Measuring the Power of Learning.®

Research Report
ETS RR–16-03

Evaluation of Different Scoring Rules for a Noncognitive Test in Development

Hongwen Guo

Jiyun Zu

Patrick Kyllonen

Neal Schmitt

June 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Evaluation of Different Scoring Rules for a Noncognitive Test in Development

Hongwen Guo,[1] Jiyun Zu,[1] Patrick Kyllonen,[1] & Neal Schmitt[2]

1 Educational Testing Service, Princeton, NJ
2 Michigan State University, Lansing

In this report, systematic applications of statistical and psychometric methods are used to develop and evaluate scoring rules in terms of test reliability. Data collected from a situational judgment test are used to facilitate the comparison. For a well-developed item with appropriate keys (i.e., the correct answers), agreement among various item-scoring rules is expected in the item-option characteristic curves. In addition, when models based on item-response theory fit the data, test reliability is greatly improved, particularly if the nominal response model and its estimates are used in scoring.

Keywords Scoring rules; item-response curve; IRT; reliability

doi:10.1002/ets2.12089

For noncognitive assessments that measure adaptability, creativity, cross-cultural competence, collaborative problem-solving ability, or other skills, it is usually challenging for test developers to identify clear keys to all the questions on the test. In the case when no master or expert keys are available, observed data can be used for scoring items by such procedures as popularity, consensus, or item-response theory (IRT)-based scoring (DeMars, 2008; McDonald, 1983; Sympson & Haladyna, 1988). These data-driven scoring rules also help to verify whether there are flaws or errors when expert keys are available. However, it is challenging to maximize test reliability by optimal selection of scoring rules for items.

Among the IRT-based scoring rules, the nominal response model (NRM; Bock, 1972) can be used to score test takers' responses when the expert keys to test items are not available or are questionable. In the case of noncognitive assessment data, Kyllonen, Zu, and Guo (2014) explored the generalized partial credit model (GPCM; Masters, 1982; Muraki, 1992) and NRM for polytomously scored items. However, when the IRT-based scoring rules are used, researchers and practitioners have to make sure the models fit the data. As in all parametric models, misfit is often observed for such reasons as multidimensionality of the data; nonmonotonic item-response curves; and multiple-group structure, as in differential item functioning (Maydeu-Olivares, 2005; Sinharay, Haberman, & Jia, 2011). Haberman and Sinharay (2013) noted that no IRT model appears to agree with any data encountered in educational testing, but the practical impact of model error is quite variable. For newly developed noncognitive assessments, miskeys or wrong keys may be a source of model misfit, and their practical impact is not negligible. Meijer and Baneke (2004) recommended that, when constructing and revising a noncognitive instrument and before using a parametric IRT model or other statistical models, nonparametric approaches should be used for better data exploration.

In this study, our primary goal is to compare and evaluate different scoring rules using both nonparametric and parametric methods. Another goal is to promote systematic applications of statistical and psychometric procedures in the development of noncognitive assessments from item analysis to score reporting. We used data collected from a situational judgment test (SJT) to present our procedures. In the Data section of this report, we describe the test and data. Three observed score-based scoring rules are introduced in this section. In the Methods section, we discuss the statistical and psychometric approaches in item and test analysis. We first introduce the functional approach proposed by Ramsay (1991, 1997), which uses nonparametric smoothing techniques to estimate the item characteristic curves (ICC), the relation between probability of choosing the key for an item and a criterion score such as total score. When there is not a clear key to a situation or an item, studies have used rules based on the *wisdom of the crowd* (Davis-Stober, Budescu, Dana, & Broomell, 2014; DeMars, 2008; Galton, 1907; McDonald, 1983). The nonparametric approach we introduce here is useful

Corresponding author: H. Guo, E-mail: hguo@ets.org

in the same way to explore the response data of a test, identify possible miskeys, and suggest more appropriate keys for some ambiguous items. It may be particularly useful when item keys are not available. In the second part of the section, we further present two parametric item-response models: the NRM and the GPCM. These parametric IRT models are used to investigate various IRT-based scoring rules for this test in the latent ability space. Three reliability definitions are also introduced for later analysis. In the Result section, we use the SJT data to demonstrate how to use the functional approach in finding item keys and identifying flawed keys. In addition, we compare different scoring rules and test reliabilities based on observed and estimated scores. In the Discussion section, we conclude the study with discussions and recommendations.

Data

We analyzed data from the SJT investigated by Oswald, Schmitt, Kim, Ramsay, and Gillespie (2004). An SJT item presents a real-life situation and then provides the test taker with choices of possible responses to the situation. Oswald et al. (2004) described the development and validation of this SJT as a standardized test that assisted in predicting college students' success in the broad set of life and academic situations facing new college students. As evidenced in mission statements and other promotional materials, colleges clearly wanted students who would succeed in the college environment academically, interpersonally, and psychologically. New selection tools and assessments with adequate criterion-related validity, less adverse impact, and greater relevance to a broader conceptualization of performance in college were in demand. This SJT was such a measure that intended to evaluate students' noncognitive attributes and predict multiple dimensions of college student performance.

The studied SJT assessment had 36 items (or situations), and 5–7 options were presented for each situation. The test takers were asked to pick the option that was most likely to occur and the option that was least likely to occur based on their judgment in that situation; that is, a test taker produced two responses to each situation/item: one answer for the most likely agreement and another for the least likely agreement. The data derived from 5,564 college students aged between 16 and 48 years (99% between 16 and 21 years, $M = 17.4$, $SD = 1.1$) who answered at least one SJT item.

