Psychological Test and Assessment Modeling, Volume 61, 2019 (2), 207-225

NRM-based scoring methods for situational judgment tests

Hongwen Guo[1], Jiyun Zu[2] & Patrick C. Kyllonen[3]

Abstract

Situational judgment tests (SJTs) show useful levels of validity as predictors for job performance. However, scoring SJTs is challenging. We proposed to use the nominal response model (NRM)-based scoring methods for SJTs. Using real data from an SJT, we illustrated how to setup the NRM-based scoring rules and their rationales, how to examine dimensionality and reliability, and how to evaluate item-, measurement- and score- invariance across subgroups at different time points. We also compared the NRM-based scores with other commonly-used scoring approaches in terms of their relationships with relevant external variables for the studied SJT test.

Keywords: NRM, SJT, scoring rules, reliability, validity

[1] *Correspondence concerning this article should be addressed to:* Hongwen Guo, Educational Testing Service, 660 Rosedale Rd MS 12-T, Princeton NJ 08541; email: hguo@ets.org
[2] Educational Testing Service
[3] Educational Testing Service

Introduction

A situational judgment test (SJT) evaluates a test taker's reaction to critical situations in real-world contexts, where even reasonable people with good intentions may disagree on the best actions. It measures tacit knowledge and practical intelligence of test takers that are not taught in schools (McDaniel, et al., 2007; Weekly & Ployhart, 2006). Compared to other traditional noncognitive tests, SJTs may be less subject to test takers' faking motived by social desirability, and thus they have relatively high validity (Lievens, Peeters, & Schollaert, 2007). Whetzel and McDaniel (2009) indicated in their overview of current SJT research that SJTs had useful levels of validity as predictors of job performance, they may have lower sub-group differences than general cognitive ability, and they had face and content validity because of work-related situations.

SJT items are frequently written in a way that there is no definitive "correct" answer (Whetzel, et al., 2009); hence scoring them is challenging. Various scoring rules have been proposed and compared by many researchers (Bergman, et al. 2006; Guo, Zu, Kyllonen, & Schmitt, 2016; and McDaniel, Psotka, Legree, Yost, & Weekly, 2011). The most popular SJT scoring is the subject matter expert (SME) scoring which is based on item keys or opinions provided by SMEs (Campion, Ployhart, & MacKenzie, 2014; MacCann, et al., 2008). SMEs, for an emotion management SJT for example, might include people with academic knowledge of emotions, experience in professions geared toward emotional healing, or in professions related to managing people's relationships and goals. However, the expert-scoring approaches are usually time-consuming and labor-intensive. In cases where expert keys are not available, observed data are used for scoring items by procedures such as popularity, consensus, or dichotomous item-response theory (IRT)-based scoring (McDonald, 1983; Whetzel, et al., 2009). Nevertheless, research shows that SJTs have relatively low internal consistency reliability for various reasons, including possible multi-dimensionality; a meta-analysis shows that the average internal consistency reliability of SJTs is about 0.57 (Campion, Ployhart, & MacKenzie, 2014).

Because there may not be absolutely correct or incorrect answers for items on SJTs, every option of an item may contain useful information for estimating a test taker's ability; for this reason, the nominal response model (NRM; Bock, 1972) is well-suited to analyze SJT data. In this study, using large data sets collected from two student cohorts on a SJT, we illustrate how to set up NRM -based scoring rules and their rationales, how to evaluate them in comparison with other commonly-used scoring rules on dimensionality, reliability, and measurement invariance between subgroups and time points. We also inspect relationship of scores obtained under different scoring rules with external variables. Results of these analyses may support SJTs' psychometric properties, test fairness, and validity, and they also provide guidance on operational uses of SJTs.

Because the NRM scoring rules rely on estimated item parameters that are data-driven, as in all the IRT calibrations, one particular concern researchers may have is related to measurement invariance: whether these NRM-based scoring rules established at an

earlier administration are appropriate to use in later administrations. That is, NRM-based scoring methods present challenges such as: (a) are NRM-based scores consistent across different test samples? In other words, do we need to recalibrate item parameters in the later administrations? (b) do NRM-based scores maintain consistent relationships with relevant external variables across administrations?

In the following sections, we first introduce six different scoring methods, and then evaluate dimensionality and reliability of the SJT scored by these scoring rules. Dimensionality of this SJT is evaluated by principal component analysis (PCA); Cronbach's alpha and IRT-based reliability are then computed to compare the internal reliabilities of scores produced by different scoring rules. Next, we use both item response theory (IRT) and confirmatory factor analysis (CFA) approaches to address the measurement invariance issue at the item-, test-, and score-levels respectively. Moreover, we inspect whether SJT scores maintain stable relationships with external criterion variables across the two student cohorts. In the discussion section, we summarize our findings and make recommendations for practical use of these scoring rules.

Methods

NRM

The NRM (Bock, 1972) is an IRT model designed for items with unordered (i.e., nominal) responses. It captures information of every response option for a multiple-choice item. Based on data, it assigns credit to each response option. Because the distributions of each option differ across trait levels, it is possible, and may be desirable, to use a model that assesses information from all item options rather than the one that assumes a test taker either knows the answer or randomly selects an incorrect alternative. For example, Thissen (1976) confirmed that there was information gained in the incorrect response. Therefore, each item option may augment the estimation of a test taker's trait by providing information about his or her level of understanding. Applications of the NRM are also seen in distractor analysis for the same reason, as Thissen, Steinberg, and Fitzpatrick (1989) emphasized that the distractors were part of the item. Application of the NRM may lead to increased understanding of the functions and behavior of distractors in general, and it is helpful in item writing and item analysis in any testing using a multiple-choice format (Penfield, 2008).

