University of Kent

Kent Academic Repository

Brown, Anna, Morley, Emma and Patterson, Fiona (2026) *Item Response Modeling of Situational Judgment Tests With Items Nested Within Scenarios.* Educational and Psychological Measurement . pp. 1-29. ISSN 0013-1644.

Downloaded from
https://kar.kent.ac.uk/117466/ The University of Kent's Academic Repository KAR

The version of record is available from
https://doi.org/10.1177/00131644261487361

This document version
Publisher pdf

DOI for this version

Licence for this version
CC BY-NC (Attribution-NonCommercial)

Additional information

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

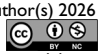
Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Check for updates

Original Research Article

Item Response Modeling of Situational Judgment Tests With Items Nested Within Scenarios

Educational and Psychological Measurement
1–29
© The Author(s) 2026

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644261487361
journals.sagepub.com/home/epm


Anna Brown[1,2], Emma Morley[2], and Fiona Patterson[2]

Abstract

Situational judgment tests (SJTs) are widely used to evaluate decision-making and problem-solving skills. These tests present respondents with situations (scenarios), and possible actions in response to each situation, which they must rate based on effectiveness or appropriateness. Despite having high face and criterion-related validities, these tests involve complex cognitive processes, and their construct validity has been notoriously difficult to establish. This is particularly challenging when several questions relate to the same scenario, thus creating local dependencies in item responses after controlling for the overall ability. Finding a suitable measurement model for SJTs would address the operational challenges of test development, maintenance, scoring, and equating the scores across multiple test forms and/or testing years. We propose to use a variant of the continuous response model (CRM; Samejima, 1973) to describe points gained at the level of scenario, regardless of the number of questions asked about the scenario. A simulation study testing the performance of this scenario-level model on data generated from categorical items nested within scenarios found that CRM fits well under a wide range of realistic conditions and resembles the true score better than a summated test score. We illustrate the proposed approach with an SJT presenting clinical dilemmas with several possible actions, each scored using four ordered categories. The SJT was completed by $N = 2{,}740$ examinees in operational assessment settings. The proposed CRM variant fitted the data well and was effective for assessing dimensionality, modeling and scoring responses, and evaluating test reliability. Implications for other operational uses such as form linking and pre-equating are discussed.

[1]School of Psychology, University of Kent, Canterbury, UK
[2]Work Psychology Group, Derby, UK

Corresponding Author:
Anna Brown, School of Psychology, Keynes College, University of Kent, Canterbury CT2 7NP, UK.
Email: a.a.brown@kent.ac.uk

Keywords

SJT, situational judgment test, continuous response model, item response theory, test equating

Introduction

Situational judgment tests (SJTs) are widely used in psychological assessment, personnel selection, and educational selection to evaluate decision-making and problem-solving skills. These tests typically present respondents with multiple situations (scenarios), each describing a problem or dilemma, and several possible actions in response to each situation, which they must rate (or rank) based on their own judgment of this action's effectiveness or appropriateness. Despite having high face and criterion-related validities, it has been notoriously difficult to establish the construct validity of these tests due to scenarios' complex and multifaceted content leading to multidimensionality (Lievens et al., 2008; Patterson et al., 2012).

Classical test theory (CTT), which is most often used for scoring and evaluation of SJTs' reliability and validity, is ill suited to address the dimensionality and construct validity challenges, because it equates the theoretical attribute with the way it is measured—that is, with the single score (Borsboom et al., 2004). Basic item response theory (IRT) models, which are commonly used in ability tests to model responses to test items conditional on person ability and item parameters such as difficulty, discrimination, and liability to guessing, struggle to accommodate the multifaceted content involved in SJTs. Their use is particularly challenging when several questions relate to the same scenario, thus creating local dependencies in item responses. In principle, such challenges can be overcome by looking beyond unidimensional IRT and modeling not a single latent ability but as many ability facets as needed based on the SJT content and as many method factors as needed to accommodate nesting of items within scenarios. However, such multidimensional IRT (MIRT) models are associated with operational and implementation challenges. They are complex to estimate, difficult to use for scoring and content management, particularly in large-scale educational testing where items must be rotated and replaced for test security reasons, and multiple forms need to be routinely anchored and pre-equated.

More recently, cognitive diagnosis models (CDMs) have been proposed to address the SJT's construct validity challenges (Sorrel et al., 2016). CDMs are psychometric models used to diagnose individuals' mastery of specific cognitive skills or knowledge components underlying test performance. Instead of estimating a single overall ability, CDMs use item-level responses to infer person's mastery or non-mastery on several discrete attributes or skills. With this, each item is typically assumed to require mastery on several such attributes; thus, CDMs are multidimensional at the item level. While this approach may be attractive and appropriate for learning about the test content and response process, or for scoring a fixed SJT, it does not have tangible advantages over MIRT models in large-scale assessment settings, where the demands are for simple item banking and measurement models.

Further search for a suitable and practicable measurement model for SJTs with items bundled within scenarios is warranted. This article explores the use of IRT at the level of scenarios and proposes to use a well-established model for bounded continuous data that can address many of the challenges and issues associated with the use of SJTs in large-scale operational testing, such as establishing dimensionality and latent structures, evaluating reliability, and providing a simple unidimensional measurement model that enables instant test scoring and pre-equating.

The rest of this article is organized as follows. We first describe the challenges of modeling responses to SJTs with rating items organized under multiple scenarios. Next, we describe a variant of the continuous response model (CRM, Samejima, 1973) and show how it can be applied to SJTs at the scenario level. We then present an extensive simulation study, manipulating six design factors (multidimensionality present at construct and testlet level, test and testlet size, number of response categories, and item difficulty) to generate item responses nested within scenarios to then analyze them at the scenario level using CRM. The simulation assesses the extent to which combining item scores into scenario scores eliminates multidimensionality of SJTs to be sufficiently described by a unidimensional CRM model, and the extent to which CRM recovers true scores on the construct that an SJT is designed to measure. This theoretical part is followed by an empirical example, where we illustrate the use of this variant of CRM with recent operational data gathered with a bespoke SJT in the context of medical assessment. We conduct a comprehensive quantitative and conceptual analysis of item responses and show the benefits of applying CRM to scenario-related item bundles compared to computationally heavy item-level alternatives. We conclude with limitations and recommendations for practice.

Method

Types of Multidimensionality in SJT Data

When an SJT intends to measure one common capacity (typically, decision-making and problem-solving capacity in the given professional or social environment), any form of multidimensionality presents a challenge. Multidimensionality can result from the ''local'' dependence of items relating to the same scenario (*within-scenario* specific factors), or from dependence between items relating to different scenarios (*between-scenario* specific factors). Within-scenario specific factors are almost inevitable in SJTs where each scenario is followed by several possible actions, and the test takers need to rate the appropriateness of each action. Responses to each action will likely correlate with responses to other actions within the same scenario even after controlling for the common capacity measured by the SJT simply because one's mental representation of the scenario will elicit more similarity between responses to actions that belong to it than actions that belong to different scenarios. In cognitive ability testing, this nesting of items within common passages is referred to as ''testlets,'' which typically need addressing in any item response modeling. Between-scenario specific factors can occur when some actions suggested as possible responses to

various scenarios involve similar behavior (say, altruistically helping others). While such a type of multidimensionality is intentional in SJTs that measure several competencies, it causes problems in SJTs that intend to measure just one capacity.

Scenarios as the Unit of Analysis

To reduce the dimensionality of SJT response data, we may consider scenario-level scores rather than item-level scores. This approach may help resolve the local dependencies created by several questions relating to the same scenario, as ''parcel'' scores in confirmatory factor models resolve the issue of correlated residuals of items with highly similar content. The approach may also help to deal with extra dimensions spanning across all scenarios, such as content similarity in some items (e.g., altruistically helping others). By considering whole scenarios as units of analysis, the unique sources of variance within them should be concealed and at the test level we expect to see only the common variance between the scenarios, measuring which is typically the intention of such SJTs.

To form scenario scores, sum scores of expert-scored items per scenario can be computed. The resulting scenario scores may have different theoretical score ranges depending on how many items they contain. For instance, SJTs used as illustrations in this article consist of scenarios followed by 4 to 7 rating items, each with the minimum expert-keyed score of 0 and the maximum of 3 or 4, yielding scenario sum scores with the theoretical minimum of 0 and varying theoretical maxima between 16 and 28. These are sufficiently large ranges of integer values to consider the scenario scores essentially continuous variables.