As Oswald et al. (2004) discussed, sometimes it might be inappropriate to rely on keys developed by a few experts (e.g., faculty or highly accomplished students specializing in just one area). Hence, in this study, we evaluated several scoring rules, including expert-provided keys, scores determined by popularity, and consensus scores (Guttman, 1941). Scores based on IRT models are discussed in the Results section.

More specifically, the test items are scored in the following ways.

Score by Expert Keys

The option chosen by a test taker as his or her most likely response to an item is scored 1 if it is one of the keys most likely to be selected by experts, − 1 if it is one of the keys most likely not to be selected by experts, or 0 otherwise.

The option chosen by a test taker as his or her least likely response is scored 1 if it is one of the keys most likely not to be selected by experts, − 1 if it is one of the keys most likely to be selected by experts, or 0 otherwise.

Score by the Most Popular Options in the Sample

If the test taker's most likely response is the same as the most popular option of the most likely item in the sample, 1 point is assigned to the test taker's most likely item score; if the response is the same as the most popular option of the least likely item in the sample, − 1 point is assigned; otherwise, 0 points are assigned.

If the test taker's least likely response is the same as the most popular option of the least likely item in the sample, 1 point is assigned to the item score; if the response is the same as the most popular option of the most likely item in the sample, − 1 point is assigned; otherwise, 0 points are assigned.

Score by the Consensus Method in the Sample

A test taker's consensus score of an item is the proportion of the test takers who choose the same option. The item consensus score is a value between 0 and 1.

The item score is the sum of the most likely and least likely scores, and the final score for an item can range from -2 to 2. Students can have three scores: sum of the most likely scores (SJT.M), sum of the least likely scores (SJT.L), and sum of both (SJT).

Methods

As Oswald et al. (2004) showed, the SJT scales seemed to reflect a single dimension, perhaps representing knowledge of how to succeed in the broad set of life and academic situations facing new college students. Therefore, in our study, we treat the test as unidimensional throughout.

In the first part of the section, we present the nonparametric method for item analysis, focusing on identifying possible miskeys and providing revised expert keys. In the second part, we focus on two parametric IRT methods, the NRM and the GPCM. In the third part, we introduce various reliabilities to evaluate different scoring rules.

Functional Approach of Item Analysis

Unlike the item-response models, the functional approach (Ramsay, 1991, 1997) has minimal constraints: It does not require unidimensionality of a latent ability variable, local independence of item responses given ability, or monotonicity of the ICC. Instead, in the functional approach, estimation of the ICC can be based on observed data, and it assumes that estimation at one score point can be facilitated by surrounding score points. The closer the test taker's score is to the studied score point for which the probability is to be estimated, the greater the relevance (weight) of his or her response to the estimation.

To facilitate the estimation, we now introduce necessary notation. Let J be the number of items of a noncognitive test given to N test takers. The item is scored polytomously by three or more ordered categories. For item j, the item scores are $1, 2, \ldots, M_j$. Let Y_j be the random variable representing the score on item j, taking values $Y_j = 1, 2, \ldots, M_j$. The total score $X = \sum_{j=1}^{J} Y_j$ is an estimate of the examinee's true score x – the expectation of the total score. Even though the observed score X is discrete, the true score x is assumed to be continuous.

Let $P_{jm}(x)$ be the probability of an examinee with a true score of x who gets a score m on item j; that is, $P_{jm}(x) = P(Y_j = m|x)$. A simple estimate of $P_{jm}(x)$ is the relative frequency of item score m on item j in the subgroup whose total score is x. These estimates are accurate if the number of observations in each subgroup is large and observations are of high reliability, but they are not useful for data that have small counts in certain subgroups. In this situation, borrowing information from nearby subgroups can lead to improved estimation (Wand & Jones, 1995). Kernel smoothing is a statistical technique for such a task.

More specifically, let $K(x)$ be a kernel function, a nonnegative and continuous function integrated to 1, such as the normal density function and the density function of a uniform distribution. The kernel estimation of $P_{jm}(x)$ is

$$\widehat{P}_{jm}(x) = \sum_{i=1}^{N} w_i(x)\, Y_{jm}^{(i)}, \tag{1}$$

where

$$w_i(x) = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{x - X_j}{h}\right)}$$

and where $Y_{jm}^{(i)} = 1$ if the ith examinee's response to item j receives a score of m, and $Y_{jm}^{(i)} = 0$ otherwise; X_i is the ith examinee's total test score; and h is the bandwidth (Wand & Jones, 1995). When h tends to infinity, $P_{jm}(x)$ becomes the proportion of test takers who score m on item j that is free of x; when h tends to zero, the estimator is the simple estimate mentioned earlier. For the particular item score m of item j, $P_{jm}(x)$ is called the option characteristic curve (OCC) as a function of x.

Large-sample results (Ruppert, Sheather, & Wand, 1995) show that, for a continuous variable, under regularity conditions, when $h \to 0$ and $Nh \to \infty$, the bias of the estimator in (1) has the order of h^2, and the variance of the estimator

has the order of $(Nh)^{-1}$. The optimal bandwidth is usually chosen to minimize the mean integrated square error, which provides an optimal $h \sim CN^{-1/5}$ for some constant C. Under additional regularity conditions, the bias can converge to zero faster than standard error (Hardle, 1990). Even though the results do not strictly hold for ordered categorical data such as test scores, the smoothing method is useful in exploring item-response behavior (see Simonoff, 1996).