This is particularly true for non-cognitive assessments such as SJTs, where absolutely correct answers may not exist, and where multiple options may be reasonable in a real life situation. In the NRM, the probability of an examinee with a trait level of θ choosing the kth category/option of item j is defined as

$$P(y_j = k|\theta) = P_{jk}(\theta) = \frac{\exp(a_{jk}\theta + c_{jk})}{\sum_{h=1}^{M_j} \exp(a_{jh}\theta + c_{jh})} \quad (1)$$

where y_j is the item response of item j, a's and c's are the item parameters analogous to

the traditional item discrimination and intercept for category $k = 1, 2, \cdots, M_j$ of item j.

Let $\mathbf{y} = (y_1, y_2, \cdots, y_J)$ be the item response vector, the latent ability is estimated by $\tilde{\theta} = E(\theta|\mathbf{y})$, the expected a posteriori (EAP) mean of θ given $\mathbf{y}$. When the options for an item are scored ordinally, an alternative score is the true score (or expected weighted sum given the latent ability)[1], $T = \sum_{j=1}^{J} kP(y_j = k|\theta)$, which is estimated by $\tilde{T} = E(T|\mathbf{y}) = \sum_{j=1}^{J} kP(y_j = k|\tilde{\theta})$, the EAP measure of T given $\mathbf{y}$. The IRT-based reliabilities for $\tilde{\theta}$ (or $\tilde{T}$) is the ratio of the estimated conditional variance given responses and the estimated variance of θ (or T; Haberman & Sinharay, 2010).

Scoring

We used six scoring methods in the illustration. For this studied SJT, *Expert scoring*, which relies on SMEs, is the consensus judgment from 17 experts. To score one item, experts were asked to rank each option from 1 to 5; and then the average of the expert scores was obtained for each option of this item (refer to MacCann, et al., 2010 for details). In *popularity scoring* (which is to a dichotomous scoring method), an examinee's item score is 1 if he or she chose the most popular option for this item in the studied sample; otherwise his or her item score is zero (the popularity scoring is somewhat similar to a dichotomous expert key method).

The remaining four scoring methods require fitting a NRM model to the data. The first two are EAP estimates ($\tilde{\theta}$) of the latent ability and the EAP estimates ($\tilde{T}$) of the true score in the NRM model as described in the previous section. The *rank scoring* method (similar to a polytomous scoring method on cognitive assessments or the Likert scale on some non-cognitive assessments) assigns an item response a score equal to the rank of the estimated NRM-slopes for this item. The last scoring method, *NRM-slope scoring*, is based on the fact that the weighted sum score ($\sum_j^J \sum_{k=1}^{M_j} a_{jk} I_{jk}$, where a_{jk} is the slope parameter of Category k of Item j, and where $I_{jk} = 1$ if the kth category is chosen, and $I_{jk} = 0$ otherwise) is the sufficient statistics for the latent ability in NRM (Glas, 2016), and using of the weighted sum does not lose information carried in the responses for the ability estimation. Therefore, we also proposed to use the *NRM-slope scoring* method that assigns an item response a score equal to its estimated NRM slope. Note that the slope scoring method has not been studied before; because of its statistical property, we expect favorable results.

An example of these scoring methods is shown in Table 1 for an item with four options (A, B, C, and D). Expert scores were obtained before test administrations (presented in the fourth row) from the 17 SMEs. Row two of Table 1 shows the observed relative

[1] The advantage of the true score is that it is the expected score on the raw score scale. Note that for true scores to be meaningful, options need to be scored ordinally. In this paper, when a true score is needed, we recode item responses based on the order of the NRM-estimated slopes of the options. More details are described under rank scoring in the next subsection. When a true score is used, we fit the NRM model to the recoded ranked data. Otherwise, we use the raw data in the NRM calibration.

Table 1:
Scoring methods for an item with four options.

Option	A	B	C	D
Relative Frequency	.21	.10	.67	.002
Estimated Slope	-.06	.48	.91	-1.33
Expert Scoring	2.82	2.88	4.71	1.35
Popularity Scoring	0	0	1	0
Rank Scoring	2	3	4	1
Slope Scoring	-.06	.48	.91	-1.33

frequencies of student responses to the item; based on these frequencies, option C is the most popular choice, so the popularity scoring method (in the fifth row) assigns a score of 1 to the item if C is chosen; otherwise a score of zero is assigned. Row three of Table 1 shows the estimated NRM slopes of the item; the rankings of D, A, B, and C are 1, 2, 3, and 4 from low to high. Therefore, the ranking scoring method assigns a score of 2, 3, 4, and 1 to option A, B, C, and D, respectively[2]. The slope scoring method assigns the item slope values to its options, as shown in the last row of the table.

Measurement invariance

To ensure that NRM-based scores are comparable across different cohorts, we investigate measurement invariance for the studied test. Two different approaches (IRT and CFA) are used in our evaluation.