Measurement Models for Continuous Scenario Scores

Linear Factor Analysis. One way of analyzing continuous variables in the context of psychometric modeling is linear factor analysis (LFA). In LFA models, linear functions are assumed to describe the relationship between latent traits (what the test measures) and observed responses. That is, the observed score Y_{ij} of examinee j on scenario i is described as follows:

$$Y_{ij} = \mu_i + \lambda_i \theta_j + \varepsilon_{ij} \quad (1)$$

where μ_i and λ_i are the mean and factor loading of scenario i, respectively; θ_j is the score of person j on the latent trait; and ε_{ij} is the error.

However, when working with bounded continuous variables such as scenario scores using linear factor models (LFM), a problem arises that the model-predicted scores Y_{ij} can take values outside the range that is physically possible for the observed variables (from 0 to the scenario maximum), and therefore prediction at the extremes is less accurate (Ferrando, 2002). If enough scenarios have significant numbers of examinees achieving maximum possible scores, this can result in skewed distributions of the overall SJT theta score.

Continuous Response Model. To overcome this issue, a CRM was proposed by Samejima (1973). The basic idea of this model is that underlying each observed (continuous-limited) score Y, there is a continuous-unlimited response variable T for which the linear factor model (1) holds:

$$T_{ij} = \beta_i + \alpha_i\theta_j + \omega_{ij} \tag{2}$$

The variable T is a ''response tendency'' for the observed variable Y; it represents the unobserved scenario score that is distributed normally in the population of examinees. However, T cannot be directly observed since we cannot measure below the floor and above the ceiling and hence only observe Y, which is the manifested number of points achieved on the scenario. The relationship between T and the continuous-bounded observed score Y is given by the following logit transformation:

$$T_{ij} = \ln\left(\frac{Y_{ij}}{k_i - Y_{ij}}\right) \tag{3}$$

where k_i is the maximum score obtainable on scenario i (the minimum score is assumed 0). In theoretical formulations of CRM, the observed scores are assumed to never include the theoretical bounds. However, in SJT scores, the bounds can actually be achieved, so to avoid the division by 0 and taking ln(0) in Equation 3, in the proposed variant of the CRM we apply a small correction of 0.01, as suggested by Shojima (2003) to the observed scores if they reach the upper and lower bounds. Thus, we recode the score of 0 to 0.01, and the score k_i to $k_i - 0.01$ (e.g., score 16 is recoded to 15.99), leaving other scores intact.[1]

The logit transformation (3) brings the scenario scores (which can be on different scales depending on the number of items included) to the same scale since the ratio of achieved points to the points ''not achieved'' has the same meaning for any scenario. For example, on a scenario with max = 24 points, the score of 12, which is exactly half of all obtainable points, will result in the same ratio $12/(24 - 12) = 1$ as the score of 10 on a scenario with max = 20 points, $10/(20 - 10) = 1$. With this, the relationship between the continuous-bounded response Y and unobserved T is given by

$$Y_{ij} = k_i \frac{\exp(T_{ij})}{1 + \exp(T_{ij})}. \tag{4}$$

Since T is a function of the latent ability θ, Equation 4 can be written as a function of θ using Equation 2:

$$Y_{ij} = k_i \frac{\exp(\beta_i + \alpha_i\theta_j + \omega_{ij})}{1 + \exp(\beta_i + \alpha_i\theta_j + \omega_{ij})}. \tag{5}$$

Figure 1 illustrates relationships between the latent ability θ and the expected scenario scores in the empirical illustration presented later (where error $\omega_{ij} = 0$). While θ

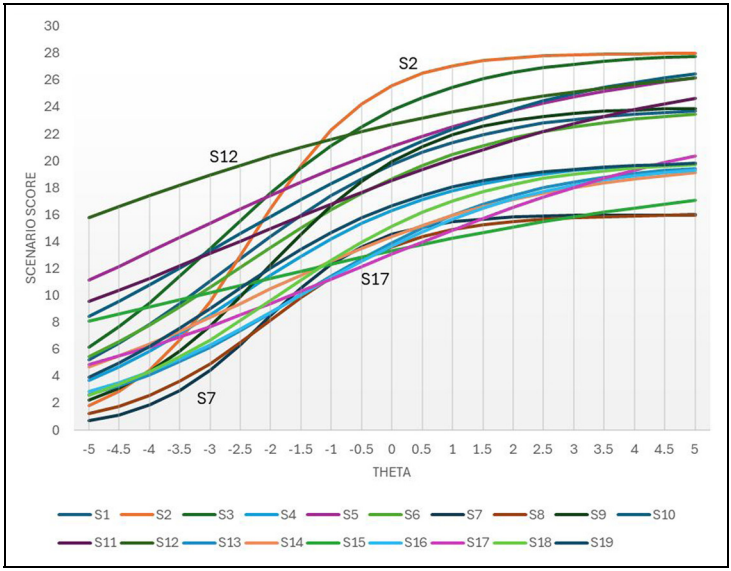


Figure 1. Expected scenario score conditional on ability under CRM in the empirical example.
Note. CRM = continuous response model.

(and T) is unbounded, the expected number of points on the scenario is bounded (in this case, by 0 from below and by 16–28 from above).

The CRM model was originally developed extending the graded response model (GRM) (Samejima, 1970) to infinite number of response categories and describes item characteristic functions conditional on the latent trait (theta), two item parameters: *difficulty* ($b = -\beta/\alpha$) and *discrimination* ($a = \alpha/SD(\omega)$), and a separate scaling parameter (α) linking the original rating score scale to the (θ) scale. The difficulty parameter (b) describes the level of ability at which the expected scenario score is $k_i/2$ (half of all points available), which is the point of inflexion of the item response function. This is the direct analogy of the difficulty parameter in the 2-parameter IRT model for binary items. The discrimination parameter (a) represents the slope at the theta level equal b.

Model Estimation. The attractive feature of CRM is its flexibility since it can be estimated in standard structural equation modeling packages (e.g., in the R package lavaan), first transforming the observed scenario scores into logits using Equation 3 and then applying usual LFA with maximum likelihood (ML) estimation. Specialist methods for model estimation and analysis of model fit have also been developed more recently, as well as other utilities enabling the use of CRM in practice. Specifically, the R package EstCRM (Zopluoglu, 2012) comprises useful features

including model estimation, person score estimation, and evaluation of model fit (Ferrando, 2010).

Person Scores, Standard Errors of Measurement and Reliability. Once the item (here, scenario) parameters are known, the ability level for person j is the maximum likelihood estimate (MLE) using the following closed expression (Shojima, 2005):

$$\mathrm{MLE}(\theta_j) = \frac{\sum_{i=1}^{m}\left[a_i^2\left(b_i + T_{ij}/\alpha_i\right)\right]}{\sum_{i=1}^{m} a_i^2} \tag{6}$$

In the EstCRM package, the MLE estimates are re-scaled to have a mean zero and Standard Deviation unity (Zopluoglu, 2022), so the rescaled ability scores $\hat{\theta}_j$ are on the same scale as the latent variable θ_j.

The information for item i in CRM is simply the squared discrimination parameter, $I_i(\theta_j) = a_i^2$ (Samejima, 1973), and the standard error of ability estimate $\hat{\theta}_j$ is the square root of the reciprocal of the total test information, which is the sum of information for all m scenarios:

$$SE(\hat{\theta}_j) = 1/\sqrt{a_1^2 + a_2^2 + \ldots + a_m^2} \tag{7}$$

Unlike in binary or polytomous item response models, the SE of the ability estimate in CRM is the same for all examinees unless an examinee has missing data. If the examinee has missing data, the ability level estimate and its standard error are obtained from the available item scores (Zopluoglu, 2022). When the SEs are the same for all examinees in the sample, reliability of theta score can be computed using the CTT definition, as the ratio of the true score variance to the observed score variance, which is true variance plus error variance. Since the true theta score variance is set to unity, the reliability of the estimated $\hat{\theta}$ is:

$$\rho(\hat{\theta}) = 1/\left(1 + SE^2(\hat{\theta})\right) \tag{8}$$

When the SEs for examinees are different (due to missing responses), the *empirical reliability* index can be computed by averaging squared SEs for the sample (du Toit, 2003):

$$\rho(\hat{\theta}_j) = 1/\left(1 + \overline{SE^2}(\hat{\theta}_j)\right) \tag{9}$$

Operational Uses of CRM. The use of CRM with scenario scores opens many possibilities. For instance, CRM can be adopted as the measurement model for SJT scoring and equating. In terms of scoring, the CRM can be used to estimate examinee theta scores and their standard errors (and thus the test score reliability). In terms of equating/linking, the CRM enables all IRT-based approaches, such as true-score IRT equating or ''fixed-item parameter calibration'' (FIPC). In FIPC, which may be the most natural and easy-to-implement pre-equating method in many SJT testing environments with ''Non-Equivalent group with Anchor Test'' (NEAT) designs,

parameters for all scenarios are estimated in the current year (or in the first test form) and then fixed for the anchor scenarios carried over to the subsequent year (or to the alternative test forms). These fixed anchor parameters carry the scale of measurement to the new forms administered, enabling free estimation of parameters for the new scenarios as well as free estimation of the ability level of the new examinees, not making any assumptions about the distribution of ability in the subsequent years (González & Wiberg, 2017). When multiple test forms are used in a large-scale assessment NEAT design, concurrent calibration can also be used, whereby item parameters for all forms are estimated at once using ML estimation of missing-by-design data from all examinees.