Estimation of variance of $\widehat{P}_{jm}(x)$ can be found in Hardle (1990). Ramsay (1991) suggested an alternative and simplified version,

$$\mathrm{V}ar\left(\widehat{P}_{jm}(x)\right) \sim \sum_{i=1}^{N} w_i^2(x)\,\mathrm{V}ar\left(Y_{jm}^{(i)}\right) = \sum_{i=1}^{N} w_i^2(x)\,P_{jm}\left(x_i\right)\left[1 - P_{jm}\left(x_i\right)\right], \tag{2}$$

by assuming the independence of responses Y (i.e., for a given item j, the examinees' responses to the item are independent from each other) and ignoring the variation in the weights (Ramsay, 1991, p. 619). Thus one has the pointwise confidence envelope for OCC:

$$\widehat{P}_{jm}(x) \pm Z_{\alpha/2}\left\{\sum_{i=1}^{N} w_i(x)^2\,\widehat{P}_{jm}\left(x_i\right)\left[1 - \widehat{P}_{jm}\left(x_i\right)\right]\right\}^{1/2}, \tag{3}$$

where $Z_\alpha = \Phi^{-1}(1-\alpha/2)$, the critical value of a standard normal distribution with a confidence level of $1-\alpha$, and where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Note that the OCC represents the probability of an examinee with ability x receiving a score of m on one item. To describe the behavior of one item as a whole, the ICC is defined as

$$\mathrm{ICC}_j(x) = \sum_{m=1}^{M_j} mP_{jm}(x) \tag{4}$$

for item j. For an examinee with ability x, $\mathrm{ICC}_j(x)$ is his or her expected score on item j. An estimate of $\widehat{\mathrm{ICC}_j}(x)$ can be easily obtained as $\sum_m m\widehat{P}_{jm}(x)$. Hence, under regularity conditions,

$$\mathrm{Var}\left[\widehat{\mathrm{ICC}_j}(x)\right] \sim \sum_i w_i(x)^2\left\{\sum_m m^2 P_{jm}\left(x_i\right) - \left[\sum_m mP_{ijm}\left(x_i\right)\right]^2\right\},$$

because the correlation between $\widehat{P}_{jm}(x)$ and $\widehat{P}_{jn}(x)$ can be ignored compared to the preceding main term. In a similar fashion as in (3), the confidence envelope for the ICC is

$$\widehat{\mathrm{ICC}_j}(x) \pm Z_\alpha\left\{\mathrm{V}ar\left[\widehat{\mathrm{ICC}_j}(x)\right]\right\}^{1/2}.$$

In a noncognitive test, a monotonically nondecreasing ICC may be desirable if we expect that higher total scorers will get higher item scores on the item. In practice, the true score x is replaced by the observed total test scores, and the practical impact is limited (Guo & Sinharay, 2010).

Parametric IRT Models

Visualization of the item-response curves using the functional approach can facilitate the use of IRT models at the next step. However, a big drawback of the functional approach is that it provides no justification for using total score to order students' latent ability. The same argument has been posed to challenge polytomously scored IRT models (parametric and nonparametric, particularly nonparametric). In contrast to dichotomous IRT models, most well-known polytomous IRT models do not imply stochastic ordering of the latent trait (SOL) by the total score. Only the partial credit models (PCMs; Masters, 1982) and special cases of the models imply SOL (Hemker, Sijtsma, Molenaar, & Junker, 1996, 1997). Van der Ark and Bergsma (2010) showed that a broad class of polytomous IRT models (including the GPCM) has a weaker form of SOL and argue that weak SOL justifies ordering respondents on the latent trait using the total score. Hence, in this report, we use GPCM to explore our data. In addition, we investigate the use of NRM (Bock, 1972; Thissen, Cai, & Bock, 2010), which does not have a true score but could potentially improve test reliability.

Nominal Response Model

NRM can be used to capture the information in a multiple-choice item or to assign credit for a partially correct answer in a polytomous model. Moreover, because the distributions of incorrect answers over the options of multiple-choice items differ across trait levels, it is possible, and may be desirable, to use a model that assesses information from all item options rather than one that assumes a test taker either knows the answer or randomly selects an incorrect alternative. For example, Tatsuoka's (1983) analysis of student misconceptions showed that incorrect responses vary in types when solving mathematics problems. In this regard, an item's incorrect alternatives may augment the estimate of a test taker's trait by providing information about his or her level of understanding.

NRM provides a direct expression for obtaining the probability of an examinee with trait level θ responding in the kth category of item j as

$$p_{jk}(\theta) = \frac{\exp\left(a_{jk}\theta + c_{jk}\right)}{\sum_{h=1}^{M_j} \exp\left(a_{jh}\theta + c_{jh}\right)} = \frac{\exp\left[a_{jk}\left(\theta - b_{jk}\right)\right]}{\sum_{h=1}^{M_j} \exp\left[a_{jh}\left(\theta - b_{jh}\right)\right]}, \tag{5}$$

where a_{jk}s are analogous to the traditional discrimination indices. The keyed answer/option should ideally have the highest a. The intercept parameters c_{jk} reflect the interaction between a category's difficulty and how well it discriminates. It appears that, in general, large values of c_{jk} are associated with categories with large frequencies and that the frequencies for the corresponding categories decrease as the value of c_{jk} decreases (Bock, 1972).

Partial Credit Model

Masters (1982) proposed the PCM, and Muraki (1992) generalized this model to allow inclusion of discrimination parameters (GPCM). For the GPCM, the probability function of scoring in category k on item j given the examinees' ability θ is defined as

$$P_{jk}(\theta) = \frac{\exp\left\{a_j \sum_{h=1}^{k}\left(\theta - b_{jh}\right)\right\}}{\sum_{c=1}^{M_j} \exp\left\{a_j \sum_{h=1}^{c}\left(\theta - b_{jh}\right)\right\}}, \tag{6}$$

where M_j is the number of score categories, b_{jh} is the item difficulty parameter associated with score category h, and a_j is the item discrimination. When $a_j \equiv 1$, GPCM reduces to PCM.