CFA-based

The multiple-group confirmatory factor analysis focuses on test level invariance. In this approach, researchers investigate the invariance of the relations between underlying latent variables and the observed responses; that is, whether the different regression parameters are equal in two or more groups (Hirschfeld, et al., 2014). Different levels of measurement invariance may be defined: the configural level evaluates whether the number of latent variables and the pattern of loadings of latent variables are similar, the weak level evaluates whether the magnitude of the loadings is similar, the strong level evaluates whether both item loadings and intercepts are similar, and the strict invariance evaluates whether item loadings, intercepts, and residual variances are similar. Weak and strong invariance are required to meaningfully compare the relationship and means between latent variables across groups, respectively.

IRT-based

The IRT-based approach focuses on item invariance. Test items provide equivalent measurement when the item response functions are the same across groups or samples. That is, when the item parameters are invariant across groups or samples, we achieve

[2]Crowd-sourcing and consensus judgments could be used to rank item options as well, which would be somewhat close to the ranking method we studied here

measurement invariance for the assessment; item parameter invariance is a sufficient condition for measurement invariance, and it can be evaluated by using statistical tests such as the likelihood ratio (LR) test (Thissen, Steinberg, & Wainer, 1993).

The LR statistic for item j is defined as

$$G_j^2 = -2 \ln \frac{L_{j0}}{L_j}, \tag{2}$$

where L_{j0} and L_j are the likelihood functions of fitting an IRT model assuming that the studied item j is invariant (i.e., all item parameters are constrained to be the same across groups) or different (i.e., the studied items are different across groups while the other item parameters are constrained to be the same, in a multi-group concurrent calibration). Under the invariant assumption, G^2 asymptotically follows a chi-square distribution where the degrees of freedom (DF) equals the difference of the numbers of parameters in the two models.

To evaluate the practical impact of item variance, we adopt the effect size proposed by Kim, Cohen, Alagoz, and Kim (2007) in the IRT-based Differential item functioning (DIF) study, which is parallel to the observed-score-based standardized mean difference (SMD; Dorans & Schmitt, 1991; Zwick et al., 1993). The effect size of item j (which is $T(1)$ in Kim, et al., 2007) is

$$\text{DIF}_j = \int [F_{rj}(\theta) - F_{fj}(\theta)] dG_f(\theta) \tag{3}$$

where $F_j(\theta) = \sum_{k=1}^{M_j} y_{jk} P(y_{jk}|\theta)$, $F_{rj}(\theta)$, and $F_{fj}(\theta)$ are the item response functions for the total, the reference, and focal groups of item j, y_{jk} is the rank of the kth category for item j, and $G_f(\theta)$ is the latent ability distribution of the focal group.

Score-based

Besides the above item and test invariance analysis, we compute the correlation coefficient between the estimated abilities (estimated true scores, rank scores, and slope scores) of the new sample by using the old-sample item parameters and those by using the new-sample item parameters. A higher correlation coefficient indicates a higher consistency between the scores.

Consistency

To evaluate consistency of the relationships between SJT scores and external test scores, we compare their correlation coefficients in the old and new cohort samples. Similar coefficients would provide additional evidence to validate the NRM-base SJT scoring rules.

Data

For illustration purposes, we used Data collected from the Situational Test of Emotional Management for Youths (STEM-Y), which was designed to measure students' ability to manage emotions, moderate negative emotions and enhance positive ones (MacCann, Wang, Matthews, & Roberts, 2010). This SJT consisted of eleven four-choice items, and each of the four options/choices to an item (i.e., a situation) was a likely action. For each item, test takers were asked to pick the option/choice that matched what they would do in the situation described in the item.

Data collections were conducted for a battery of non-cognitive assessment multiple times, where the SJT was embedded. The test battery contained six non-cognitive self-rating tests along with the SJT test (Petway, et al., 2016). These self-rating tests measure the following six traits respectively: time management (TM; 26 items), team work (TW; 25 items), resilience (Re; 39 items), intrinsic motivation (IM; 28 items), creativity (Cr; 23 items), and ethics (Et; 22 items). The numbers of items in each test may be slightly different between the old and new administrations because of test revision. Items on the tests were scored by a 4-point Likert scale (such as 1 = never or rarely, 2 = sometimes, 3 = often, and 4 =usually or always). This online assessment battery is to provide schools with the opportunity to examine and monitor the development of non-cognitive skills in their students from Grade 6 to Grade 8. Note that participating schools are members of the Independent School Data Exchange (INDEX) organization, and they did not receive individual student scores but school reports. Their participation was on volunteer basis.

In this study, we focused on this SJT and used two sets of data collected in 2011 (initial collection) and 2013 (larger scale collection). The old (initial) sample contained 2081 students' responses to SJT from 18 participating schools along with their responses to other measures collected in the fall of 2011. Only data from the subsample (N = 2048; 51% female; 99% between 11 and 14 years old) who responded to all eleven items on the STEM-Y were included in our analysis. The new sample was collected for the same test administered to 15,590 students from 68 participating schools (including the 18 schools in the old sample) in the fall of 2013 (Petway, Rikoon, Brenneman, Burrus, & Roberts, 2016). Only a subset of 11,723 students (49% female; 98.3% between ages 11 and 14) who responded to all eleven items were analyzed in this study. All the schools in both the old and the new data sets were private schools. The schools in the old sample were from 11 different states, while those in the new sample were from 29 states (refer to Table 2 for more details).

