Behavior Research Methods (2024) 56:8080–8090
https://doi.org/10.3758/s13428-024-02472-7

ORIGINAL MANUSCRIPT

Customizing Bayesian multivariate generalizability theory to mixed-format tests

Zhehan Jiang[1,2] · Jinying Ouyang[1,3] · Dingjing Shi[4] · Dexin Shi[5] · Jihong Zhang[6] · Lingling Xu[1,2] · Fen Cai[1,2]

Accepted: 27 June 2024 / Published online: 29 July 2024

Abstract

Mixed-format tests, which typically include dichotomous items and polytomously scored tasks, are employed to assess a wider range of knowledge and skills. Recent behavioral and educational studies have highlighted their practical importance and methodological developments, particularly within the context of multivariate generalizability theory. However, the diverse response types and complex designs of these tests pose significant analytical challenges when modeling data simultaneously. Current methods often struggle to yield reliable results, either due to the inappropriate treatment of different types of response data separately or the imposition of identical covariates across various response types. Moreover, there are few software packages or programs that offer customized solutions for modeling mixed-format tests, addressing these limitations. This tutorial provides a detailed example of using a Bayesian approach to model data collected from a mixed-format test, comprising multiple-choice questions and free-response tasks. The modeling was conducted using the Stan software within the R programming system, with Stan codes tailored to the structure of the test design, following the principles of multivariate generalizability theory. By further examining the effects of prior distributions in this example, this study demonstrates how the adaptability of Bayesian models to diverse test formats, coupled with their potential for nuanced analysis, can significantly advance the field of psychometric modeling.

Keywords Bayesian modeling · Generalizability theory · Mixed-format test · Assessment · Stan

Introduction

Mixed-format tests encompass assessments that combine various formats, such as multiple-choice questions (MCQs) and free-response tasks (FRTs), aiming to encompass a broader spectrum of knowledge and skills through diverse response types. In contrast to single-format tests, mixed-format tests offer a more comprehensive evaluation of students' abilities, mitigating potential bias that might stem from relying solely on one question type (Rodriguez, 2003). A notable example of mixed-format tests is the AP German Language exam, featuring 65 MCQs and 4 FRTs (Bischof, 2005), with responses assessed by raters. However, the analysis and interpretation of mixed-format test results can be intricate, given the necessity to address distinct designs and sources of error for each item format.

Unlike traditional univariate generalizability theory (G-theory; Cronbach, 1972; Brennan, 2001a) that decomposes test or measurement variance, multivariate generalizability theory (MG theory) serves as a statistical framework to account for multivariate response data (Shavelson & Webb, 1991). As MG theory applies to studies involving multiple subtests or subdomains, its outcomes include covariance and correlation data alongside variance estimates from the corresponding decomposition. MG theory empowers researchers to precisely model and analyze sources of error and variability tied to each facet (e.g., rater effect, task effect, and station

✉ Zhehan Jiang
jiangzhehan@bjmu.edu.com

✉ Jinying Ouyang
ouyangjinying@bjmu.edu.cn

1 Institute of Medical Education, Health Science Center, Peking University, Haidian District, 38 Xueyuan Rd, Beijing, China

2 National Center for Health Professions Education Development, Peking University, Beijing, China

3 School of Public Health, Peking University, Beijing, China

4 Department of Psychology, University of Oklahoma, Norman, OK, USA

5 College of Psychology, University of South Carolina, Columbia, SC, USA

6 College of Education and Health Professions, University of Arkansas, Fayetteville, AR, USA

effect), thereby shedding light on test reliability and validity. MG theory has been extensively applied across diverse studies, including instructional evaluations (Gillmore et al., 1978), teaching behavior assessment (Shavelson & Dempsey–Atwood, 1976), psychotherapy process ratings (Wasserman et al., 2009), clinical research (Lakes & Hoyt, 2009), and personality inventory studies (Arterberry et al., 2014).

Emerging from the mainstream use of MG theory, which primarily focuses on subtests/subdomains, Brennan et al. (2022) have adapted this method to investigate mixed-format tests (including AP German Language exam used in the latter section), which are considered intricate design structures (Sinharay, 2015). When considering the same population responding to both formats, this configuration is represented as the notation $\{p^\bullet \times i\}\{p^\bullet \times (r : i)\}$ in the language of G theory. In this context, p represents the person (random) effect, i stands for the item/task (random) effect, and r denotes the rater (random) effect. The notation "facet1 × facet2" is used to denote crossed facets, while "facet1 : facet2" is used to denote that facet1 is nested within facet2. The superscript bullet symbolizes that an identical group of individuals responds to both item types, under the assumption that each rater potentially rates different items in the second format. The utilization of two sets of braces indicates two levels (MCQs and FRTs) within the fixed facet. In MG theory notation, this same design is denoted as $p^\bullet \times [i^\circ \cup (r^\circ : i^\circ)]$, where ∪ represents the union of the two fixed facets, namely MCQs and FRTs. The open circle superscript signifies that measurement conditions differ between these two formats. The MG theory notation can be visually explained through Venn diagrams depicted in Fig. 1. In G theory, Venn diagrams are used to visually represent the relationships between different facets (sources of variability) in a measurement design. As each facet is represented by a big circle, the relationships between facets (crossed, nested, or independent) are depicted using overlapping, concentric, or separate circles, respectively. Crossed facets are represented by overlapping circles, indicating that each level of one facet can be combined with each level of the other facet. Nested facets are represented by concentric circles, with the nested facet positioned inside the circle representing the facet in which it is nested. Independent (disconnected) facets are represented by separate, non-overlapping circles, indicating that the levels of one facet are completely unrelated to the levels of the other facet. MG theory's diagrams also include dots within the circles to represent the variables being measured. Filled dots indicate that a variable is present within a particular facet or combination of facets. For example, if a filled dot is placed within the overlapping region of two crossed facets, it means that the variable is measured for all combinations of levels of those two facets. Blank dots, on the other hand, indicate that a variable is absent within a particular facet or combination of facets. If a blank dot is placed within a circle representing a facet, it means that the variable is not measured for any levels of that facet. The placement of filled and blank dots within the Venn diagram helps researchers understand which variables are measured within each facet and how the variables are related to the facets. If the two formats were modeled individually instead of being treated as a whole, like Fig. 1 presents, one can translate the MG theory design into its univariate version: $p \times i$ and $p \times (r : i)$ for MCQs and FRTs, respectively.