Reliability

Reliability quantifies the precision of test scores and other measurements. It is concerned with how the scores resulting from a measurement procedure would be expected to vary across replications of that procedure (Haertel, 2006). There are many versions of reliability definitions, but we only introduce three.

Cronbach's Alpha

In classical test theory (CTT), Cronbach's coefficient alpha is the best-known and most widely used internal consistency reliability estimate. The formula is usually given in the following form:

$$\rho_a = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J}\sigma_{Y_j}^2}{\sigma_X^2}\right), \tag{7}$$

where σ_X^2 is the total score variance and $\sigma_{Y_j}^2$ is the variance of item scores on item j. Under certain conditions, coefficient alpha is a lower bound to the internal consistency reliability (Novick & Lewis, 1967). The standard error of measurement

(SEM) indicates the extent to which test takers' scores differ from their true scores. It is a function of the reliability of a test and the standard deviation of the scores; that is,

$$\text{SEM} = \sigma\sqrt{1-\rho_a}, \tag{8}$$

where σ is the standard deviation of the observed total score on the test.

Reliability Based on Item Response Theory Models

In the IRT world, reliability may be defined as in Haberman and Sinharay (2010) in the unidimensional IRT case. Let Y_i be the item-response vector of test taker i on the test. Let $\widetilde{\theta}_i$ be the random variable $E(\theta_i|Y_i)$ with value $E(\theta_i|y)$ if $Y_i = y$. Then $\widetilde{\theta}_i$, the expected a posteriori (EAP) mean of θ_i given Y_i, is often used to approximate the latent ability of test taker i. The mean squared error for $\widetilde{\theta}_i$ is

$$E\left(\left[\theta_i - \widetilde{\theta}_i\right]^2\right) = E\left(\text{Var}\left(\theta_i|Y_i\right)\right), \tag{9}$$

and the reliability of $\widetilde{\theta}_i$ is

$$\rho^2 = 1 - \frac{\text{Var}\left(\theta_i - \widetilde{\theta}_i\right)}{\text{Var}\left(\theta_i\right)} = 1 - \frac{E\left(\left[\theta_i - \widetilde{\theta}_i\right]^2\right)}{\text{Var}\left(\theta_i\right)}, \tag{10}$$

which is also the ratio of reduction in MSE from approximation of θ_i by $\widetilde{\theta}_i$, instead of by a constant, and MSE from approximation of θ_i by a constant.

As in Haberman and Sinharay (2010), to determine the reliability of the true score on the scale of the observed score, we define the true score as

$$T_i = \sum_{j=1}^{J} mP\left(Y_{jm}|\theta_i\right)$$

in the unidimensional case. Approximate T_i by the EAP mean

$$\widetilde{T}_i = E\left(T_i|Y_i\right).$$

The reliability of $\widetilde{T}_i$ is

$$\rho^2 = 1 - \frac{\text{Var}\left(T_i - \widetilde{T}_i\right)}{\text{Var}\left(T_i\right)}. \tag{11}$$

For detailed definitions and estimations of (10) and (11), readers can refer to Haberman and Sinharay (2010).

Results

In this section, we use the SJT data to illustrate how to use the statistical and psychometric procedures introduced in the previous section.

Item Analysis

Table 1 is a summary of our data. In the table, Exp.K, Pop.K, and Con.K refer to the scoring rules introduced in the Data section based on the expert key, popularity, and consensus, respectively; SJT, SJT.M, and SJT.L are the total test, the most likely portion, and the least likely portion of the test, respectively; and α and SEM are the Cronbach's alpha and standard error of measurement.

Table 1 shows that reliability of consensus scoring is slightly higher than that of the expert key and the most popular scoring. This observation agrees with Guttman (1941) that, when the option is scored as the proportion, the coefficient alpha reliability is maximized (DeMars, 2008). Because an item score in consensus scoring is always a value between 0 and 1, the score range, variance, and SEM of consensus scoring are the smallest ones. However, because the scores are not on the same scale, this SEM is not comparable with those produced by expert scoring and popular scoring. The biggest

Table 1 Summary of Reliability, Standard Error of Measurement, and Score Range Under Different Scoring Rules

	Exp.K			Pop.K			Con.K		
	SJT	SJT.M	SJT.L	SJT	SJT.M	SJT.L	SJT	SJT.M	SJT.L
α	0.74	0.61	0.67	0.73	0.58	0.58	0.79	0.64	0.72
SEM	5.33	3.47	3.39	5.23	3.40	3.40	1.58	0.98	1.12
Max	53.00	29.00	25.00	54.00	30.00	30.00	32.71	16.32	17.55
Min	−16.00	−6.00	−14.00	−12.00	−7.00	−7.00	11.14	4.93	4.31
M	29.46	16.23	13.20	30.89	14.85	14.85	26.01	12.41	13.61
SD	10.50	5.57	5.89	10.04	5.26	5.26	3.42	1.63	2.10

drawback of consensus scoring, as well as the most popular scoring, is that test takers' scores are sample dependent; that is, a test taker's score has the undesirable property that it depends on the sample of test takers. On the contrary, expert keys are not dependent on the sample of test takers (though they might be dependent on the sample of experts). If the expert key scoring rule could be improved to achieve a higher reliability, it would be the most appropriate scoring rule to use in practice.

To evaluate expert keys, we investigated OCC plots of the 36 items. In the item analysis, OCCs for each item were plotted against SJT.M and SJT.L, respectively. In the smoothing procedure, the bandwidth was chosen as $h = 1.1 \times N^{-.2}$ (Ramsay, 1991). In addition, test takers who answered fewer than half of the items were removed from the analysis. Furthermore, for each item, missing responses were removed in computing OCCs.