Results

Because of operational constraints, we used the stand-alone IRT program, *MIRT*, developed by Haberman (2013), to run all of the following IRT analyses[3]. *MIRT* is a general

[3] For R users, the packages *mirt* (Chalmers, 2012) and *sirt* (Robitzsch, 2019) can be used for IRT calibrations as well.

Table 2:
Sample information of the SJT test takers

	Old cohort sample	New cohort sample
Year	2011	2013
No. of Students	2081	15590
Gender (F:M)	51%:49%	49%:51%
Grades 6:7:8	37%:32%:31%	32%:34%:35%
Hispanic	5%	7%
White:Black:Asian	74%:4%:6%	62%:7%:10%
No. of Schools	18	68
No. of States	11	29
State list	California	Alabama, California
	Connecticut	Colorado, Connecticut
	District of Columbia	Delaware, District of Columbia
	Georgia	Georgia, Illinois
	Kentucky	Kentucky, Louisiana
	Massachusetts	Maine, Massachusetts
	New Jersey	Michigan, Minnesota
	New York	Mississippi, Missouri
	North Carolina	New Jersey, New York
	Ohio	North Carolina, Ohio
	Tennessee	Oklahoma , Pennsylvania
		South Carolina, Tennessee
		Texas, Utah
		Virginia, Washington
		Wisconsin

program for item response analysis that uses the stabilized Newton-Raphson algorithm and the adaptive Gauss-Hermite quadrature to accelerate computation speed. *MIRT* facilitates computation of estimated asymptotic standard deviations of parameters and thus facilitates examination of parameter identification. In addition, generalized residual analysis is implemented in this software package for better model identification and fit analysis (Haberman, 2009). The rest of the analyses were carried out in R (R Core Team, 2018).

Because four of the six scoring rules are NRM-based, we first evaluated the NRM fit to the data. The NRM fit was reasonable in terms of convergence, fit indices, estimation errors, and residuals. For example, in Figure 1, the differences between the observed and expected proportions of each options for the eleven items were very small, and the generalized residuals at each raw score point are mostly around ± 3 (Haberman, 2009) or less for the new sample.

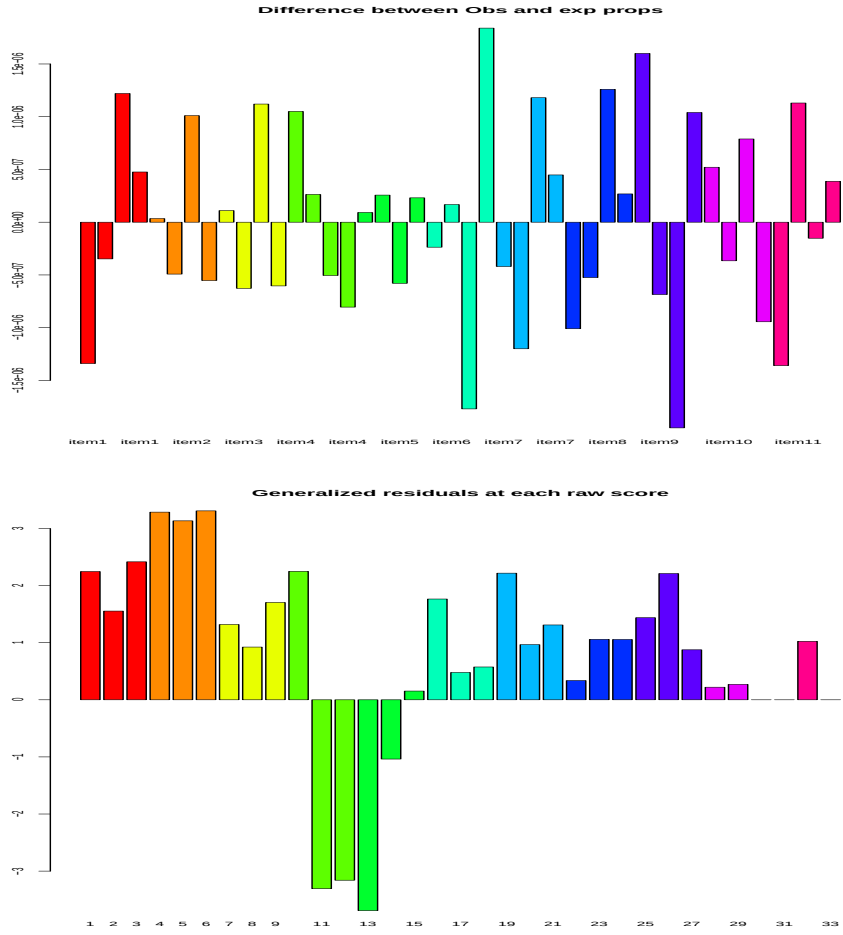

Figure 1:
The differences between the observed and expected proportions for the eleven items (upper panel, where each item had 4 options for students to choose from) are very small. Generalized residuals (lower panel) are mostly within ± 3.

