTM and constrained 2PL model are both less successful than the independence model.

The LLTM based on the item attributes from stepwise regression is not very successful, but it appears more successful than the LLTM based on the IC and IT attributes. The R^2_{0L1} statistic is 0.50, $\nu_{0L} = 0.0037$, and $\nu_{L1} = 0.0019$, so that the LLTM is substantially less effective than the full Rasch model, but it does reduce MEELPPI per independent parameter compared to the 0PL model somewhat better than in the case of the LLTM based on IC and IT. Results for the constrained 2PL case are somewhat similar. Nonetheless, a substantial selection bias is involved due to the choice of item attributes by stepwise regression. 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. For each combination of 13 nontrivial indicators, the MEELPPI was computed along with the R^2 statistics for prediction of item difficulty from the indicator variables. The sample mean of the MEELPPI statistics was 0.6101, and the sample standard deviation was 0.0063. The smallest MEELPPI observed from the 20 additional models was 0.5966, and the corresponding value of R^2_{0L1} was 0.38, so that the results of stepwise regression were a bit better than those typically derived by a random use of a comparable number of indicator variables for item attributes. On the other hand, some reason still exists for concern about the reality of even the modest result achieved from the stepwise regression. A regression of MEELPPI on R^2 for the 20 models yields an estimated regression line of $0.6305 - 0.0820R^2$ for estimation of MEELPPI. The corresponding coefficient of determination is 0.88. The R^2 for the 13 nontrivial item attributes from stepwise regression is 0.52, so that the regression predicts an MEELPPI of 0.5880, a close approximation to the observed 0.5851. Thus it is quite plausible that the results based on stepwise regression merely reflect the tendency of the stepwise regression procedure to admit more predictors than is appropriate. Similar remarks also apply to the constrained 2PL case.

Table 3 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 MEELPPI for the LLTM and the constrained 2PL model are much larger than that for the Rasch and 2PL models, respectively, and are even larger than that for the independence model. These results, like the regression results above, show that IT and IC classifications are poor predictors of either item difficulty or item discrimination.

Table 3

Minimum Estimated Expected Log Penalty per Item (MEELPPI) for the Different Models for Form 2

Model	Number of parameters	MEELPPI	Cor($\hat{b}, p+$)	Cor($\hat{b}$,$\hat{b}$)	Cor($\hat{a}, R_{bis}$)	Cor($\hat{a}$,$\hat{a}_2$)
0PL	2	0.6282				
Independence	40	0.5741				
Rasch	41	0.5175	-0.98			
LLTM-IT&IC	7	0.6161	-0.34	0.33		
2PL	80	0.5137	-0.98		0.86	
2PL-C-IT&IC	12	0.6159	-0.34	0.36	0.03	0.00

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 was found tedious by the participants.

Conclusions

This paper describes a set of simple statistical and psychometrics techniques that can be used to examine if a scale anchoring study will come up with useful information. The techniques involve fitting of simple linear regression and IRT models to examine whether appropriate item attributes can predict the item difficulty and item discrimination. The application of the techniques to four forms of a teacher licensing examination show that the item attributes do not predict the item difficulty and item discrimination adequately for these data. So scale anchoring is not expected to provide much useful information to the

examinees for this test.

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 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 elicit 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 report PLDs, 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. If that is done, the methodology suggested in this paper can be used in the initial stages of a test construction, probably after a trial administration and before an operational administration. Attempts to report PLDs from a test which was not built to do so usually will not result in much useful information.

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.

References

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* (pp. 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.

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.

Haberman, S. J. (2005). *Latent-class item response models* (ETS Research Rep. No. RR-05-28). Princeton, NJ: ETS.

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: ETS.

Haberman, S. J. (2007). *The information a test provides on an ability parameter* (ETS Research Rep. No. RR-07-18). Princeton, NJ: ETS.

Haberman, S. J., & Sinharay, S. (2008). *Sample size requirements for automated essay scoring* (ETS Research Rep. No. RR-08-32). Princeton, NJ: ETS.

Kelley, D. L. (2002). Application of the scale anchoring method to interpret the TIMSS scales. In D. F. Robitalle & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 375–390). Amsterdam, the Netherlands: Springer-Verlag.

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.

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.

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 annual 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 Rep. No. RR-94-14). Princeton, NJ: ETS.

Sinharay, S., & Haberman, S. J. (2008). How much can we reliably know about what examinees know? *Measurement*, *6*, 46–49.

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. (1998). *Some paths toward making praxis scores more useful* (ETS Research Rep. No. RR-98-44). Princeton, NJ: ETS.