Simulation Study

This simulation study was designed to investigate the suitability of CRM to typical SJT data with rating items nested within scenarios. We emphasize that by applying CRM to summated scenario scores, we intentionally ignore complex item structures to simplify operational test scoring. Therefore, the objective of this study is to investigate how this simplified model performs on item data generated with the true (complex) nested model.

This simulation study is somewhat unusual in that it is not intended to recover item parameters because the fitted model, which considers the scenario summated scores only, is by design mis-specified, ignoring the item level of the generating model. However, we are interested in the impact of misspecification on model fit, which would reflect the extent to which combining item scores into scenario scores eliminates multidimensionality of the SJT to be sufficiently described by a unidimensional CRM model. But the primary focus of the simulation is on the recovery of person parameters (true scores on the construct that an SJT is designed to measure) under varying degrees of testlet structure and local dependence, with other factors included to ensure ecological validity.

To appraise the model fit, we analyzed 100 replications of $N = 1{,}000$ cases per simulation condition ($64 \times 100 = 6{,}400$ samples in total). This should provide a realistic picture of how the CRM model would fit typical data collected in practice. To appraise the recovery of person parameters, we analyzed one giant sample of $N = 100{,}000$ cases per simulation condition (64 giant samples in total). This should allow examining distributions of true generating ability scores against estimated CRM scores essentially at the population level, rather than through averaging smaller samples with their random variations. Analyses were conducted in R 4.4.1 (R Core Team, 2023) and Mplus 9.1 (Muthén & Muthén, 2017). All data generation and analysis scripts are available on osf.io/dn42h

Simulation Study Design

SJTs used in educational contexts exhibit substantial heterogeneity in their structural design, particularly in the configuration of scenarios, items, and response formats. Operational assessments such as the AAMC PREview Professional Readiness Exam (Ellison et al., 2024) and the University Clinical Aptitude Test (Patterson et al., 2017) typically comprise between 15 and 30 scenarios, with roughly 4 to 7 potential actions (items) following each scenario, although both shorter and longer item clusters are sometimes observed. Response formats used for rating appropriateness of each action vary, with 4- or 5-category Likert-type scales widely used, and occasionally longer 7-point scales (Kepes et al., 2025). Since SJT items are often scored using expert-defined keys, the original response options may not convert into the same number of scored options. For instance, the empirical example in this article showcases a scoring protocol whereby 5-point responses are converted into 4-category scores after expert keys are applied. At the construct level, SJTs are typically designed to sample multiple correlated attributes, yet are often operationally scored as one construct, creating potential model–data mismatch. In addition, items within scenarios exhibit local dependence due to shared context, the magnitude of which varies across instruments and scenarios, and observed score distributions are often skewed toward higher performance due to item targeting and socially desirable responding.

Accordingly, the present simulation study varies six key design factors reflecting this observed heterogeneity: (a) latent dimensionality, (b) number of scenarios, (c) number of items per scenario, (d) magnitude of scenario-level dependence, (e) number of item score categories, and (f) item difficulty targeting to reproduce both well-targeted and ceiling-skewed score distributions. These factors, their levels and rationale are given in Table 1. Fully crossed, the simulation design results in 64 conditions.

Data Generation

Our approach was identical for both generating giant samples for examining person parameter recovery and generating smaller samples for model fit appraisal. First, true scores on one general ''ability'' and four further specific factors uncorrelated with one another were randomly sampled from standard normal $Z \backsim N(\mu = 0, \sigma = 1)$. With this approach, the same variables can be used for generating either unidimensional structure or multidimensional correlated factors by adding influences from both general and specific factors, as we explain below. Second, 20 random testlet effects (enough for the largest number of scenarios) were sampled from $N(\mu = 0, \sigma = 1)$. Lastly, 120 random unique effects/error terms (enough for the largest number of items $20 \times 6 = 120$) were sampled also from $N(\mu = 0, \sigma = 1)$.

With these ''true scores,'' which were generated just once for $N = 100{,}000$ cases to form the giant sample, and 100 times for $N = 1{,}000$ cases to form smaller samples, we produced continuous item response tendencies by weighting all factors (i.e.,

Table 1. Factors Manipulated in the Simulation Study.

Factor (R notation)	Conditions	Rationale
1. Latent dimensionality	1D; 4D (ρ = .5)	Reflects unidimensional (1D) vs. intended multidimensional (4D) structure with factors correlated at .5.
2. Number of scenarios	10, 20	Reflects short (10) vs. typical (20) tests
3. Items per scenario (testlet size)	4, 6	Reflects typical (4) vs. dense (6) clusters
4. Scenario local dependence (testlet loading)	0.0, 0.32	Reflects absent (0.0) vs. moderate (0.32) local dependence. Loading 0.32 translates to residual item correlations of .1.
5. Response categories	2, 4	Reflects dichotomous (2) and polytomous (4) item scores
6. Item difficulty targeting	Well-targeted; easy (ceiling)	Reflects normal vs. negatively skewed score distributions, created by average vs. low category thresholds

general, specific, testlets, and errors) by coefficients chosen according to specific simulation conditions as described below.

In the 1D latent trait condition, we created a continuous item response tendency T_{ikj} of person j for item i nested within scenario k, by using a linear combination of true person's trait/ability θ_j (with factor loadings α_{ik} differing for each scenario and item—see section ''Item Parameter Generation''), person's testlet effect t_{kj} (with either 0 or 0.32 loading according to levels in factor #4), and the unique effect ω_{ikj}, as follows:

$$T_{ikj} = \begin{cases} \alpha_{ik}\theta_j + \omega_{ikj}, & \text{if testlet effect is 0} \\ \alpha_{ik}\theta_j + 0.32t_{kj} + \omega_{ikj}, & \text{if testlet effect is 0.32} \end{cases} \quad (10)$$

In the 4D latent trait condition, person's j scores on four correlated factors f^1, f^2,f^3,f^4 were created via a second-order structure with equal influences from one general and four specific dimensions s^1, s^2, s^3, s^4 as follows: $f_j^d = \theta_j + s_j^d$. When all—the general and specific factors—have unity variances and uncorrelated with one another, correlations between any two factors, for example f^1 and f^2, equal 0.5 as intended in the 4D condition:

$$\mathrm{cor}\left(f^1, f^2\right) = \frac{\mathrm{cov}(f^1, f^2)}{SD(f^1)SD(f^2)} = \frac{\mathrm{var}(\theta)}{\sqrt{(\mathrm{var}(\theta) + \mathrm{var}(s_1))(\mathrm{var}(\theta) + \mathrm{var}(s_2))}} = \frac{1}{\sqrt{(1+1)(1+1)}} = 0.5 \quad (11)$$

Thus, in the 4D condition, item response tendency T_{ikj} of person j for item i nested within scenario k was presented by extending the 1D linear combination above as follows:

$$T_{ikj} = \begin{cases} \alpha_{ik}\left(\theta_j + s_j^d\right) + \omega_{ikj}, & \text{if testlet effect is 0} \\ \alpha_{ik}\left(\theta_j + s_j^d\right) + 0.32t_{kj} + \omega_{ikj}, & \text{if testlet effect is 0.32} \end{cases} \quad (12)$$

In the 4-item-per-testlet condition, item i within each testlet loaded on a specific dimension s^i. In the 6-item-per-testlet condition, all items were sequentially loaded on s^1, s^2, s^3, s^4, repeating this pattern across all testlets.

To produce categorical item responses typically observed in SJTs, we first standardized the continuous response tendencies and then cut them into two or four categories (according to levels in factor #5) using one or three thresholds, respectively—see section ''Item Parameter Generation.''

Item Parameter Generation

Two sets of thresholds were generated reflecting two levels of item difficulty targeting (factor #6). Set 1 contained well-targeted items with the first (leftmost) thresholds sampled from $N(\mu = -0.8, \sigma = .6)$, the increments from first to second threshold sampled from $N(\mu = 1.0, \sigma = .2)$, and the increments from second to third threshold sampled from $N(\mu = .6, \sigma = .2)$. Positive increments rather than actual thresholds were sampled to ensure the thresholds are ordered. Set 2 contained ''easy'' items with the first (leftmost) thresholds sampled from $N(\mu = -1.9, \sigma = .6)$, increments from the first to the second threshold sampled from $N(\mu = .7, \sigma = .2)$, and increments from the first to the second threshold sampled from $N(\mu = .9, \sigma = .2)$.