Dimensionality and reliability

The dimensionality of the scores is shown in the scree plots in Figure 2. A scree plot displays the eigenvalues (on the y-axis) associated with components or factors in descending order versus components or factors (on the x-axis).

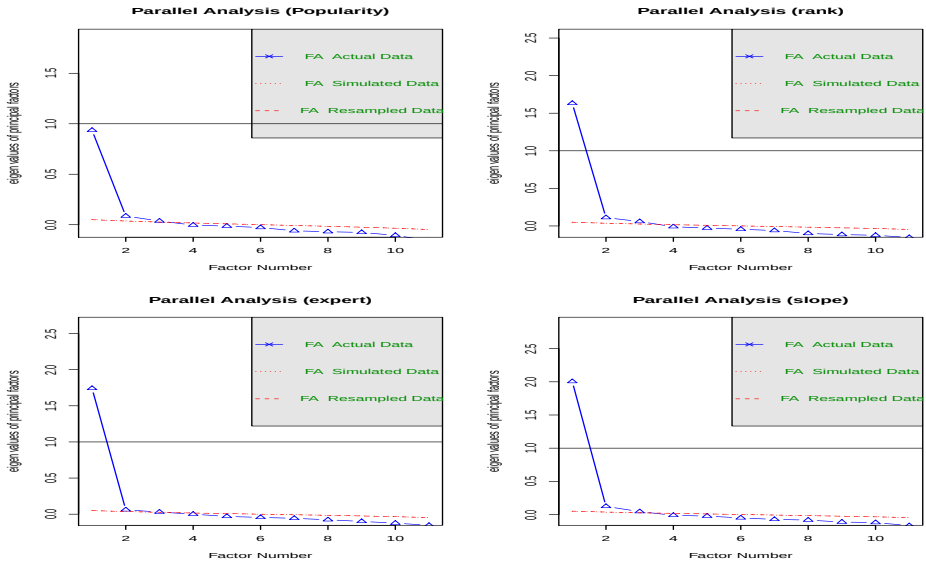

Figure 2:
The Scree plots of SJT (from top left clockwise: popularity-scored data, rank-scored data, slope-scored data, and expert-scored data).

To produce the scree plots, we ran the factor analysis (the function *fa.parallel()*) in the R-package *psych* (Revelle, 2017). From the top left, clockwise, data used to produce the scree plots were popularity scores, rank scores, slope scores and expert scores of the new sample. As expected, the scree plots of NRM-based scores, as well as the expert scores, in Figure 2, show a sudden drop in eigen values from the second factor (with values significantly less than one), indicating one dominating factor[4] (Revelle, 2017).

In the new sample, the Cronbach's alpha reliabilities based on the expert keys, popularity-, rank-, and slope-scoring methods were 0.67, 0.41, 0.65, and 0.71, respectively. Applying NRM to both the raw data and to the rank-scored data, the IRT reliability coefficients of estimated EAP latent abilities and true scores were 0.60 and 0.72, respectively. Overall, the true scores and the slope scores had the highest reliability (refer to Table 3).

Measurement invariance

To assess measurement invariance of the NRM scoring methods, we investigated its invariance on the item level (i.e, DIF) and on the test score level (i.e., comparability of the NRM-based test scores).

[4]There is potentially a second factor, but very weak.

Table 3:
Cronbach's Alphas, estimated IRT reliabilities and standard error of measurement (SEM) for the new samples (J=11 items, n=11,723 students).

Scoring	expert	popularity	rank	slope	$\tilde{\theta}$	$\tilde{T}$
Reliability	0.67	0.42	0.65	0.71	0.60	0.72
SEM	2.30	1.48	2.56	1.47	0.80	2.50

Test level (CFA-based)

We prepared three data sets for each of the old sample and the new sample, respectively, by using the expert scoring, rank scoring, and slope scoring methods. We used the R package, *lavaan* (Rosseel, 2012), to run the following analyses. Note that the Chi-square tests of group invariance applied to nested models are directly affected by sample sizes, so for large samples, even trivial differences may become significant. As a result, all the chi-square tests were statistically significant in our large samples.

Table 4 shows the model fit indices under the constraints of equal loadings (weak invariance), or equal loadings and intercepts (strong invariance), when using the same scoring rule for both the old and new samples. The model fits were satisfactory (CFI$\geq$ .95, RMSEA$\leq$ .05 and SEMR$\leq$ .05), and the scales under these three scoring methods exhibited weak to strong invariance across the two samples collected at different time points for this SJT.

Table 4:
Measurement Invariance under different scoring methods

N (old group)	2047	2047	2047
N (new group)	11723	11723	11723
Equal loadings	Rank Scoring	Slope Scoring	Expert Scoring
Number of free parameters	66	66	66
Number of equality constraints	10	10	10
Model Fit Test Statistic	638.475	785.568	476.547
Degrees of freedom	98	98	98
Comparative Fit Index (CFI)	0.952	0.956	0.969
Tucker-Lewis Index (TLI)	0.946	0.950	0.965
RMSEA	0.028	0.032	0.024
SRMR	0.022	0.024	0.019
Equal loadings & Intercepts	Rank Scoring	Slope Scoring	Expert Scoring
Number of free parameters	67	67	67
Number of equality constraints	21	21	21
Model Fit Test Statistic	746.939	884.557	562.129
Degrees of freedom	108	108	108
P-value (Chi-square)	0	0	0
Comparative Fit Index (CFI)	0.943	0.95	0.963
Tucker-Lewis Index (TLI)	0.942	0.949	0.962
RMSEA	0.029	0.032	0.025
SRMR	0.023	0.025	0.02

Item level (IRT-based)

For the item level analyses of our nominal data, we assumed that the old and new groups were samples from the same population and conducted multi-group concurrent calibrations for each studied item. We first present the log likelihood ratio G^2 test defined in Equation (2) in the previous section. Because each item had four options and each option had two parameters in NRM, the statistic G^2 followed a chi-square distribution with 6 = $(4-1) \times 2$ degrees of freedom.