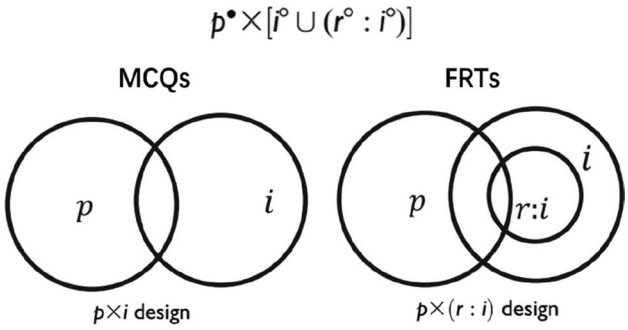

Fig. 1 The Venn diagram of multivariate generalizability theory for a mixed-format test. *Note*: Overlap indicate crossing design facets and concentric circles indicate nested facets

From a statistical perspective, facets within (M)G theory are regarded as random effects, while (traditional) intercepts/means and slopes are treated as fixed effects. When both effect types are present, they suggest linear mixed-effect modeling (LMM). MG theory can be understood as LMM. Using the FRT's univariate G-theory $p \times (r : i)$ design for example, a response/score of a person rated by a rater on a task/item can be decomposed into $x_{p(r:i)} = \mu + \theta_p + \theta_i + \theta_{r:i} + \theta_{pi} + e$, while e is the residual that can be also termed as $\theta_{p(r:i)}$ and μ is the intercept/mean. Each θ is assumed to follow a normal distribution with a zero mean and a corresponding variance (e.g., $\theta_p \sim N(0, \sigma_p^2)$). The multivariate component in MG theory, on the other hand, requires some $\mathbf{\Theta} \sim MVN(0, \Sigma)$ while the subscript should be added to indicate the facet. These θ s and $\mathbf{\Theta}$ s are the random effects and the intercepts/means are the fixed effect of LMM. LMM estimation can be conducted through established methods like maximum likelihood (ML), restricted maximum likelihood (REML), and Bayesian estimation (Jiang & Skorupuski, 2018). Alternatively, a well-known method called expected means squares (EMS) remains prevalent, particularly in studies utilizing the mGENOVA software (Brennan, 2001b) for MG theory analysis.

However, these estimations are predicated on the assumption that responses in G theory follow a normal distribution, which limits the practical application of (M)G theory. Many response data, particularly in mixed-format tests, can be binary or ordinal, rendering this assumption less suitable. In such cases, if responses do not adhere to a normal

distribution but the preferred linear-like model is still desired for (M)G theory analyses, a link function is necessary to transform a non-linear relationship into a linear one. Common link functions and their descriptions are outlined in Table 1. It is important to note that LMM with link functions is referred to as generalized LMM (GLMM). GLMM, like its name suggests, is highly related to generalized linear models (GLMs), which are a flexible class of statistical models that extend the concept of linear regression to accommodate response variables with various distributions, such as binary, count, or categorical data. GLMs allow for the modeling of the expected value of the response variable as a linear combination of predictor variables, linked through a specified link function. The choice of the link function depends on the distribution of the response variable. For example, logistic regression, a common type of GLM, uses the logit link function to model binary responses. The general formula for a GLM can be expressed as:

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

where $g(\cdot)$ is the link function, E(Y) is the expected value of the response variable Y, β_0 is the intercept, β_i are the regression coefficients, and X_i are the predictor variables. GLMMs extend GLMs by incorporating random effects in addition to fixed effects. Random effects are used to account for clustered, nested, or repeated measures data, where observations within a group are correlated. GLMMs can handle various types of response variables, similar to GLMs, but they also model the variability between groups or clusters. The general formula for a GLMM can be expressed as:

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + b_1 Z_1 + b_2 Z_2 + ... + b_k Z_k$$

where $g(\cdot)$, E(Y), β_0, β_i, and X_i are the same as in the GLM formula, and b_j are the random effects coefficients associated with the random effects variables Z_j. GLMs and GLMMs are widely used in various fields, such as social sciences, ecology, and medical research, when the response variable is not normally distributed, and the relationships between the predictors and the response are not necessarily linear. These models provide a powerful and flexible framework for analyzing categorical, count, and other types of data while accounting for the specific characteristics of the response variable and the structure of the data. For a comprehensive treatment of GLMs, GLMMs, and their applications to categorical and ordinal data analysis, readers are encouraged to refer to the textbook by Agresti (2012).