Only one set of factor loadings λ_{si} representative of existing SJTs was generated regardless of the item difficulty condition. These were generated from $N(\mu = .5, \sigma = .2)$. The generated parameters are given in the Supplemental Table S1.

Evaluating Model Fit With the 100-Replications Approach

For each of the 64 fully-crossed simulation conditions, 100 replications with $N = 1{,}000$ were used to generate response tendencies and categorical item scores as described above; then, scenario scores were computed and submitted to CRM estimation. To evaluate how the design factors influenced the fit of CRM, we examined four popular indices: the χ^2/df ratio, the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Standardized Root Mean Square Residual (SRMR). Three complementary analyses were conducted. First, marginal means of each fit index were computed for every level of every factor,

averaging over all other factors. These provided a straightforward summary of the direction and magnitude of each factor's main effect.

Second, a fully crossed factorial ANOVA was performed for each fit index, including all main effects and two-way interactions between design factors. Because the design contained only 64 unique conditions (i.e., $N = 64$ observations), we used partial omega-squared (ω_p^2) as an unbiased effect size estimate (Lakens & Bakker, 2013; Olejnik & Algina, 2003). Partial ω^2 quantifies the proportion of variance uniquely accounted for by a factor or interaction, corrected for small sample, and allows direct ranking of the relative importance of the design features.

Third, interaction plots were constructed using the layout that emerged as most informative from the effect size rankings. The dominant factors were mapped to the *x*-axis, line color, and line type, while the remaining factors defined the facets. All figures were generated using R package ggplot2 (Wickham, 2016).

Evaluating True Score Recovery With the One-Giant-Sample Approach

For each of the 64 fully-crossed simulation conditions, a single replication with $N = 100{,}000$ was used to generate response tendencies and categorical item scores as described above, and finally scenario scores were submitted to CRM item and person parameter estimation. To evaluate correspondence between true (θ) and CRM-estimated ability scores ($\hat{\theta}$) in each condition, three aspects of score recovery were examined: the magnitude and pattern of estimation error, the functional form of the $\theta - \hat{\theta}$ relationship, and rank-order preservation.

First, estimation error was quantified as the difference $\varepsilon = \hat{\theta} - \theta$. Overall bias was computed as the mean error (ME), and overall accuracy as the root mean squared error (RMSE). To detect systematic deviations across the ability continuum, conditional bias was examined by dividing the θ distribution into deciles using the ntile function from dplyr (Wickham, et al. 2026) and calculating the ME within each decile. The maximum and the mean absolute conditional ME were recorded as summary indices of conditional bias.

Second, to determine whether the regression of $\hat{\theta}$ on θ followed a linear trajectory as expected, we fitted a linear regression model to all conditions using ordinary least squares. The relative fit was evaluated via the Akaike information criterion (AIC). A lower AIC for a particular condition was interpreted as evidence of a relationship closer to linear. A difference of at least 10 between models is usually considered decisive evidence (Burnham & Anderson, 2004).

Third, rank-order preservation was assessed using Spearman's rank-order correlation (ρ) between θ and $\hat{\theta}$. Spearman's coefficient was preferred over Pearson's *r* because it directly measures the consistency of the rank ordering, which is the property of interest for score interpretation when strict linearity is not assumed.

Table 2. Marginal Means (Standard Errors) of Fit Indices Across 100 Replications by Simulation Condition.

Factor	Level	Chi^2/df	CFI	RMSEA	SRMR
multidim	0 = 1D	1.047 (0.005)	0.996 (0.000)	0.007 (0.000)	0.020 (0.000)
	1 = 4D	1.224 (0.028)	0.992 (0.001)	0.013 (0.001)	0.021 (0.000)
testlet_n	10	1.153 (0.030)	0.994 (0.001)	0.011 (0.001)	0.019 (0.000)
	20	1.117 (0.019)	0.994 (0.001)	0.009 (0.001)	0.022 (0.000)
testlet_size	4	1.078 (0.009)	0.995 (0.000)	0.008 (0.000)	0.020 (0.000)
	6	1.193 (0.032)	0.993 (0.001)	0.012 (0.001)	0.021 (0.000)
testlet_load	0	1.161 (0.030)	0.994 (0.001)	0.011 (0.001)	0.020 (0.000)
	0.32	1.109 (0.019)	0.994 (0.001)	0.009 (0.001)	0.021 (0.000)
categories	4	1.152 (0.027)	0.994 (0.001)	0.011 (0.001)	0.020 (0.000)
	2	1.118 (0.023)	0.994 (0.001)	0.009 (0.001)	0.021 (0.000)
easy	0	1.128 (0.024)	0.995 (0.000)	0.010 (0.001)	0.019 (0.000)
	1	1.142 (0.027)	0.993 (0.001)	0.010 (0.001)	0.022 (0.000)

Results

Model Fit Across Simulation Conditions (100 Replications Approach)

Marginal Means of Fit Indices. Table 2 displays the marginal means and standard errors of each fit index averaged over all other factors. Overall, the CRM model showed excellent fit across all conditions, with mean Chi-squared/df ratio ranging from 1.047 to 1.224, RMSEA ranging from 0.007 to 0.013, CFI from 0.992 to 0.996, and SRMR from 0.019 to 0.022. Figure S1 in the Supplemental Material shows heatmaps of fit indices for all 64 simulation conditions. According to χ^2/df ratio, which is the most differentiating among the indices, the worst conditions for CRM fit are: 4D with larger testlets (6 items rather than 4). Even in most unfavorable conditions, χ^2/df ratio reached only 1.59 at its worst; RMSEA and SRMR never exceeded 0.02, and CFI never dropped below 0.98.

Factorial ANOVA and Interaction Patterns. Partial omega-squared values for all main effects and two-way interactions are reported in Supplemental Table S2. Most important main effects were latent dimensionality (Factor #1), which uniquely accounted for large proportions of variance in Chi-squared/df ratio (ω_p^2 = .69), RMSEA (ω_p^2 = .82), CFI (ω_p^2 = .82), and less so in SRMR (ω_p^2 = .44). Testlet size (Factor #3) had an impact on Chi-squared/df ratio (ω_p^2 = .48), RMSEA (ω_p^2 = .64) and CFI (ω_p^2 = .47), but not SRMR (ω_p^2 = .00). The number of scenarios (Factor #2) was uniquely important for SRMR (ω_p^2 = .80) but no other fit indices. Finally, item difficulty (Factor #6) had an important influence on CFI (ω_p^2 = .57) and SRMR (ω_p^2 = .76) but no other fit indices. Most notable interactions were between Factor #1 and Factor #3 (dimensionality: testlet size), and Factor #5 and Factor #6 (item categories: item difficulty). Line plots in Supplemental Figure S2 display the fit index as a function of latent dimensionality, clearly showing the interaction patterns.

Table 3. Marginal Means (Standard Errors) for True Score Recovery Indicators by Simulation Factor.

Factor	Level	RMSE (*SE*)	Mean abs cond. bias (*SE*)	Max abs cond. bias (*SE*)	Spearman rho (*SE*)
multidim	0	0.430 (0.015)	0.094 (0.005)	0.296 (0.025)	0.908 (0.007)
	1	0.613 (0.010)	0.156 (0.004)	0.434 (0.021)	0.809 (0.006)
testlet_n	10	0.568 (0.018)	0.142 (0.007)	0.409 (0.026)	0.836 (0.011)
	20	0.475 (0.020)	0.108 (0.006)	0.320 (0.024)	0.881 (0.010)
testlet_size	4	0.543 (0.020)	0.131 (0.008)	0.379 (0.028)	0.846 (0.011)
	6	0.500 (0.021)	0.118 (0.007)	0.351 (0.023)	0.871 (0.011)
testlet_load	0	0.507 (0.021)	0.121 (0.007)	0.355 (0.026)	0.867 (0.011)
	0.32	0.536 (0.020)	0.128 (0.007)	0.374 (0.026)	0.850 (0.011)
categories	2	0.549 (0.020)	0.129 (0.008)	0.405 (0.026)	0.841 (0.011)
	4	0.494 (0.021)	0.120 (0.007)	0.324 (0.024)	0.877 (0.010)
easy	0	0.487 (0.020)	0.111 (0.007)	0.261 (0.018)	0.876 (0.010)
	1	0.556 (0.019)	0.138 (0.007)	0.469 (0.018)	0.841 (0.011)

Note. RMSE = root mean squared error.