Table 5:
Test statistics, p-values, and the DIF sizes of SJT items

Item	1	2	3	4	5	6	7	8	9	10	11
G^2	47.98	12.08	24.42	35.18	9.55	41.47	16.34	11.11	7.16	7.51	35.12
p	0.00	0.15	0.00	0.00	0.30	0.00	0.04	0.20	0.52	0.48	0.00
DIF	0.13	-0.02	0.06	0.08	0.03	-0.09	-0.04	-0.02	-0.01	-0.02	-0.07

The p-values in Table 5 show that about one half of the items functioned statistically

different between the two groups (p-value ≤ 0.05). The largest test statistic was 47.98. Because of the large sample sizes, it was difficult to judge whether these statistically significant differences indicated practical problems with the items. Hence, using the new sample as the focal group, we used the DIF effect size in (3) on the rank-scored responses to evaluate item performance across the two administrations. A DIF size of 0.1 has been recommended as the threshold to flag items for dichotomous (Dorans & Schmitt, 1991) and polytomous responses (Kim, et al., 2007). The DIF sizes in the studied data, shown in the last row of Table 5 on a score range of 1 to 4, indicate that only the first item may require close examination[5].

The first item is similar to the one presented below[6] (Note that on the operational test, girls' names are used for female test takers, and boys' names for male test takers): Your friend, Emma, was given a difficult project by her teacher. She asked you to help her and you did. Emma then received a bad grade on the project, and she blamed you. What would you do in this situation? (a) Keep your relationship, but stop helping Emma with her school work in the future; (b) Apologize for not doing a better job and promise to do better next time; (c) Tell Emma that even when you help her she is the one responsible for her grade; (d) Ignore Emma and spend time with other friends.

Figure 3 provides a statistical description of Item 1 in the new sample. The left panel of Figure 3 presents the summary statistics of this item. For example, it shows that 67% of the test takers picked option C, and their average EAP ability was .24, and the polyserial coefficient (which measures the association between the option and the ability; Drasgow, 1986) was .54. The plot on the right panel shows the estimated item option characteristic curves (OCCs), which are the conditional probabilities of the chosen option for Item 1 given estimated EAP-ability, using the kernel smoothing method (Guo, et al., 2016; Ramsay, 1991). Option C behaved like a key because its conditional probability monotonically increased as ability grew, while the conditional probability of other options decreased eventually. In addition, for this item, option C was the most discriminant option, and it had the highest ability mean in the new sample.

[5] A free-baseline approach may be necessary when any of the items have DIF (a purification procedure in DIF analysis).

[6] The item presented here is a disguised version of the actual (operational) item, but the rewritten form presented here maintains its essential character.

Option	a	b	c	d
Freq	0.21	0.1	0.67	0.02
Mean	−0.56	−0.1	0.24	−1.46
Polyserial	−0.48	−0.08	0.54	−0.62

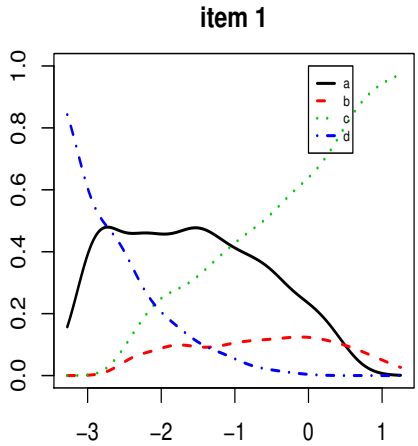

Figure 3:
Description of Item 1 for the new group.

Option	a	b	c	d
Freq	0.18	0.1	0.69	0.03
Mean	−0.64	−0.02	0.22	−1.52
Polyserial	−0.49	−0.01	0.51	−0.63

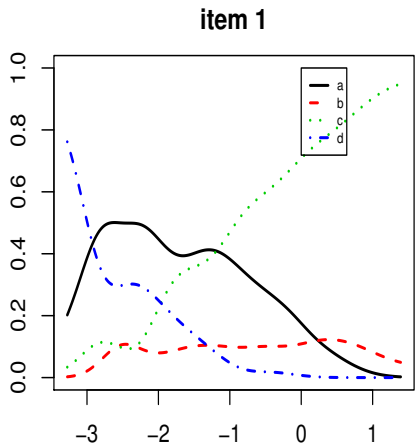

Figure 4:
Description of Item 1 for the old group.

Figure 4 shows the statistical description of Item 1 as well, but for the old group. The overall patterns were somewhat similar between the old and new groups. Rank scoring of the item options may have exaggerated the OCC differences between the two groups in the item-level DIF analysis. Information presented in the two figures could assist content experts to decide whether Item 1 needs revision.