Figure 1 displays the basic item statistics and OCCs for Item 28. Figure 1 (top) shows the frequency distribution of each option of the item and mean test scores of students who chose the option: The top left figure displays the most likely item and the top right figure the least likely item. For this item, students can choose one option from A, B, C, and D as their most likely agreement and one as their least likely agreement response, respectively. As shown in Figure 1, the expert key for the most likely question is D, and the expert key for the least likely question is B. In Figure 1 (bottom), OCCs of this item are plotted based on scores obtained by different scoring rules: SJTMost and SJTLeast are the most and least scores based on the expert keys; Pop.Key.Most and Pop.Key.Least are those based on the most popular scoring rule; Con.Key.Most and Con.Key.Least are those based on the consensus scoring rule; and the last column in the figure are the OCC plots based on NRM, which we discuss later. In the most likely/least likely item plot, the *x*-axis stands for the test taker's test score (which equals the sum of item scores divided by the total number of items answered) on the most likely/least likely items; the *y*-axis stands for the proportion of students who chose each option. As shown in Figure 1, all OCC curves (the first three nonparametric estimation by different scoring rules and the fourth parametric estimation) agree with each other. The plots show that Option D is the right key for the most likely item and that Option B is the right key for the least likely item with desirable properties.

Conversely, the OCC plots of Item 21 in Figure 2 show that there may be a miskey. In Figure 2 (top), the OCC of Option C decreases as the test score increases, which is not the characteristic of a desired key. However, in Figure 2 (bottom), Option C shows the monotonically increasing trend as the test score increases. Therefore Option C should be the key for the least likely item instead of the most likely item.

We revised the expert keys for a few items based on content and agreement of OCCs of the four scoring methods. Summary statistics are shown in Table 2. After revision, the reliability of the test scores produced by expert-key scoring was greatly improved.

Item Response Theory Model Fitting and Reliability

To fit IRT models to data, we used the stand-alone software package, MIRT, developed by Haberman (2013). MIRT is a general program for item-response analysis that uses the stabilized Newton–Raphson algorithm, and the adaptive Gauss–Hermite quadrature is used to accelerate computation speed. MIRT facilitates computation of estimated asymptotic standard deviations of parameters and thus examination of parameter identification (Haberman & Sinharay, 2010). In addition, generalized residual analysis is implemented in this software package for better model identification and fit analysis (Sinharay et al., 2011).

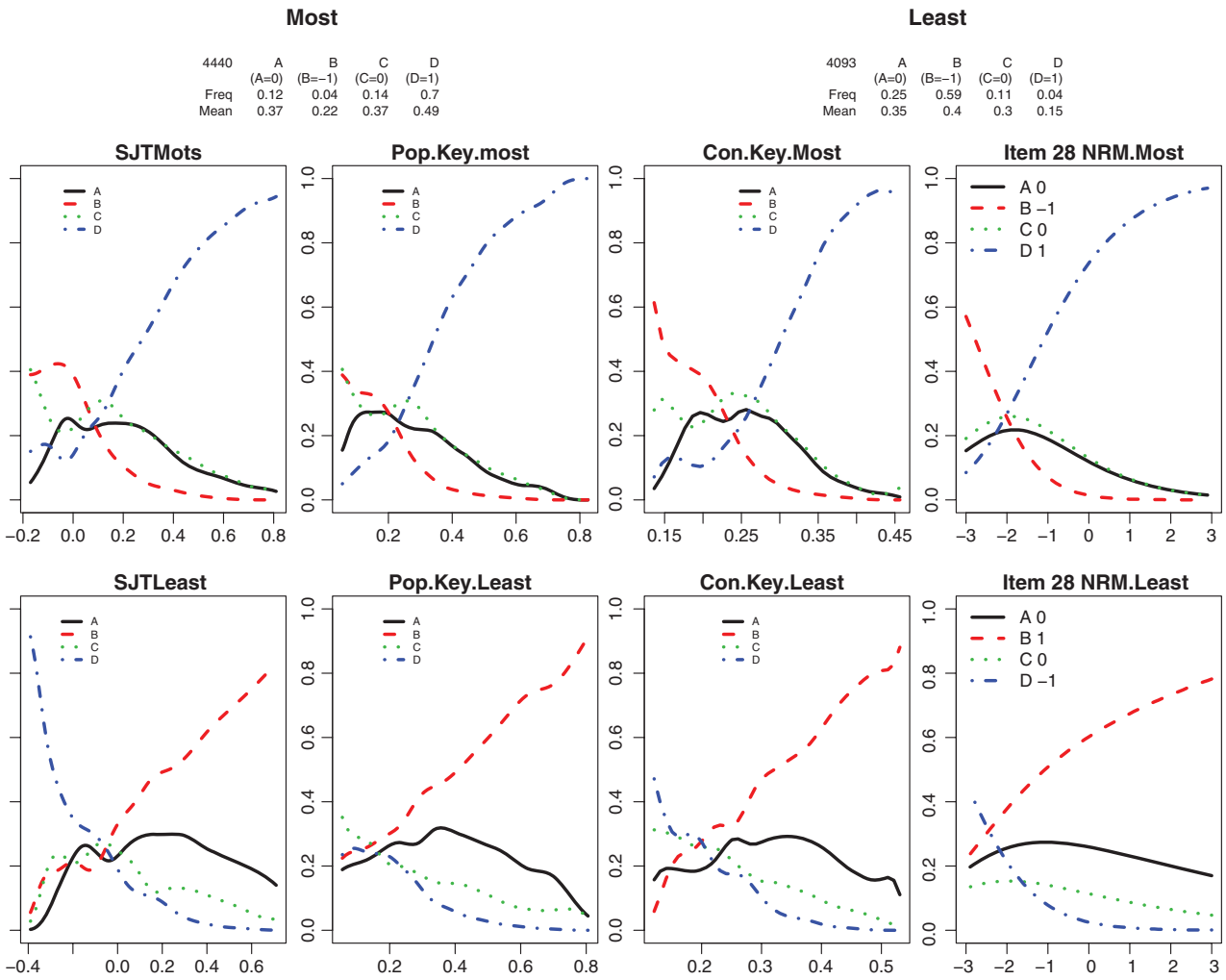

Figure 1 Option characteristic curves of Item 28.