Summary of Fit. Taken together, the CRM model's fit to scenario scores, which were derived from categorical items nested within scenarios, was primarily driven by latent dimensionality, testlet size, and their interaction. The number of categories or in-testlet dependence produced almost no change in fit, while the number of testlets were uniquely impacting only SRMR, and item difficulty only impacting CFI ad SRMR. However, since fit indices showed favorable marginal means and very little variation across all conditions, we conclude that the CRM is likely to be retained in cases when categorical items nested in testlets measure something in common.

True Score Recovery Across Simulation Conditions (One Giant Sample Approach)

Table 3 reports the marginal means and standard errors for all key indicators of true score recovery (RMSE, conditional bias, AIC, and Spearman rho), and Figure S3 in the Supplemental Material displays a heatmap of these indicators for all conditions. Table S3 in the Supplemental Material reports partial omega-squared values for all effects of simulation factors on the score recovery indicators, and line plots in Figure S4 show corresponding interactions.

Estimation Error. ME was uniformly negligible (–0.0007) across all conditions, reflecting the mean-preserving nature of the CRM scoring algorithm. Consequently, we focus on RMSE and conditional bias to quantify estimation accuracy. RMSE varied markedly by design features (range: 0.245–0.743, median = 0.547) as heatmaps in Figure S3 show. Overall, longer 1D tests with larger testlets and more response

categories showed smaller errors. A factorial ANOVA attributed the largest unique variance in RMSE to latent dimensionality (ω_p^2 = .99), followed by number of testlets (ω_p^2 = .96). Other main effects also yielded high partial omega squared, indicating that all factors were important to the score recovery. The most notable interaction effect was between latent dimensionality and number of testlets (ω_p^2 = .62).

Analysis of conditional bias—ME within deciles of θ—revealed that the CRM model systematically over-estimated low ability individuals (mean bias in the 1st decile = 0.227 across conditions) and underestimated high-ability individuals (mean bias in the 10th decile = –0.279), see Figure S5 in the Supplemental Material. Mean absolute conditional bias ranged from 0.038 to 0.223 across conditions (median = 0.125). The factorial ANOVA revealed that latent dimensionality (ω_p^2 = 0.81), number of testlets (ω_p^2 = 0.55) and item difficulty (ω_p^2 = 0.44) were the strongest determinants of mean absolute conditional bias, with lower (better) values under 1D, 20 testlets, and moderate item difficulty. Both the largest positive bias (0.532, 1st decile) and the largest negative bias (–0.702, 10th decile) occurred under 4D, fewer testlets, in-testlet dependency, and overly easy items. In these least favorable conditions for CRM, extreme true scores were most attenuated toward the mean. Supplemental Figure S6 shows the relationship between θ and $\hat{\theta}$ at the midpoint of each decile of theta in the worst condition, where the regression to the mean is clearly visible. In contrast, the correspondence is almost perfect in the best condition, also shown in Figure S6.

Shape of the θ–$\hat{\theta}$ Relationship. As the heatmap in Supplemental Figure S3 shows, AIC was smallest (indicating best relative fit to a linear model) in longer 1D tests with larger testlets, more response categories, less within-testlet dependency, and well-targeted difficulties. Figure 2 presents two scatter plots: one between true θ and estimated $\hat{\theta}$, and the other between true θ and summated test score, in this most favorable condition for N = 100,000 cases. While the θ–$\hat{\theta}$ relationship follows a linear trajectory, the summated test score has a clear logistic shape of relationship with θ. This resembles test characteristic functions in categorical IRT models, illustrating why the logit transformation of summated categorical item responses employed by CRM is beneficial.

Rank-Order Preservation. As shown in the heatmap in Figure S3, Spearman's rho ranged from 0.714 (4D, in-testlet dependence, short test with fewer items in each scenario, two response categories, easy items) to 0.974 (1D, no in-testlet dependence, long test with more items per scenario, four categories and moderate item difficulty), with a median of 0.850. According to ANOVA, all factors had strong independent influences ($\omega_p^2 > .89$) on rank-order preservation and so did some of their interactions (see line plots in Figure S4).

Summary of Recovery Influences. Overall, the CRM model applied to scenario sum scores recovered the true scores with high fidelity under most conditions. The

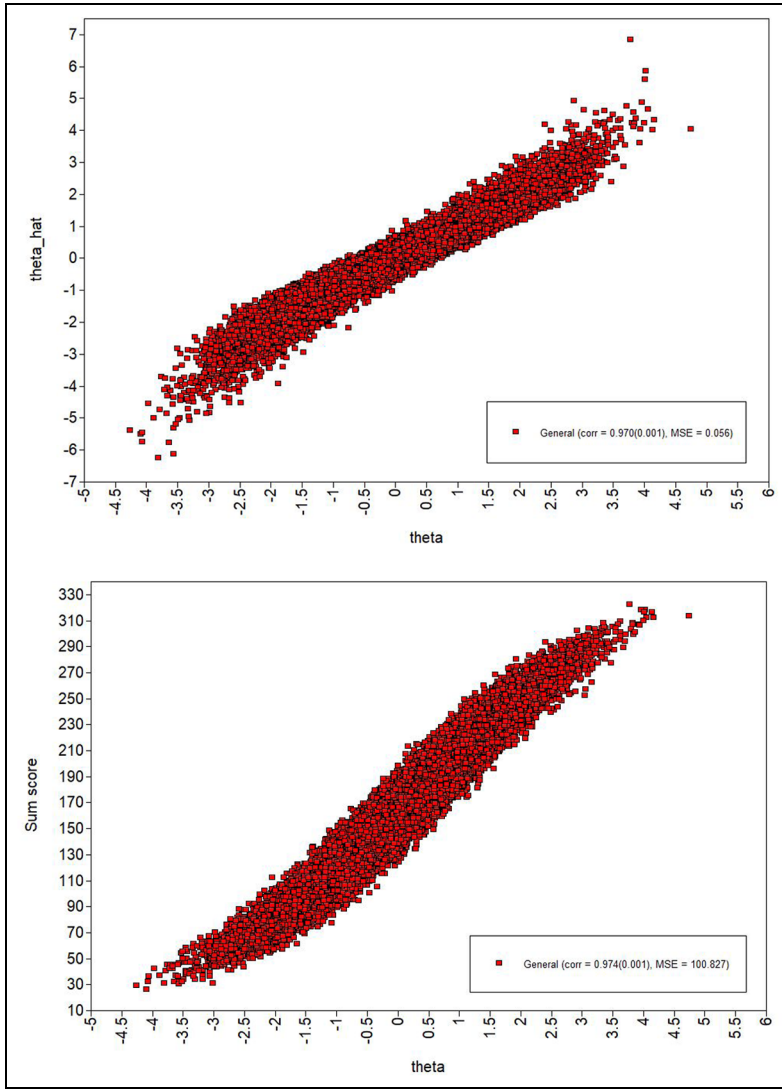


Figure 2. Relationship between θ and $\hat{\theta}$, and θ and sum score, in the most favorable condition for true score recovery (1D, testlet load = 0, 20 testlets, 6 items per testlet, 4 categories, moderate difficulty items).

patterns of all score recovery indices (RMSE, conditional bias, AIC for $\hat{\theta} - \theta$ regression, and rank-order correlations) across conditions were remarkably similar, as heatmaps in Supplemental Figure S3 and ANOVA Table S3 show. The most favorable

conditions for true score recovery were single measured dimension, longer tests, more response categories, and moderate difficulty targeting. Under such conditions, the rank ordering of individuals remained almost perfectly preserved ($\rho > .95$, corresponding to reliability greater than .90), and the regression shape was linear with virtually no compression at the trait extremes. These findings underscore the practical viability of the scenario-level CRM model for score estimation, provided the assessment context avoids the most adverse design combinations identified here. We note that under most unfavorable conditions (multidimensionality, short test, binary responses and poorly targeted item difficulty), the score precision would be equally poor using summated test scores.

Empirical Illustration

To illustrate the use of CRM in SJTs where items are nested within scenarios, we analyzed $N = 2{,}740$ operational responses to a bespoke SJT. The SJT intends to assess an applicant's ability to respond appropriately to a range of professional dilemmas set in the hospital environment. The test consists of multiple scenarios describing such dilemmas, all with several nested items, each representing behavioral responses to the given dilemma.

Participants

The sample analyzed consisted of $N_1 = 2{,}740$ applicants to postgraduate medical training, who took the SJT in 2023 in the English language and met the operational filter set for this test (having no more than six non-attempted items). There were 64.9% women, 33.3% men, and the remaining 1.8% did not disclose their gender.

Materials

The bespoke SJT consisted of 109 items arranged under 19 scenarios, with 4 to 7 items per scenario. The examinees were asked to rate each item according to how desirable such a behavioral response would be in the given situation, using a 5-point scale: 1 (*Very desirable*), 2 (*Somewhat desirable*), 3 (*Neither desirable nor undesirable*), 4 (*Somewhat undesirable*) and 5 (*Very undesirable*). A sample scenario is given below.