Score level

In order to evaluate the comparability of estimated latent ability and true scores in the NRM across administrations at the two time points, we ran NRM to obtain two sets of

item parameters from the old and new samples, respectively. We then obtained two sets of estimated abilities (and two sets of true scores) of the new sample using the above two sets of item parameters. We computed the correlation of the two sets of ability estimates based on the new sample. We observed that the two sets of EAP ability estimates were highly correlated, and the correlation coefficient was 0.99. The two sets of EAP true scores for the new sample had a correlation coefficient of 1.00.

In addition, using the two sets of item parameters, we rank-scored and slope-scored the new sample, and the score correlation coefficients were 0.94 and 0.97 for the rank scores and the slope scores, respectively.

Correlation with external variables

To evaluate whether NRM-based SJT scores maintained a consistent relationship with external variables, we calculated the correlation coefficients of students' SJT scores (NRM ability estimates $\tilde{\theta}$ based on raw data, slope scores, rank scores, and expert scores) with their six external variables (sum scores) that measure the six traits (TM, TW, Re, IM,Cr, and Et), introduced in the Data section.

Table 6 displays the summary statistics of the six external tests (in the first six columns) and their correlation coefficients with SJT scores (in the last four columns).

Table 6:
Summary of the six external variables and their correlation with SJT scores,

Test	Sample	J	Size	Mean (STD)	Rel(SEM)	r($\tilde{\theta}$)	r(Slp)	r(Rnk)	r(Exp)
TW	Old	25	2023	75.58 (10.22)	.87(3.62)	.35	.36	.35	.34
	New		11736	74.67 (10.82)	.89(3.60)	.35	.32	.30	.30
TM	Old	26	2019	76.34 (11.64)	.88(4.02)	.31	.32	.30	.29
	New		11744	77.40 (13.68)	.92(3.87)	.35	.34	.33	.33
Re	Old	39	2008	115.98 (14.09)	.89(4.63)	.35	.35	.35	.35
	New	36	11733	108.29 (14.74)	.91(4.34)	.32	.31	.30	.30
IM	Old	28	1800	74.04 (16.92)	.95(3.76)	.34	.32	.32	.31
	New		11679	75.27 (15.53)	.94(3.90)	.31	.31	.30	.29
Cr	Old	23	1782	70.28 (13.75)	.95(3.14)	.24	.26	.24	.24
	New	22	11702	64.95 (12.05)	.93(3.25)	.19	.19	.18	.17
Et	Old	22	1787	69.83 (12.55)	.95(2.88)	.41	.42	.39	.40
	New		11720	68.95 (10.96)	.93(2.97)	.39	.39	.37	.36

Note 1: The six external tests are team work (TW), time management (TM), resilience (Re), intrinsic motivation (IM), creativity (Cr), and ethics (Et).
Note 2: STD stands for standard deviation of scores; Rel and SEM stand for reliability and standard error of measurement; and r(·) is the correlation coefficient between the external test scores (in the first column) and SJT scores.
Note 3. Column J shows the number of items in each test. Slp, Rnk, and Exp stand for slope scoring, ranking scoring, and expert scoring, respectively.

From Table 6, we observed that all the six tests had high internal reliability (Cronbach Alpha). Despite the sample size differences, the old and the new samples had similar scores and reliabilities on these six external tests. The SJT latent ability estimates, slope scores, rank scores, and the expert scores all had a moderate association with the six tests themselves (with correlation coefficients ranging from .17 to .42 with an average of .32 and standard deviation of .06). Among the six external skills, ethics had the strongest correlation with SJT scores, and Creativity had the lowest one. Among the three scoring rules (slope, rank, expert), SJT scores produced by slope scoring generally had slightly higher association with the external variables.

Overall, the NRM-based scores on the SJT showed reasonable consistency in relationships with the six test scores between the old and new samples.

Discussion

Even though studies show that SJTs show useful levels of validity as predictor for job performance, these tests face challenges. The validity of an SJT partly depends on its scoring, and that poor choices could lead to the conclusion that SJTs are not valid when it may only be that the scoring key is not valid. In our study, we focused on the scoring methods for SJTs in practice to improve their psychometric properties. We proposed to use the NRM-based scoring methods that may fully use the information carried in the response data, particularly the newly-defined slope-scoring. Real data sets were used to illustrate how to setup the NRM-based scoring rules and how to evaluate the test properties (dimensionality, reliability, measurement invariance at item, test, and score levels, and consistency) when using scores produced by different scoring rules.

Our analysis showed that the NRM-based scoring methods produced scores of one dominating factor, and they produced higher test reliability than other commonly used scoring methods. For example, in the new sample, expert scoring produced a internal consistency reliability (Cronbach's alpha) of 0.67, NRM-slope-scoring produced a internal consistency reliability of 0.71, and the IRT reliability of the NRM true scores in rank-scored data was 0.72. As to measurement invariance, even though the IRT-based DIF analysis showed that some items in the SJT test may function statistically different between the old and new samples (i.e., the item parameters may be somewhat different in the NRM calibrations for the two groups), the effect sizes were small, except for one item. In addition, the CFA results showed that the SJT test had weak to strong measurement invariance when each of the rank-, slope-, and expert-scoring rules was used. In addition, score consistency was maintained in these NRM scores; that is, the NRM scores using two sets of item parameters had correlation coefficients larger than 0.99. Stable relationships between SJT scores and the six external tests were maintained as well across the two sample cohorts, and the correlation coefficients were statistically significant and were mostly around .30, which supported the validity of the SJT test (Motowidlo, et al. 1990). Our analyses suggest that for the studied SJT, scores based slope scoring or the latent true score were preferable in terms of the test psychometric

properties.