We first fitted GPCM to the data scored by the revised expert keys (expt in short), and then we fitted NRM. Estimated item slopes and item difficulties for each item option from NRM are used in plotting the OCCs in the fourth columns of Figures 1 and 2. As shown in Figure 1, the three observed item-scoring rules and the NRM item-scoring rule agree with each other for well-defined item keys. According to the slopes estimated in NRM, for each item option, we created a new IRT-based scoring rule: assign an item score equal to the order of the option slope for this item (we call this scoring rule the slope scoring, or slp, in short). For example, Item 28 has four options, A, B, C, and D. By NRM calibration, the order (from small to large) of slopes of the four options is 2, 1, 3, 4 for the most likely agreement. If a student chooses D as his or her response to the most likely question, his or her item score is assigned 4. We also fitted GPCM to the slp scored responses. All three IRT models (GPCM-by-expt, NRM, and GPCM-by-slp) fitted the data fairly well. The marginal residuals produced in MIRT are all close to zero for the three models, and thus they are not reported. In Table 3, we present the reliability of the latent variable defined in (10). As expected, NRM produces the highest reliability for the latent ability estimates, followed by GPCM-by-slp. GPCM-by-expt has the lowest reliability.

In Table 4, we report estimated score reliability in (11) under IRT models (for definitions, refer to Haberman, 2009, 2013) for GPCM-by-expt and GPCM-by-slp. Estimated score reliability in GPCM-by-expt is in the first row of Table 4 for the most likely items and the least likely items, respectively. The estimated score reliability in GPCM-by-slp is displayed in the second row of the table. The estimated reliability for the total test is not reported because the number of the item-response categories is too large to provide a reliable estimate. Estimated score reliability from NRM is not reported either, because the values of item scores are not meaningful in NRM. In the last two rows, the Cronbach's alphas are computed

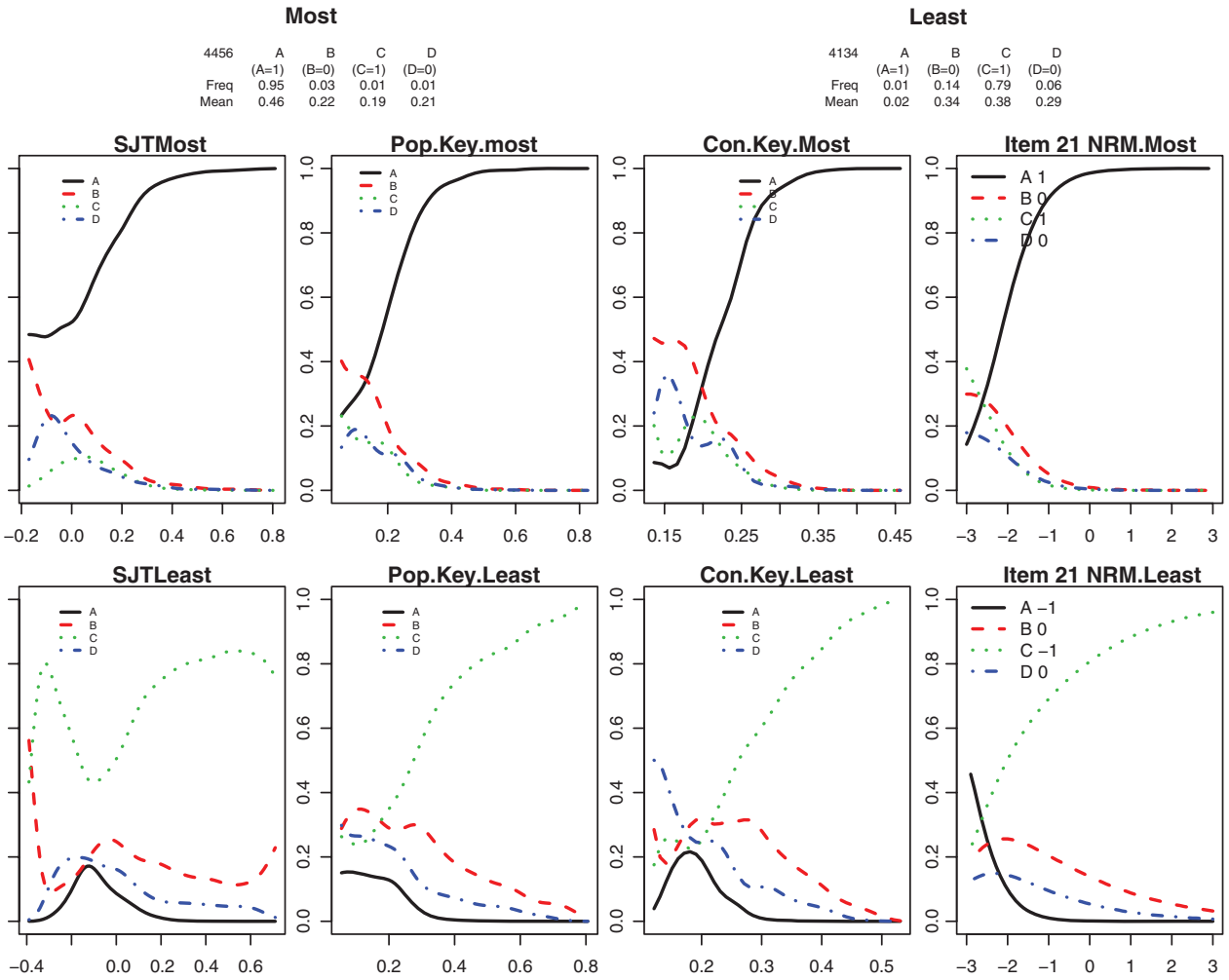

Figure 2 Option characteristic curves of Item 21.