> Sharmin is working on a ward in a hospital during her placement. Over the past week, Sharmin has observed Larry, a nursing student, share his opinion with many patients about the medication that doctors on the ward have prescribed. He tells the patients that they should opt for herbal remedies instead of the prescribed medication because prescribed medication is ''full of toxic side-effects.'' He has also given strong recommendations to patients to refuse taking the prescribed medication.

How desirable are each of the following responses by Sharmin in this situation?

A. Tell Larry that he should not be discussing choice of medication with patients.
B. Tell a senior colleague on the ward about her observations of Larry.
C. Advise the patients that they should not trust Larry's judgment about their medication.
D. Explore Larry's reasons for giving his recommendations to patients.

In operational use, these raw item responses are scored using expert keys, which assign between 0 and 3 or 4 points using a ''near-miss'' scoring convention according to how close the response was to the key.

Procedure

Applicants registered to take the SJT using a unique identifier and sat the test through a remote testing platform. Applicants had 75 min to complete the test as standard, with additional time available as a reasonable adjustment. The tests were remotely proctored, using image capture technology.

Analysis Overview

To explore the dimensionality of responses (how many factors—including both substantive constructs and methodological artefacts—underlie the item sets), we carried out exploratory item response modeling (aka exploratory factor analysis [EFA] with categorical outcomes). Raw item responses as well as expert-scored item responses were considered as units of analysis. These analyses were carried out using general modeling software Mplus (Muthén & Muthén, 2017). Based on the results of these analyses, we discussed suitable measurement models for item-level data from the MIRT family.

Next, we considered the use of scenarios as units of analysis, and repeated the analyses of dimensionality, reliability and measurement model evaluation looking at sums of expert scores for item bundles (scenarios) rather than individual items. These analyses were carried out using two different approaches. First, the raw scenario scores were logit-transformed as per Equation 3 and subjected to confirmatory factor analysis using LFA in Mplus. This approach allows obtaining goodness of fit (GOF) indices, which can be compared to GOF of LFA with non-transformed scenario scores. Significant improvement of model fit would suggest benefits of using CRM rather than LFA. Second, the raw scenario scores were submitted to CRM estimation using the R package EstCRM. This software provides convenient functionality for

estimating model parameters in the conventional CRM parametrization and estimating person SJT scores and their standard errors.

Considering Item Responses as Units of Analysis

EFA of Item Responses. EFA of the expert-scored SJT rating items, with responses analyzed as ordered categories (ranging from 0 for least ''correct'' to 3 or 4 for most ''correct'' response) using the ML estimation with logistic link, yielded the first 10 eigenvalues of 11.72, 8.35, 5.79, 2.74, 2.65, 2.28, 2.12, 1.96, 1.68, and 1.60, suggesting three major factors explaining similar amounts of variance, plus about five further minor factors. Essentially the same solution emerged from the analysis of raw (unscored) rating items. Scree plots for both expert-scored and raw responses are depicted in Supplemental Figure S7 (panels A and B, respectively).

To explore the nature and meaning of the major factors underlying the expert-scored data, an exploratory three-factor solution with the Oblimin (oblique) rotation was obtained and interpreted. The three factors spanned across all scenarios. Factor 1 was indicated most strongly by items that describe some advice or arbitration that the given situation calls for (e.g., ''Advise Asha to speak to her supervising physician about …''). Factor 2 was indicated by items that describe amicable or compliant behavior (e.g., ''Consider ways to make Olivia feel more welcome.''). Factor 3 was indicated by items that describe some confrontational behavior (e.g., ''Inform others in the team about Asha's marital problems.''). While Factor 2 was dominated by items constituting ''appropriate'' behaviors according to the scoring key, Factor 3 consisted largely of ''inappropriate'' behaviors.

EFA of the raw (unscored) item responses revealed the same three factors emerging, with essentially the same mapping of factors to items and predictable differences in factor loadings (i.e., reversed loadings for ''inappropriate'' behaviors). This suggests that the major factors found are not artefacts of expert-defined scoring but are based on the test content. Further minor factors could be attributed to local dependencies within scenarios. This was clear from examining the model's residual correlations, where largest residuals were observed for pairs of items relating to the same scenario. An important feature of both EFA solutions for raw and expert-scored data was that many items cross-loaded on two or even three factors, highlighting the challenges in finding a parsimonious measurement model for these data.

Measurement Models for Item Responses. To address the ordered-categorical nature of raw rating data, multidimensional variants of the generalized partial credit model (GPCM; Muraki, 1992) that allow for varying item discriminations and non-ordered category thresholds can be fitted, using ML estimation with numerical integration. Alternatively, multidimensional variants of the GRM (Samejima, 1970) can be fitted, again using ML estimation with numerical integration or less computationally demanding least squares estimation. The same models can be fitted to the ordered

categorical *expert-scored* item data. In either case, to model the three substantive factors, items with strong loadings on that factor can be identified from EFA and used as indicators. The first challenge in this MIRT modeling will be cross-loadings evident from the EFA. The second challenge will be the common scenarios potentially causing local dependencies between their items after controlling for the substantive factors (aka ''testlet effects''). To model these local dependencies, 19 additional factors uncorrelated with the substantive factors and with each other would be required. In this model with testlets, the number of dimensions is large (3 substantive traits + 19 testlets = 22 factors), therefore when using the ML estimation, Montecarlo integration with randomly selected integration points would be necessary. Although technically available and achievable with general SEM software such as Mplus, fitting such multidimensional models does not address the objective of establishing a simple measurement model enabling easy scoring and equating/linking.

Considering Scenario Scores as Units of Analysis

To reduce the dimensionality of response data, we now consider scenario scores rather than individual item scores. Since the three major dimensions found in item responses span across all scenarios, they must be item (rather than scenario) specific. By considering whole scenarios as units of analysis, the item-specific sources of variance within them cannot be separated and all that should be visible at the test level is the common variance between the scenarios. To form the scenario scores, expert-defined item scores within each scenario were summed. The resulting 19 scenario scores have a theoretical range between 0 and 16 to 28 depending on the number of items within each scenario and represent essentially continuous (but bounded) data.

EFA of Scenario Scores. When the scenario scores were submitted to linear EFA, one major factor and a minor second factor clearly emerged with the first five eigenvalues of 4.76, 1.58, 1.06, 1.00, and 0.89 (for scree plot, see Supplemental Figure S8). A single-factor linear model fitted acceptably according to RMSEA and SRMR but poorly according to CFI; RMSEA = .058, 90% CI [0.056, 0.061], CFI = 0.839, SRMR = 0.052. A two-factor exploratory model fitted well according to all fit indices; RMSEA = 0.033, 90% CI [0.030, 0.036], CFI = 0.952, SRMR = 0.024. In the two-factor model with bifactor rotation (which maximizes communalities for the first factor and all remaining variance is assigned to the second factor uncorrelated with the first), all scenarios loaded on the first factor positively and significantly with standardized loadings between 0.35 and 0.58, and the second factor was defined by three scenarios with non-trivial standardized loadings –0.43, –0.32, and 0.36. On inspection, all three scenarios referred to hospital activities involving ''fellow residents'' and thus probably caused content-related local dependencies. It appears that

Table 4. CRM Parameter Estimates for 19 Rating Scenarios in the Empirical Example.

Scenario	*a*	*b*	Alpha
S1	0.415	–2.721	.564
S2	0.530	–2.352	1.009
S3	0.469	–2.874	.599
S4	0.418	–2.516	.595
S5	0.447	–3.638	.307
S6	0.434	–2.530	.496
S7	0.419	–2.117	1.080
S8	0.430	–2.051	.848
S9	0.497	–2.050	.778
S10	0.618	–2.728	.369
S11	0.488	–2.524	.267
S12	0.392	–6.077	.239
S13	0.448	–1.525	.545
S14	0.411	–2.225	.425
S15	0.340	–3.207	.215
S16	0.500	–1.511	.514
S17	0.354	–0.579	.309
S18	0.467	–1.878	.607
S19	0.353	–2.680	.604

Note. CRM = continuous response model.

the secondary factor is not meaningful nor important and can be ignored in analysis and scoring.

CRM Fitted to Rating Scenarios

Model Fit and Parameter Estimation. We tested the CRM with 19 rating scenarios using two alternative methods. In Method 1, we logit-transformed the scenario scores using Equation 3 and fitted a simple linear one-dimensional CFA model as described in Equation 2 using Mplus. This is equivalent to fitting the CRM to observed scenario scores. In Method 2, we submitted the scenario scores to the CRM parameter calibration function EstCRMitem() from package EstCRM (Zopluoglu, 2022). Both methods yielded identical model parameters. Table 4 reports the item (scenario) parameters in the CRM-specific parameterization obtained from package EstCRM. Figure 1 illustrates the expected scores on the 19 scenarios conditional on the SJT theta score according to the fitted CRM model (we can call these functions ''scenario characteristic curves'').