Overall, the NRM-based scoring rules are very promising for scoring SJTs. They showed relatively high internal consistency reliability and stable relationships with external measures of human judgments, skills, and attitudes for the studied test, which agree with those presented in previous studies (Guo, et al., 2016; Kyllonen, et al., 2014). To use the NRM scoring rules in practice, we would recommend using a large and representative sample of the target population to set up the NRM-based scoring rules. Methodologies similar to those demonstrated here can be used to evaluate these scoring rules for decision making. Even if the testing program decides to use expert scoring, results obtained from the NRM analyses are helpful for validating and improving the expert keys.

Our study has a few limitations. First, because the SJT test under study is relatively short (eleven items), we only achieved a reliability of 0.7 or so. To improve the test reliability for operational use, we would need to add more items to the test and develop items that have higher polyserial correlation coefficients. Second, measurement invariance in demographic subgroups was not conducted for our data because of inconsistent demographic definitions between the two samples. Also, because of the particular SJT and the studied samples, generalizations about psychometric properties are not necessarily true to other SJTs. However, the methodology presented here applies to subgroup analysis and other situations to investigate which scoring methods work better. Third, for multidimensional SJTs, test-retest reliabilities should be investigated with appropriate data, instead of the internal consistency reliability, under different scoring rules.

In addition, in this study, we focused on evaluation of the psychometric properties of the NRM-based scoring rules (such as reliability, measurement invariance, and consistency). Further evidence needs to be collected for validating SJTs (Lievens, Peeters, & Schollaert, 2007; Whetzel and McDaniel, 2009).

References

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.

Chalmers, R. P. (2012). *mirt*: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48(6)*, 1-29. doi:10.18637/jss.v048.i06

Campion, M., Ployhart, R. & MacKenzie, W. (2014). The state of research on situational judgment tests: a content analysis and directions for future research. *Human Performance, 27*, 283-310.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item function: A pragmatic approach* (Research Report No. RR-91-47). Princeton, NJ: Edu-

cational Testing Service.

Drasgow, F. (1986). Polychoric and polyserial correlations. In Johnson, N. and Kotz,S. (eds.), *Encyclopedia of Statistical Sciences*. New York, Wiley.

Glas, Cees A.W. (2016). *Maximum-Likelihood Estimation*. In: Wim J. van der Linden (Ed.), Handbook of Item Response Theory, Volume Two: Statistical Tools. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. CRC Press, 197 - 216.

Guo, H., Zu, J., Kyllonen, P., & Schmitt, P. (2016). *Evaluation of Different Scoring Rules for a Non-cognitive Test in Development* (ETS Research Report No. RR-16-03). Princeton, NJ: Educational Testing Service.

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

Haberman, S. (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.

Kyllonen, P., Zu, J.,& Guo, H. (2014, July). *Nominal Response Model for Scoring Situational Judgment and Other Personality Tests*. Paper presented at the IMPS Annual Meeting, Philadelphia, PA.

Lievens, F., Peeters, H., & Schollaert, E. (2007). Situational judgment tests: a review of recent research. *Personnel Review, 37*, 426-441.

MacCann, C., Wang, P., Matthews, G., & Roberts, R. D. (2010). Emotional intelligence and the eye of the beholder: Comparing self- and parent-rated situational judgments in adolescents. *Journal for Research in Personality, 44*, 673-676.

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

McDaniel, M., Hartman, N., Whetzel, D., & Grubb, W. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91. doi: 10.1111/j.1744-6570.2007.00065.x

McDaniel, M.A., Psotka, J., Legree, P.J., Yost, A.P., & Week, J.A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology 96*, 327-336.

Motowidlo, S. J., Dunnette, M. D., & Carer, G. W. (1990). An alternative selection procedure: the low-fidelity simulation. *Journal of Applied Psychology, 75*, 640-647.

Petway, K. T., Rikoon, S. H., Brenneman, M. W., Burrus, J., & Roberts, R. D. (2016). *Development of the Mission Skills Assessment and Evidence of Its Reliability and Internal Structure* (Research Report No. RR-16-19). Princeton, NJ: Educational Testing Service.

R Core Team, (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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

Revelle, W. (2017). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.7.8.

Robitzsch, A. (2019). *sirt*: Supplementary item response theory models. R package version 3.2-39.

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48(2)*, 1-36. URL http://www.jstatsoft.org/v48/i02/.

Thissen, D., Steinberg, L. & Fitzpatrick, A. (1989). Multiple-choice models: the distractors are also part of the item. *Journal of Educational Measurement, 26*, 161-176.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale NJ: Erlbaum.

Weekly, J. & Ployhart, R. (Eds.) (2006). *Situational judgment test: Theory, measurement and application*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.