Table 2 Summary of Reliability, Standard Error of Measurement, and Score Range Under the Revised Expert Keys

	SJT	SJT.M	SJT.L
α	0.80	0.68	0.72
SEM	5.44	3.58	3.46
Max	60.00	31.00	31.00
Min	−17.00	−9.00	−12.00
M	33.92	16.54	17.38
SD	12.01	6.36	6.56

based on the revised expert key scoring and the slope scoring rules, respectively. SEM is not compared across different scoring rules because they are not on the same scale.

From Table 4, we can observe that, when we score an item using the slope scoring rule based on the NRM output, the estimated reliability is higher than it is for those based on expert keys. As expected, an increased number of item score categories increases the reliability of the test in the IRT models (Culpepper, 2013). Comparing Tables 3 and 4, the reliabilities of true scores are higher than those of latent ability for both expert scoring and slope scoring based on IRT models.

In addition, in the observed scores of the CTT world, the test reliability (α) using the slope scoring rule is even higher than that using the expert keys, and the combined test scores have a reliability of .85. Note that, because of the large-sample

Table 3 Summary of Latent Ability Reliability

	SJT.M	SJT.L
GPCM-by-expt	0.719	0.751
NRM	0.809	0.807
GPCM-by-slp	0.795	0.789

Table 4 Summary of Score Reliability

	SJT	SJT.M	SJT.L
GPCM-by-expt		0.737	0.775
GPCM-by-slp		0.802	0.812
α (CTT by expt)	0.80	0.68	0.72
α (CTT by slp)	0.85	0.77	0.78

size in the study, a very small difference in α may be statistically significant (van Zyl, Neudecker, & Nel, 2000). Therefore statistical tests for differences in reliability may not be useful. Instead, an absolute difference is more meaningful.

The correlation coefficients between estimated latent abilities obtained from NRM and GPCM (by slp) were both .988 for the most likely data and the least likely data.

Discussion

In this study, we used a systematic procedure involving statistical and psychometric models to analyze a SJT test and its item responses. The nonparametric smoothing technique was used to evaluate item performance, identify potentially flawed keys, and create item-scoring rules. Even though the smoothing technique described here is for continuous variables, it is useful in exploring item and test score behavior.

Scoring rules based on popularity and consensus are particularly useful in the situation when it is hard to create and choose an appropriate item key. The item OCC plots based on these two scoring rules described in the study can assist test developers in exploring and evaluating the behavior of the items even when item keys are not available. They can also help test developers find the appropriate item keys, if necessary. Both scoring rules are somewhat based on the wisdom of the crowd, which was first documented by Galton (1907). After revising the expert keys, the OCC plots based on the expert keys and those based on the two remaining scoring rules show greater consistency, which confirms, again, the wisdom of the crowd, the surprising accuracy of a group's aggregated judgments, which has been demonstrated in numerous studies and anecdotes (Davis-Stober et al., 2014). In addition, the OCC plots produced by NRM agree with those produced by the scoring rules based on observed scores. Inspection of the OCC plots generated by both the nonparametric (functional) approaches and the parametric (IRT) approach and their agreement provide a visual evaluation of item quality.

When observed test scores are reported to test takers, comparison of test reliability and measurement error produced from different scoring rules can guide the test program to make reasonable decisions on the scoring rule. Generally, the consensus scoring rule produces slightly higher reliability, α, and smaller SEM because the item scores are within a small range (0 – 1). However, because this score is heavily dependent on the test sample, it is hard to compare across different sample groups. The popularity scoring rule is also somewhat sample dependent. Another issue with popularity scoring is that, for either a very hard item or a tricky item, the most popular option may not possess the desirable properties as a key that guarantees that higher scorers have a higher probability to succeed. Conversely, the appropriate expert keys are sample independent, and the test reliability is comparably high (refer to Tables 1 and 2). The biggest advantage for expert-key scoring is the easy interpretation to test takers.

When we used the NRM-generated slope scoring rule to score items, compared to expert-key scoring, the observed score reliability, α, was sizably higher (refer to the last two rows in Table 4). The increased item score categories by slope scoring increased test reliability. Therefore, if observed scores are to be reported, the NRM-generated slope scoring rule may have advantages over the expert-key scoring rule.

However, because a test taker will get a higher score when the test is easy and a lower score when the test is hard, observed scores, even true scores, are test dependent. Scores will not be comparable without test equating. Therefore,

in this regard, reporting latent abilities may overcome the test-dependency disadvantage given a well-calibrated item pool. The IRT models are convenient tools to obtain latent ability estimation. In this study, we compared GPCM and NRM using the studied data. In addition, within GPCM, the expert-key scoring and the NRM-generated slope scoring rules were compared. Similar to the observed scores in CTT, the increased number of item-response categories in slope scoring in GPCM-by-slp improved the latent ability reliability compared to expert-key scoring in GPCM-by-expt (refer to Table 3). Nevertheless, the highest reliability of latent ability was produced by the NRM model among the three IRT models (NRM, GPCM-by-slp, and GPCM-by-expt), and reliabilities of NRM and GPCM-by-slp are higher than the reliability of GPCM-by-expt. Intuitively, NRM uses all information one could possibly extract from an item and its response categories to assign item scores, GPCM-by-slp uses only the slope of all response categories of the item to assign item scores, and GPCM-by-expt uses only some item-response categories (key options) of the item. Therefore the latent ability reliability decreased in that order. Kyllonen et al. (2014) also reported that latent ability estimates obtained from NRM have the highest correlation with external variables for this dataset. Therefore, when the reported score is latent ability or its transformation, NRM may be a reasonable choice for the studied test. Even though NRM is sample dependent, estimates from a large and representative sample may be reserved for future use to avoid the problem.