Importantly, Method 1 enabled calculation of traditional fit indices, which showed that the CRM fitted the observed scenario scores well according to RMSEA and SRMR but not CFI, with RMSEA = 0.041, 90% CI [0.039, 0.044], CFI = 0.879, SRMR = 0.035. We note that the single-factor linear model's fit to the scenario

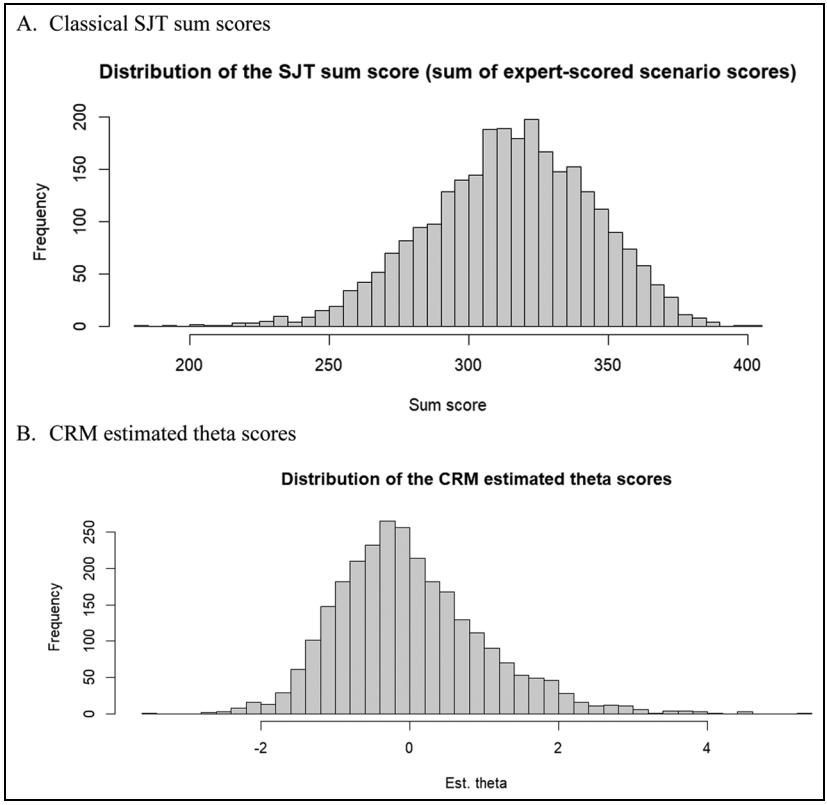


Figure 3. Distribution of the classical SJT sum scores and the CRM estimated theta scores in the empirical example. (A) Classical SJT sum scores. (B) CRM estimated theta scores.
Note. CRM = continuous response model; SJT = situational judgment test.

scores before logit transformation was notably worse (reported earlier in the section ''EFA of Scenario Scores''). This confirms advantages of the CRM with its adjustments around the score boundaries for these data.

CRM Person Parameter Estimation and Reliability. Person scores were estimated using EstCRMperson() function of package EstCRM, based on the established item parameters. Figure 3 presents the distribution of the classical SJT total score (sum of all scenario scores) and the theta score distribution. Figure 4 presents the relationship between the theta scores and the classical SJT sum scores for all examinees. Despite a very high correlation between the classical and IRT scores (r = .938), this relationship is not linear and resembles the logistic function assumed for its constituent scenarios. Finally, Figure 5 illustrates the expected SJT sum score conditional on the theta score, which is

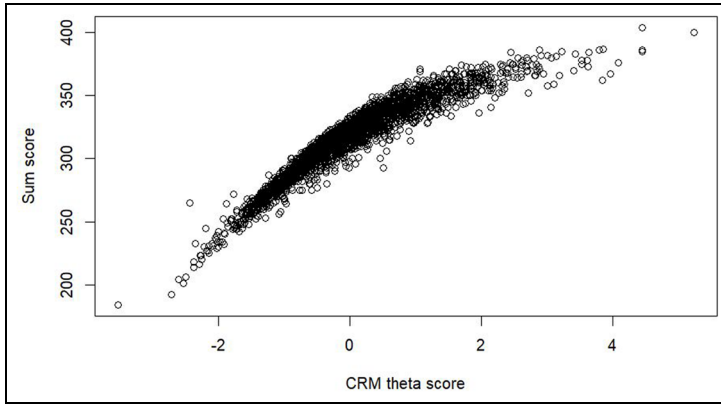


Figure 4. Observed SJT sum scores against CRM-estimated theta scores in the empirical example.
Note. CRM = continuous response model; SJT = situational judgment test.

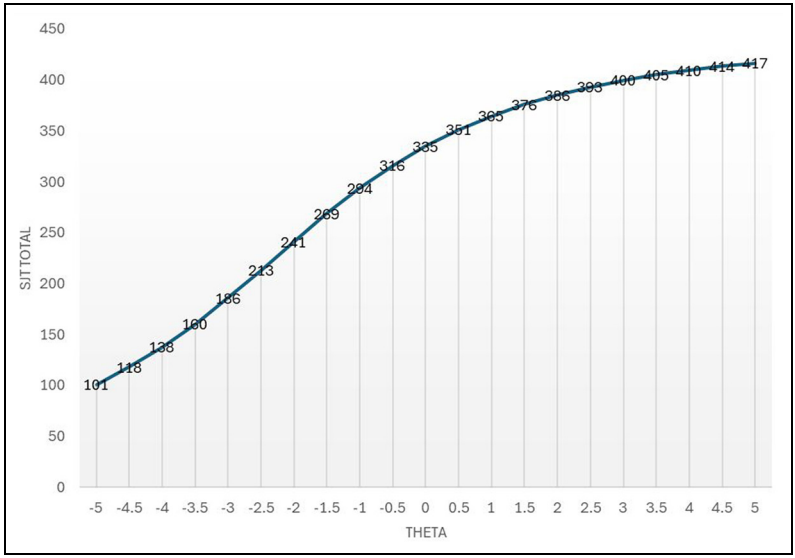


Figure 5. Expected total SJT score conditional on ability under CRM in the empirical example.
Note. CRM = continuous response model; SJT = situational judgment test.

the result of adding all expected scenario scores presented in Figure 1 (we can call this function ''test characteristic curve''). This function resembles the observed relationship between the theta scores and classical scores in Figure 4.

Since the examinees in this sample had no missing data, the standard error of measurement (SEM) was the same for all, SEM = 0.512. Theta score reliability computed using Equation 8 is then $\rho(\theta) = .793$.

Discussion

This article presented an approach to IRT modeling of SJT data with rating items nested within scenarios. We suggest using summated scenario scores as units of analysis and fitting Samejima's (1973) CRM to them. If the summated scenario scores have sufficient score ranges (e.g., 0–20) and the SJT in question is designed to measure a single capacity, then the CRM is theoretically suitable for such essentially continuous but bounded data.

This approach has multiple benefits. Perhaps the most important one is its simplicity and accessibility. Data preparation for this modeling is straightforward—it involves simply producing sums of expert-scored item scores within each scenario, making scenario scores. Fitting CRM is also straightforward, for example, using the R package EstCRM, alternatively fitting a CFA model to logit-transformed scenario scores as per Equation 3 in any general SEM package such as R's lavaan, or Mplus. Mplus also facilitates estimating and saving person theta scores in the same model run and allows computing additional a/b/alpha parameters using the ''MODEL CONSTRAINT:'' command. An example syntax is provided in Supplemental Material.

Once the item (scenario) parameters have been established, person ability scores may be computed using the closed expression in Equation 6. Then, the standard error of ability scores computed using Equation 7 and the test reliability can be estimated using Equation 8. Thus, instant scoring of test administrations is easy to implement outside of specialized software. This is in contrast with difficulties establishing construct validity, dimensionality and finding an appropriate model when working with items as units of analysis, resulting in questionable scoring foundations and biased reliability estimates.