Browne et al. (2001) highlighted the potential pitfalls if link functions are not accurately specified in LMM. Additionally, Broatch et al. (2018) demonstrated that jointly modeling different data types, as seen in mixed-format tests, results in significantly improved median log-loss and absolute residuals of cross-validation predictions when correlated pairs of estimated random effects are employed. From both statistical and psychometrical perspectives, benefits of the joint estimation method are multifold (DeCarlo, 2024; Lee et al., 2020; Wei et al., 2023):

(1) Increased statistical power: By combining data from multiple sources or response formats, joint estimation methods can increase the overall sample size, leading to more precise and reliable estimates.
(2) Accounting for shared variance: Joint estimation methods can account for the shared variance between different response formats, which may lead to more accurate estimates of the underlying constructs being measured.
(3) Improved model fit: Joint estimation methods may provide a better fit to the data by incorporating information from multiple sources, leading to more valid and reliable results.
(4) Reduced bias: By using information from multiple response formats, joint estimation methods may help reduce bias that could arise from relying on a single format.
(5) Enriched validity evidence: Construct validity: High correlations between the estimates derived from different formats or sources can provide evidence of convergent validity, strengthening the overall validity of the measurements, while if the correlations between estimates of distinct constructs are low, it provides evidence of discriminant validity, indicating that the measurements are indeed assessing separate constructs.

Table 1 Common link functions and their uses

Link name	Data type	Distribution	Function
Logit	Integers: {0, 1}	Bernoulli	XB = ln(μ/(1–μ))
Logit	integers: {0, 1, …, N}	Binomial	XB = ln(μ/(n–μ))
Identity	real: (-∞, ∞)	Normal Distribution	XB = μ
Negative inverse	real: (0, ∞)	Exponential Distribution	XB = -μ-1
Negative inverse	real: (0, ∞)	Gamma Distribution	XB = -μ-1
Log	integers: 0, 1, 2, …	Poisson Distribution	XB = ln(μ)
Probit	Integers: {0, 1}	Bernoulli	XB = Φ^{-1}(μ)

Thus, the need for an easily implementable estimation method that facilitates incorporating link functions for modeling mixed-format tests within the MG theory framework becomes apparent.

In addition to examining the statistical attributes of response data, this research draws inspiration from the growing trend of value-added models (VAM; McCaffrey et al., 2003). In the realm of education, VAM advances existing theories by estimating the relationships between diverse teacher inputs and various student achievements, while also allowing for the assessment of tangible real-world outcomes like graduation. The exploration of simultaneously incorporating continuous and binary outcomes has been comprehensive, encompassing both methodological and empirical considerations (Park & Beretvas, 2020).

Similar to G theory, VAM employs linear mixed-effects modeling (LLM) to analyze data, considering teacher and/or school effects as random effects (Raudenbush, 2004). However, while the VAM literature offers plausible strategies for handling different data types in LMM, most available options lack the capacity to accommodate crucial elements of MG theory (Broatch et al., 2018). These elements include (1) cross-classified structures that involve more than three random effects with interactions, and (2) distinct components for different formats, such as the item effect for modeling MCQs and both item and rater effects for modeling FRTs.

Building on the foundations of applying MG theory to mixed-format tests (Brennan et al., 2022), this study methodologically expands modeling to incorporate various response data types (formats). This is achieved through a convenient realization in the R software within a Bayesian framework. Specifically, the study presents a solution for multivariate cross-classified mixed-effect modeling when both continuous and discrete data coexist. Ultimately, this solution can be applied to the analysis of mixed-format tests within the MG theory framework.

Method

Bayesian estimation is chosen over traditional estimations (i.e., frequentist methods) for two primary reasons: the intrinsic advantages of Bayesian estimation and the capabilities of the brms package (Bürkner, 2017) in the R software. A prominent feature of Bayesian modeling is its ability to incorporate prior information and beliefs about parameters into the analysis. This is particularly beneficial when there is existing knowledge or expert opinions, enhancing the interpretability and reliability of estimates (Marsman & Wagenmakers, 2017). Additionally, Bayesian estimation produces posterior distributions for model parameters, offering a transparent representation of the uncertainty linked with parameter estimates. This aids in comprehending the variability in parameter values and their implications (Wagenmakers et al., 2018). Bayesian methods are also well suited for small sample sizes, effectively utilizing available information and integrating prior knowledge (Smid et al., 2020). Moreover, Bayesian estimation adeptly handles missing data by treating missing values as parameters to be estimated, leading to more comprehensive use of available information (Enders, 2022). Notably, Bayesian methods offer flexibility in model specification, enabling the incorporation of various data distributions and intricate relationships among variables (Austerweil et al., 2015).

From a software standpoint, the brms package (Bürkner, 2017) offers a streamlined approach to estimate Bayesian LLM by leveraging Stan, a widely used probabilistic programming language for Bayesian estimation (Gelman et al., 2015). This versatile package supports a wide range of distributions and link functions, enabling users to model diverse scenarios, including binomial, Poisson, survival, response times, ordinal, quantile, zero-inflated, hurdle, and even nonlinear models within the GLMM framework. Notably, the brms package (Bürkner, 2017) facilitates the creation of Stan codes with a syntax familiar to users of the lme4 package (Bates et al., 2015), a dominant tool for LMM estimation in the R community. While the brms package does not inherently provide solutions for jointly modeling different data types as multivariate dependent variables, its feature for generating complete Stan codes simplifies the process of customization and adaptation. In addition, it is worth noting that the cumulative family is available for ordinal outcomes in the brms package, while unavailability for frequentist (ML) estimators limits other counterpart software packages to one random intercept only.