One drawback for score reporting using NRM is that the latent ability scores have to be reported separately; that is, each test taker would have two ability scores: one for the most likely portion of test and the other for the least likely portion. One way to overcome this difficulty is to use the GPCM-by-slp model by combining the most-likely and least-likely scores for each item. Another advantage to using the GPCM-by-slp is the existence of true scores in this model, which NRM does not possess. Considering that latent ability estimates from NRM and GPCM-by-slp have a correlation as high as .988, the GPCM-by-slp may be a more feasible and flexible choice in practice. Alternately, one can score items by their estimated slope parameters in NRM and then combine the most-likely and least-likely scores. Further investigation of this approach is under way.

Note that the reliability obtained from IRT models is expected to be higher because of model constraints. Therefore it is crucial to check model-fit statistics. Otherwise, the improved reliability may be artificial. MIRT produces various model-fit indices. Inspection of the fit indices indicates a reasonable fit for our studied data. In addition, the ICC plots based the functional approach can be used to evaluate IRT model fit. Strong agreement between the ICC plots based the functional approach and those based on the NRM method also supports the use of IRT models in our dataset.

In our study, the SJT test has a relatively large sample. When the test sample size is small, caution should be exercised when using NRM. As DeMars (2008) pointed out, the drawbacks of using the NRM are (a) increased error variance around the estimated parameters, because NRM has more parameters to estimate; (b) chance of overfitting the model, especially for small samples; and (c) hard-to-interpret item parameters. Particularly, when we fit GPCM-by-slp to the combined item scores, for the studied test, the number of item categories can be as high as 14. Sparseness in data is often observed for some item categories, and this may easily cause any IRT program to malfunction. Conversely, in GPCM-by-expt, the number of item categories is limited to five, and IRT programs run without difficulty. Further studies can investigate methods for reducing the number of item-response categories in GPCM-by-slp but maintaining test reliability.

Overall, if model fit is reasonably good, NRM and its item parameter estimates are very useful in scoring items, particularly for some noncognitive assessments without clear keys.

Acknowledgments

The authors would like to thank Shelby Haberman for his guidance on and review of the study and for his help with appropriately fitting data with the computer program MIRT.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29 – 51.

Culpepper, S. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement*, *37*, 201 – 225.

Davis-Stober, C., Budescu, D., Dana, J., & Broomell, S. (2014). When is a crowd wise? *Decision*, *1*, 79 – 101.

DeMars, C. (2008, March). *Scoring multiple choice items*: *A comparison of IRT and classical polytomous and dichotomous methods*. Paper presented at the National Council on Measurement in Education (NCME) Conference, New York, NY. Retrieved from https://www.jmu.edu/assessment/CED/NCME/Paper/2008.pdf

Galton, F. (1907). Vox populi. *Nature*, *75*, 450 – 451.

Guo, H., & Sinharay, S. (2010). Nonparametric item response curve estimation with correction for measurement error. *Journal of Educational and Behavioral Statistics, 36*, 755 – 778.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment* (pp. 319 – 348). New York, NY: Social Science Research Council.

Haberman, S. J. (2009). *Use of generalized residuals to examine goodness of fit of item response models* (Research Report No. RR-09-15). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton – Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*, 209 – 227.

Haberman, S. J., & Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of the American Statistical Association, 108*, 1435 – 1444.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65 – 110). Westport, CT: Praeger.

Hardle, W. (1990). *Applied nonparametric regression*. Cambridge, England: Cambridge University Press.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*, 679 – 693.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331 – 347.

Kyllonen, P., Zu, J., & Guo, H. (2014, July). *Nominal response model for scoring situational judgment and other personality tests*. Paper presented at the annual meeting of the Psychometric Society, Philadelphia, PA.

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

Maydeu-Olivares, A. (2005). Further empirical results on parametric vs. non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40*, 275 – 293.

McDonald, R. P. (1983). Alternative weights and invariant parameters in optimal scaling. *Psychometrika, 48*, 377 – 391.

Meijer, R., & Baneke, J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354 – 368.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159 – 176.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1 – 13.

Oswald, F., Schmitt, N., Kim, B., Ramsay, L., & Gillespie, M. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*, 187 – 207.

Ramsay, J. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611 – 630.

Ramsay, J. (1997). A functional approach to modeling test data. In Linden, W. J., & Hambleton, R. (Eds.), *Handbook of modern item response theory* (pp. 381 – 394). New York, NY: Springer.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association, 90*, 1257 – 1270.

Simonoff, J. (1996). *Smoothing methods in statistics*. New York, NY: Springer.

Sinharay, S., Haberman, S. J., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests* (Research Report No. RR-11-29). Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Haladyna, T. M. (1988, April). *An evaluation of "polyweighting" in domain referenced testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345 – 354.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 43 – 75). New York, NY: Routledge.

van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*, 272 – 279.

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*, 271 – 280.

Wand, M., & Jones, M. (1995). *Kernel smoothing*. London, England: Chapman & Hall.

Suggested Citation:

Guo, H., Zu, J., Kyllonen, P., & Schmitt, N. (2016). *Evaluation of different scoring rules for a noncognitive test in development* (Research Report No. RR-16-03). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12089

Action Editor: John Sabatini

Reviewers: Shelby Haberman and Sam Rikoon