These points are illustrated using an empirical example of SJT presenting professional dilemmas in a hospital setting. The dimensionality analysis revealed that the rating items in the bespoke SJT test measured three major factors, and there were five further factors, some of which related to local dependencies of items within scenarios. Theoretically suitable IRT models could be recommended for raw and expert-scored item responses; these were multidimensional variants of GPCM or GRMs, with scenario-based local dependencies modeled as uncorrelated method factors. However, if any of these suggested MIRT models were adopted for SJT item parameter and person parameter estimation, the biggest challenge would be their item-level multidimensionality. When measurement of one common capacity is intended, it is unclear how to interpret or operationalize three content-specific factors spanning all scenarios, as the items cannot be neatly separated into three unidimensional models, because there are 19 scenario-specific factors to account for. This model is

''inseparable'' and requires iterative optimization procedure to estimate person scores, which are conditional on the testlet factors. Operationally, this is hardly sustainable, particularly when running large testing projects with multiple test forms and complex linking designs. An alternative approach using scenarios as units of analysis proved more beneficial, as relatively simple unidimensional models from the continuous IRT family were found suitable. As illustrated in the data analysis example, it is easy to prepare the data by summing suitably scored item responses within scenarios, to check the goodness of fit of scenario-level data to CRM, to compute person scores and their standard errors, and to estimate reliability.

However suitable the CRM could prove in an empirical example, it was important to investigate its suitability more broadly to any rating items nested within scenarios, under varying degrees of multidimensionality, testlet structure, number of response categories, and other test design features. The simulation study, in which we manipulated six fully crossed design factors resulting in 64 conditions, showed that CRM is suitable in a wide range of applications, because its fit remains very good even on data with intended construct multidimensionality, as long as a single common dimension underlies all items, which is typically the case in SJTs.

The simulation study highlighted conditions most and least favorable to the ability of CRM to recover true score on that intended single dimension. As expected, the CRM will recover the latent trait very closely in applications with no intended construct multidimensionality, more and longer testlets, more response categories, and difficulty targeting appropriate to the tested population. But even in less favorable conditions, the CRM can be used due to its good fit to data. Although then its ability to recover the true score will be less impressive, it is no worse than for its counterpart summated test score.

The key benefit of the CRM over classical summated score or linear factor models is in its ability to differentiate at the test floor and ceiling, through the log-odds transformation of observed scenario scores, which compensates for the non-linearity of relationship between true and summated test score at the extremes shown in Figure 2 (second panel). This is particularly pronounced for scenario scores reaching lower and upper bounds, in which case they are corrected by adding or subtracting a small constant, respectively. Shojima (2003) suggested using 0.01 for this correction, which is implemented in the EstCRM package (Zopluoglu, 2012). Interestingly, in another article by Shojima (2005), a correction of 0.1was suggested, and the rationale for the choice for one or the other value is not clear. Since the magnitude of the correction impacts the degree of ''stretching'' the boundary scores, determining the optimum correction value would be an excellent topic for future research.

Implications for Assessment Practice

Many large-scale testing challenges can be addressed with the proposed approach. Simple modeling using CRM has the potential to overcome the challenges of designing SJT forms and maintaining SJT scenario/item banks. As with any model-based

measurement, the CRM model fit to empirical data can be checked, unidimensionality confirmed, and any poorly fitting scenarios can be removed or reconsidered. Revised scenarios can then be calibrated using a concurrent estimation of all forms with data missing by design, with some data overlapping between forms (these are usually anchoring scenarios) and some data specific to individual forms. This process will yield a calibrated bank of all trialed scenarios, populated with CRM parameters (a/b/alpha or alpha/beta/SD(omega) as desired) and expert scoring keys for each constituent item. In this bank, all scenarios are placed on the common scale by virtue of having CRM parameters estimated with the common basis (van der Linden & Barrett, 2016). For ongoing item/scenario calibration, newly developed items (i.e., pretest items) are calibrated concurrently with operational items by fixing the CRM parameters for the operational items (which are already known and stored in the bank) and freeing any constraint on the mean and SD of theta scores. With this, parameters for any newly developed items are on the common scale with the existing item bank—these are added to the item bank for future operational uses. For new form construction, the existing scenarios and scenarios newly added to the bank can be used, with their saved parameters. For scoring examinees, raw sum scores of each scenario included in that particular form are converted into CRM theta scores using the closed expression in Equation 6. This means no special software is needed for computing SJT theta scores; this is basically formula scoring that is established and implemented for each form, which is a linear combination of the raw scenario scores of examinees weighted by scenario parameters from the item bank. These theta scores are already on the common scale, and no equating is needed, neither across forms within the same testing year nor longitudinally across years. These theta scores can be then transformed to any other standard scores as desired for score reporting.

The process outlined above is of course only one of many possibilities. Other methods of equating SJT test scores across years and multiple test forms can be used in conjunction with the CRM, using both the common-person designs and the common-item (anchor) designs. Under the common-person designs, Shojima (2003) proposed a method for estimating the equating coefficients by integrating out common examinees' theta scores over the posterior distribution of theta. Under the common-item designs, the weighted least squares method (Haebara, 1980) and the test characteristic curve method (Stocking & Lord, 1983) can be applied using CRM-estimated item parameters and item response functions. These methods allow placing scores from different administrations or forms on a common scale while properly handling continuous item responses, including missing-by-design data. We look forward to innovations in this area.

ORCID iDs

Anna Brown https://orcid.org/0000-0002-5612-2627
Emma Morley https://orcid.org/0000-0001-6289-4490
Fiona Patterson https://orcid.org/0000-0002-1031-130X

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychology, University of Kent (No. 2026177083209210397) on February 11, 2026, with the need for written informed consent waived. The Chair of Ethics determined that this research involved minimal risk and waived the need for participant consent for the collection, analysis and publication of the retrospectively obtained and anonymized data used in the empirical illustration in this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Work Psychology Group designs the SJT which was used to illustrate this methodology.

Data Availability Statement

All data generation and analyses scripts for the simulation study are available on osf.io/dn42h. Data used in the empirical illustration cannot be shared due to test security reasons.

Supplemental Material

Supplemental material for this article is available online.

Note

1. The same correction is applied in EstCRM package (Zopluoglu, 2012), which we confirmed by examining the package code.

References

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

du Toit, M. (Ed.) (2003). *IRT from SSI*. SSI International.

Ellison, H. B., Grabowski, C. J., Schmude, M., Costa, J. B., Naemi, B., Schmidt, M., ... & Westervelt, M. (2024). Evaluating a situational judgment test for use in medical school admissions: two years of AAMC PREview exam administration data. *Academic Medicine*, *99*(*2*), 183–191. https://doi.org/10.1097/ACM.0000000000005548

Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research*, *37*(4), 521–542. https://doi.org/10.1207/S15327906MBR3704_05

Ferrando, P. J. (2010). Some statistics for assessing person-fit based on continuous-response models. *Applied Psychological Measurement*, *34*(4), 219–237. https://doi.org/10.1177/0146621609343288

González, J., & Wiberg, M. (2017). *Applying test equating methods using R*. http://www.springer.com/series/13206

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144–149. https://doi.org/10.4992/psycholres1954.22.144

Kepes, S., Keener, S. K., Lievens, F., & McDaniel, M. A. (2025). An integrative, systematic review of the situational judgment test literature. In *Journal of Management*, *51*(6), 2278–2319. https://doi.org/10.1177/01492063241288545

Lakens, D., & Bakker, M. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Muraki, E. (1992). *A generalized partial credit model: Application of an EM algorithm*. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1992.tb01436.x

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education*, *46*(9), 850–868. https://doi.org/10.1111/j.1365-2923.2012.04336.x

Patterson, F., Cousans, F., Edwards, H., Rosselli, A., Nicholson, S., & Wright, B. (2017). The predictive validity of a text-based situational judgment test in undergraduate medical and dental school admissions. *Academic Medicine*, *92*(9), 1250–1253. https://doi.org/10.1097/ACM.0000000000001630

R Core Team. (2023). *R: A language and environment for statistical computing*. https://www.r-project.org/

Samejima, F. (1970). Graded response model of the latent trait theory and tailored testing. *Psychometrika*, *35*(1), 5–17. https://doi.org/10.1007/BF02290599

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*(2), 203–219. https://doi.org/10.1007/BF02291114

Shojima, K. (2003). Linking tests under the continuous response model. *Behaviormetrika*, *30*(2), 155–171.

Shojima, K. (2005). A noniterative item parameter solution in each EM cycle of the continuous response model. *Educational Technology Research*, *28*, 11–22.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*(3), 506–532. https://doi.org/10.1177/1094428116630065

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/01466216 8300700208

van der Linden, W. J., & Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika*, *81*(3), 650–673. https://doi.org/10.1007/s11336-015-9469-6

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2026). dplyr: A grammar of data manipulation. *The R Foundation*. https://doi.org/10.32614/CRAN.package.dplyr.

Zopluoglu, C. (2012). EstCRM: An R package for Samejima's continuous IRT model. *Applied Psychological Measurement*, *36*(2), 149–150. https://doi.org/10.1177/0146621612436599

Zopluoglu, C. (2022). Package ''EstCRM.'' *The R Foundation*. https://doi.org/10.32614/ CRAN.package.EstCRM