We obtained samples from a testing site of the Standardized Competence Test for Clinical Medicine Undergraduates (SCTCMU) to demonstrate the proposed MG theory modeling for mixed-format tests. The SCTCMU consisted of 300-item MCQs with the well-known 1/0 scoring and six-tasks/station performance assessment with 12 raters (i.e., each task was presented in a testing station, while a pair of raters were assigned to grade the performance of each person on a task). The scaled score made of a checklist for each task ranged from 0 to 16. In total, 533 persons were included into the study; their responses were the data for further modeling and estimation. Note that the random effect components of the SCTCMU were similar to the AP German Language exam as the same design was adopted. Therefore, the notations and the schema used in the AP German Language exam example can be adhered to the present demonstration.

In this context, the MCQs are treated as binary responses, while the FRTs are regarded as continuous variables. Specifically, the $p^{\bullet} \times [i^{\circ} \cup (r^{\circ} : i^{\circ})]$ design can be decomposed to five covariance matrices accordingly: the left panel of

Fig. 2 shows the specifications where *X* implies existing estimates and *NA* indicates missing for not applicable to those measurements. It can be seen that Σ_p is the only dense matrix where others are sparse matrices. Σ_i is a diagonal matrix because the items of the MCQs and the tasks in the FRTs are both labelled as *i* and they are independent to each other. $\Sigma_{r:i}$ implies that the rater effect, denoted as *r*, is nested within tasks in the FRTs; it contains *NA* as this effect is unique to the FRTs only, of which the logic applies to $\Sigma_{p(r:i)}$. Note that the first element of Σ_{pi} setting to *NA* credits to the assumption that MCQs follow a binary distribution. Therefore, deploying the logit/probit link function becomes necessary according to the assumption, while no residual effect is calculable due to the statistical identifiability issue, because both the expected value and variance of a binary variable are a function of a single parameter (the probability of success). However, if MCQs are considered to be normally distributed, the first element of Σ_{pi} should be marked as X instead of NA. In this section, the right panel of Fig. 2 shows samples of the data, of which the expression is in a long type (i.e., each row only contains one response value). If it is an MCQ, the rater column is *NA* because these items are not measured by different raters and the response is either 1 or 0. On the other hand, when it comes to an FRT, one can also track the information about raters and tasks (i.e., items). Note that the nested structure of $r^\circ : i^\circ$ is also apparent from the example data in Fig. 2, as different sets of raters assess each item.

Based on the setting outlined in Fig. 2, we split the complete data into two subsets, SCTCMU_MCQ and SCTCMU_FRT for analytical purposes and deployed the brms package to model the two formats separately, of which the corresponding lines below were executed in the R software.

```
fit1_bayes <- brm(MCQ ~ (1| Person) + (1| Item), data = SCTCMU_MCQ, family = bernoulli)
fit2_bayes <- brm(FRT ~ (1| Person) + (1| Item) + (1| Rater) + (1| Person : Item), data = SCTCMU_FRT, family = gaussian)
```

Note that other configurations were left default for the estimation (e.g., iteration number and priors for parameters); we refer readers to Bayesian literature (e.g., van de Schoot et al., 2021), the Stan manual (Carpenter et al., 2017), and/or any other materials related to practical Bayesian estimation for the configuring details. The parameter estimates could be investigated from fit1_bayes and fit2_bayes, the R objects containing the outcomes of the separate modeling practice. Correspondingly, the frequentist methods can be achieved by executing the following lines. The results can be substantially different from the ones obtained via Bayesian estimation.

```
fit1_frequentist <- glmer(MCQ ~ (1| Person) + (1| Item), data = SCTCMU_MCQ, family = binomial)
fit2_ frequentist <- lmer(FRT ~ (1| Person) + (1| Item) + (1| Rater) + (1| Person : Item), data = SCTCMU_FRT)
```

Although the separate models can provide useful descriptions about the test, modeling responses from both formats jointly is the pursue of this paper. The steps outlined below are optional because one can directly write codes from scratch. However, they surely deliver a faster way to construct the models without making inadvertent errors

$$\Sigma_p = \begin{bmatrix} X & X \\ X & X \end{bmatrix} \quad \Sigma_i = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} \quad \Sigma_{r:i} = \begin{bmatrix} NA & 0 \\ 0 & X \end{bmatrix} \quad \Sigma_{pi} = \begin{bmatrix} NA & 0 \\ 0 & X \end{bmatrix} \quad \Sigma_{pr:i} = \begin{bmatrix} NA & 0 \\ 0 & X \end{bmatrix}$$

Person	Item	Rater	Format	Response
1	1	NA	MCQ	1
1	2	NA	MCQ	1
…	…	…	…	…
1	1	1	FRT	3
1	1	2	FRT	2
1	2	3	FRT	4
1	2	4	FRT	4
…	…	…	…	…
2	1	NA	MCQ	0
2	2	NA	MCQ	1

Fig. 2 Covariance components and sample data of SCTCMU and AP German Language exam. *Note*: $\Sigma = \begin{bmatrix} \sigma^2_{MCQ} & \sigma_{MCQ,FRT} \\ \sigma_{MCQ,FRT} & \sigma^2_{FRT} \end{bmatrix}$; the rater column and first element of $\Sigma_{r:i}$ and $\Sigma_{pr:i}$ is *NA* because MCQs are not measured by different raters; the first element of Σ_{pi} is *NA* because MCQs follow a binary distribution

in the model specification. To begin with, one can use the make_stancode and the make_standata functions from the brms package to generate codes to accommodate the specific requirement for the Stan modeling and Stan data lists. In terms of the modeling, one should (1) rename, (2) merge, and (3) revise the codes produced from the make_stancode and the make_standata functions. Renaming can help distinguish variables from the models. For instance, when it comes to the number of the item effect's levels, by default, Z_2_1 is used by the brms package for both models, but Z_2_1 of MCQs is almost always different from that of FRTs. Therefore, they need to be renamed as Z_2_1 _MCQ and Z_2_1 _FRT to ensure correct representations of the corresponding variables as following:

Original code： vector[N] Z_2_1;
Code after renaming： vector[N_**FRT**] Z_2_1_**FRT**; vector[N_**MCQ**] Z_2_1_**MCQ**;

Merging after the renaming from the two models requires allocating codes to correct Stan chunks: "data", "transformed data", "parameters", "transformed parameters", and "model". This step is fairly straightforward and paves the foundation for the following step. Revising implies the changes needed for jointly modeling, primarily reflecting on "parameters", "transformed parameters" and "model" chunks. Specifically, in addressing the person effect's covariance matrix into the model, a "cov_matrix" should be specified in "parameters" chunk, its random effect for each person (i.e., bivariate latent variables Θ_p) should be specified in "transformed parameters" chunk, and "model" chunk should include the relation that Θ_p follows a multivariate normal distribution with a mean vector of zeros and the "cov_matrix" covariance. The complete codes with an analysis for a toy dataset are shown in the Supplementary Material; as one can see, no priors were set for the parameter estimates.

Analysis

Joint estimates of the covariance components for FRT and MCQ data from the SCTCMU were obtained using the previously mentioned method. Reporting guidelines such as ROBUST (Reporting Of Bayes Used in clinical STudies) and BARG (Bayesian analysis reporting guidelines) were recommended in practice. Following such guidelines would help researchers to communicate their Bayesian analyses more clearly and consistently (Sung et al., 2005; Kruschke, 2021). The brms package was used to generate the initial stan code and appropriately formatted data for stan. This code and data were subsequently renamed, merged, and revised, resulting in an amalgamated code and data (available at https://osf.io/wud3x). The rstan package was employed for analysis with four chains, 6000 iterations for warmup and a total of 9000 iterations.

To enhance the robustness of the Bayesian estimator's convergence, data-dependent priors (similar to empirical Bayes priors) were utilized. We commenced by pre-estimating the mixed-format data using stan's default flat priors (Serang et al., 2015; Shi & Tong, 2017). The preliminary estimates of the covariance components are presented in Table 2. The estimated parameter means μ_{np} and standard deviations sd_{np} were then employed in the prior distributions in the formal estimation process, as depicted in Table 3. A normal distribution, with means and standard deviations matching those from the initial estimation, was chosen as the

Table 2 Covariance components of pre-estimation with flat priors

Covariance Component	M	SD	95% CI
Σ_p	$\begin{bmatrix} 0.686 & 0.112 \\ 0.112 & 0.522 \end{bmatrix}$	$\begin{bmatrix} 0.036 & 0.028 \\ 0.028 & 0.044 \end{bmatrix}$	$\begin{bmatrix} [0.606, 0.775] & [0.057, 0.167] \\ [0.057, 0.167] & [0.456, 0.595] \end{bmatrix}$
Σ_i	$\begin{bmatrix} 0.175 & 0 \\ 0 & 0.077 \end{bmatrix}$	$\begin{bmatrix} 0.015 & NA \\ NA & 0.061 \end{bmatrix}$	$\begin{bmatrix} [0.147, 0.206] & NA \\ NA & [0.021, 0.234] \end{bmatrix}$
$\Sigma_{r:i}$	$\begin{bmatrix} NA & 0 \\ 0 & 0.034 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & 0.011 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & [0.017, 0.057] \end{bmatrix}$
Σ_{pi}	$\begin{bmatrix} NA & 0 \\ 0 & 0.188 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & 0.701 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & [0.002, 1.111] \end{bmatrix}$
$\Sigma_{p(r:i)}$	$\begin{bmatrix} NA & 0 \\ 0 & 0.588 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & 0.014 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & [0.560, 0.615] \end{bmatrix}$

$\Sigma = \begin{bmatrix} \sigma^2_{MCQ} & \sigma_{MCQ,FRT} \\ \sigma_{MCQ,FRT} & \sigma^2_{FRT} \end{bmatrix}$; the first element of $\Sigma_{r:i}$ and $\Sigma_{p(r:i)}$ is *NA* because MCQs are not measured by different raters; the first element of Σ_{pi} is *NA* because MCQs follow a binary distribution

Table 3 Covariance components of pre-estimation with flat priors

Parameter	Prior distribution	M	SD
$Intercept_{FRT}$	Normal (11.949, 0.216)	11.949	0.216
$Intercept_{MCQ}$	Normal (0.016, 0.046)	0.016	0.046
Σ_p	Inverse-Wishart (425.500, $\begin{bmatrix} 220.545 & 47.320 \\ 47.320 & 289.84 \end{bmatrix}$)	$\begin{bmatrix} 0.522 & 0.112 \\ 0.112 & 0.686 \end{bmatrix}$	$\begin{bmatrix} 0.036 & 0.030 \\ 0.030 & 0.047 \end{bmatrix}$
$\sigma^2_{i,FRT}$	Inverse-gamma (3.593,0.200)	0.077	0.061
$\sigma^2_{i,MCQ}$	Inverse-gamma (138.111,23.994)	0.175	0.015
$\sigma^2_{r:i,FRT}$	Inverse-gamma (11.554,0.359)	0.034	0.011
$\sigma^2_{pi,FRT}$	Inverse-gamma (2.072,0.202)	0.188	0.701
$\sigma^2_{p(r:i),FRT}$	Inverse-gamma (1766,1038)	0.588	0.014

$\Sigma = \begin{bmatrix} \sigma^2_{MCQ} & \sigma_{MCQ,FRT} \\ \sigma_{MCQ,FRT} & \sigma^2_{FRT} \end{bmatrix}$.

prior for the intercepts. The inverse-Wishart distribution was used for covariance components for persons with degrees of freedom d and 2 × 2 matrix S. The inverse-Wishart distribution, characterized by degrees of freedom (d) and a 2 × 2 scale matrix ($\boldsymbol{S}$), was used for the covariance components of individuals. The hyperparameters for the inverse-Wishart distribution were computed using the formula described by Muthén and Asparouhov (2012), ensuring consistency of the marginal means with those from the noninformative Bayesian estimation. The marginal standard deviation for the first element of the covariance matrix was also retained. For variance components of FRT items, FRT raters, FRT person-item interactions, FRT residuals and MCQ items, an inverse-gamma distribution was utilized with hyperparameters α and β, setting to align the means and standard deviations with those from noninformative Bayesian estimation. Mean, standard deviation (SD) and 95% central credible interval (CI) of variance components were reported in following.

The results of the G study are exhibited in Table 4, shown as 2 × 2 covariance matrices for each facet with rows and columns ordered as MCQ and FRT. Examples of MCMC plot are shown in Fig. 3. Variance components for persons in MCQ and FRT were 0.683 (SD = 0.025, 95% CI = [0.623, 0.747]) and 0.509 (SD = 0.032, 95% CI = [0.462, 0.560]) separately, suggesting that variability among persons explained the total variability of MCQ and FRT scores largely. The covariance between MCQ and FRT is 0.111 (SD = 0.020, 95% CI = [0.073, 0.152]), showing a substantial linked or crossed facet. Variance components for items in MCQ and FRT were 0.369 (SD = 0.013, 95% CI = [0.344, 0.397]) and 0.202 (SD = 0.047, 95% CI = [0.134, 0.315]), which suggests there is some variability among rater scores. For FRT, $\sigma^2_{r:i} = 0.088$ (SD = 0.005, 95% CI = [0.078, 0.099]), suggesting there is some variability in rater scores for any item; $\sigma^2_{pi} = 0.268$ (SD = 0.084, 95% CI = [0.160, 0.483]), suggesting there is some variability among persons with respect to their

Table 4 Results of covariance components of SCTCMU data

Covariance Component	M	SD	95% CI
Σ_p	$\begin{bmatrix} 0.683 & 0.111 \\ 0.111 & 0.509 \end{bmatrix}$	$\begin{bmatrix} 0.025 & 0.020 \\ 0.020 & 0.032 \end{bmatrix}$	$\begin{bmatrix} [0.623, 0.747] & [0.073, 0.152] \\ [0.073, 0.152] & [0.462, 0.560] \end{bmatrix}$
Σ_i	$\begin{bmatrix} 0.369 & 0 \\ 0 & 0.202 \end{bmatrix}$	$\begin{bmatrix} 0.013 & NA \\ NA & 0.047 \end{bmatrix}$	$\begin{bmatrix} [0.344, 0.397] & NA \\ NA & [0.134, 0.315] \end{bmatrix}$
$\Sigma_{r:i}$	$\begin{bmatrix} NA & 0 \\ 0 & 0.088 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & 0.005 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & [0.078, 0.099] \end{bmatrix}$
Σ_{pi}	$\begin{bmatrix} NA & 0 \\ 0 & 0.268 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & 0.084 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & [0.160, 0.483] \end{bmatrix}$
$\Sigma_{p(r:i)}$	$\begin{bmatrix} NA & 0 \\ 0 & 0.712 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & 0.007 \end{bmatrix}$	$\begin{bmatrix} NA & NA \\ NA & [0.699, 0.726] \end{bmatrix}$

$\Sigma = \begin{bmatrix} \sigma^2_{MCQ} & \sigma_{MCQ,FRT} \\ \sigma_{MCQ,FRT} & \sigma^2_{FRT} \end{bmatrix}$; the first element of $\Sigma_{r:i}$ and $\Sigma_{p(r:i)}$ is *NA* because MCQs are not measured by different raters; the first element of Σ_{pi} is *NA* because MCQs follow a binary distribution

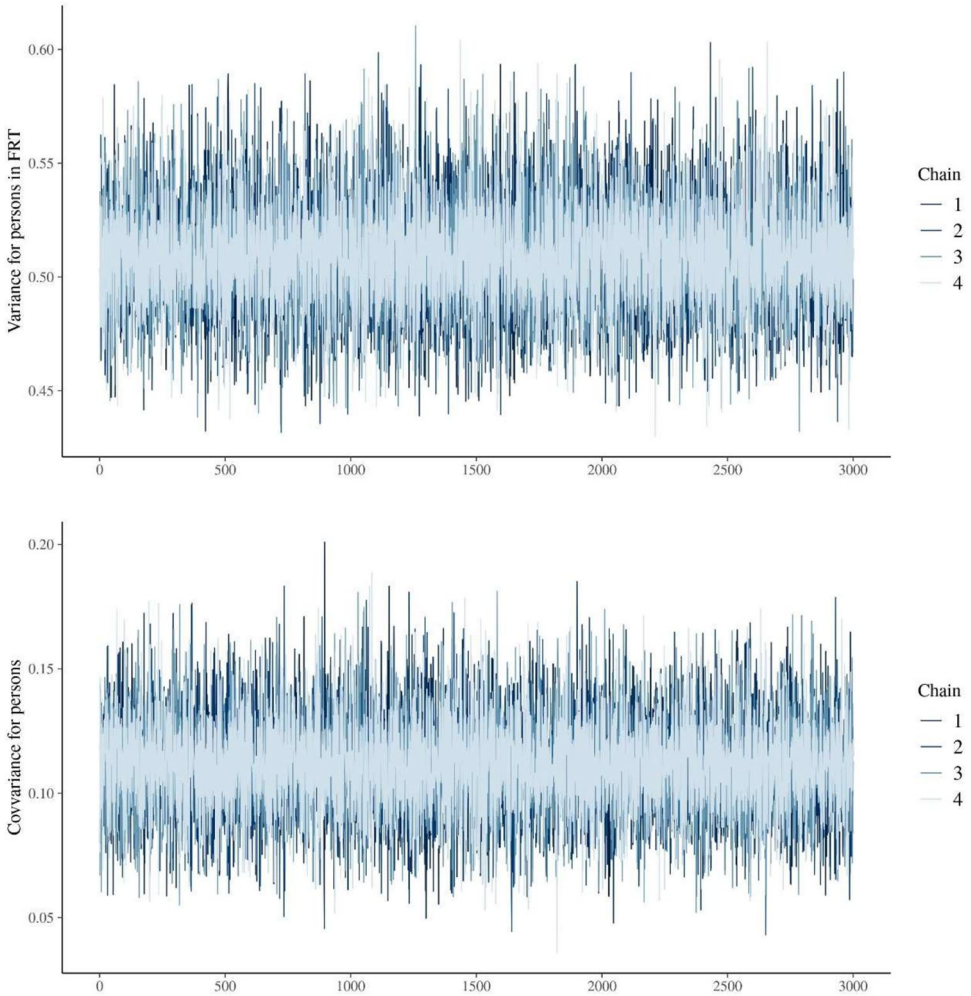

Fig. 3 MCMC plots of main parameters

scores of items; $\sigma^2_{p(r:i)} = 0.712$ (SD = 0.007, 95% CI = [0.699, 0.726]), suggesting there is some variability in residual. The R-hat values for all parameters were below 1.01, indicating satisfactory convergence. The relative and absolute error variances for FRTs were 0.0322 and 0.0637, respectively, while the generalizability and phi coefficients were 0.941 and 0.889. For MCQs, the absolute error variance was 0.0012, with a phi coefficient of 0.998, indicated a highly reliable composite even with relatively noisy measurements.

Discussion and conclusion

With appropriate modeling and estimation for a mixed-format tests, the analysis can provide many benefits. The first advantage is that understanding the correlation between target constructs allows educators and evaluators to gain a more nuanced view of a student's overall competencies and learning needs. In addition, the knowledge of the relationship between these constructs aids in designing tests that are both valid and reliable: If there is a strong correlation, it might suggest that both constructs are measuring similar aspects of knowledge or skill, while conversely, a weak correlation could indicate that they are assessing distinct domains. Similarly, it provides deeper insights into the association between target constructs that can inform teaching methods and curriculum design: stronger correlation suggests that improvements in one area might lead to improvements in the other. Following the same vein, more accurate estimates of the model are vital for providing tailored feedback and support to students (e.g., if a student excels in MCQs but struggles with writing tasks, this disparity can highlight specific areas where the student needs further support). Finally, the estimates are critical for setting appropriate standards and making informed decisions about educational policies,

particularly those related to tests and accountability. In the analysis, Σ_p could be further converted to a correlation matrix for a more straightforward interpretation: the correlation between MCQs and performance assessment was near 0.2. Ideally, one would expect the correlation to be moderate to reflect theoretical association as well as distinction of the corresponding constructs according to subscore literature: the range (after disattenuating via a psychometric model) from 0.3 to 0.8 makes a meaningful correlation for a test with subdomains (Feinberg & Jurich, 2017; Haberman & Sinharay, 2010; Raymond & Jiang; 2020). Apparently, 0.2 does not belong to the ideal range: the weak value may imply further revamping on this test.

When Bayesian methods are employed, the estimation is advantageous in that it enables the incorporation of prior information into the modeling process. Bayesian analysis incorporates prior knowledge or beliefs into estimation and updates the posterior estimation of the model parameters of interests. In medical education research, knowledge about medical tests is dynamic in that the tests are assessed over time among different test-takers. Test takers' knowledge in specific domains as well as test administrators' insights on the test takers both evolve over time. This study uses Bayesian methods to incorporate the prior knowledge about these information pieces. Specifically, we used inverse-Wishart distribution as the prior for the variance-covariance parameter matrix (e.g., Barnard et al., 2000). As a conjugate prior for the variance-covariance matrix under normal likelihood (e.g., Gelman et al., 2013), inverse-Wishart distribution is used to lead to closed-form posterior distributions of the parameter estimates of the relationships between MCQs and performance assessment. Furthermore, to improve estimation performance, we chose the hyperpriors by first using noninformative priors for the variance-covariance matrix. We then used the estimated parameters as a type of informative prior and re-analyzed the model to obtain the final estimates for the variance-covariance matrix. Our study is advantageous in at least two ways. First, using the conjugate inverse-Wishart distribution for the variance-covariance matrix for the relationship between MCQs and performance assessment, we obtain an improved convergence rate. Second, in the choice of hyperpriors, we used a concept similar to data-dependent priors (DDP; Serang et al., 2015), however, the key distinction lies in how the data-dependent estimates were obtained. While DDP used the frequentist method for obtaining the sample-based estimates, our study uses Bayesian techniques for the estimation, making our study a fully Bayesian approach for the analysis. Recent research proceedings suggest that using the data dependent priors improve estimation performance in mixed-effects models under nested data structure (e.g., McNeish, 2016; Shi & Tong, 2017); therefore, we believe the analytic approach in the current study is advantageous in improved estimation performance.

The modeling should align with the design, which turns out to be the determinant of the statistical and programming complexity. In the analysis, matrices such as Σ_i and $\Sigma_{r:i}$ were simple. Yet in designs with both cross-classified and higher-order nested structures, the modeling decompositions and covariance constructions could be highly intricated. One needs to be able to discern what components should be included. For instance, the analysis implied that no additional Σ_r should be specified because it was subsumed in $\Sigma_{r:i}$, else the redundant part may cause wrongful estimates and/or convergence failure.

Combining binary and continuous response variables, as demonstrated in this study, is a common practice in the realm of educational assessments and psychometrics. However, the generalization of our methodology to encompass more than two dimensions, alongside the integration of diverse link functions, offers substantial practical benefits, especially in the context of more complex testing scenarios. By extending the model to include multiple dimensions, one can capture a broader range of skills and abilities, reflecting the multifaceted nature of educational achievement in alignment with the theoretical framework. This multi-dimensional approach allows for a more nuanced understanding of student performance, acknowledging that proficiency is not a monolithic construct but rather a composite of various interrelated competencies. Furthermore, the adoption of diverse link functions can pave the way for a more flexible and accurate representation of the relationships between different types of assessment data: link functions can be tailored to specific types of response data, thereby enhancing the model's ability to accurately reflect the underlying processes of educational assessments. For instance, a cumulative logistic link function might be ideal for binary data, while a cumulative probit or complementary log-log link could be more suitable for ordinal values, which are not rare in G theory modeling (Ark, 2015; van der Ark et al., 2023; Vispoel et al., 2019).

Finally, in a typical MG theory study, composite reliability indexes are often used to indicate the psychometric quality of a test with multidimensional constructs. There are variations of the composite reliability indexes for different purposes, for instance, ***G*** index recently proposed by Jiang and Raymond (2018) can be used to evaluate the appropriateness of subscore reporting. There is no known index, however, to serve the similar function for mixed-format tests via MG theory. Therefore, devising indexes to accommodate the specific design is necessary for easier decision-making process, despite that a systematic investigation of the indexes as well as suggested cut-off values are needed prior to the actual application.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-024-02472-7.

Funding This work was supported by the National Natural Science Foundation of China for Young Scholars under Grant 72104006; Peking University Health Science Center under Grant BMU2021YJ010.

Declarations

Ethics Approval The studies involving human participants were reviewed and approved by Biomedical Ethics Committee of Peking University (IRB00001052-22070).

Consent to participate Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Consent for publication No identifying information is included in this article. Informed consent for publication was not required for this study in accordance with the national legislation and the institutional requirements.

Conflicts of interest No potential conflict of interest was reported by the authors.

Reference

Agresti, A. (2012). *Categorical data analysis* (792nd ed.). John Wiley & Sons.

Ark, T. K. (2015). *Ordinal generalizability theory using an underlying latent variable framework [Dissertation]*. University of British Columbia. https://doi.org/10.14288/1.0166304

Arterberry, B. J., Martens, M. P., Cadigan, J. M., & Rohrer, D. (2014). Application of generalizability theory to the big five inventory. *Personality and Individual Differences, 69*, 98–103.

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (2015). Structure and flexibility in Bayesian models of cognition. In J. R. Busemeyer, J. T. Townsend, Z. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 187–208). Oxford University Press.

Barnard, J., McCulloch, R., & Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica, 10*, 1281–1311.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bischof, D. L. (2005). Validating the AP® German language exam through a curricular survey of third-year college language courses. *Die Unterrichtspraxis/Teaching German, 38*, 74–81.

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Brennan, R. L. (2001). *Manual for mGENOVA*. University of Iowa, IA Testing Programs. (Iowa Testing Programs Occasional Papers No. 50).

Brennan, R. L., Kim, S. Y., & Lee, W. C. (2022). Extended multivariate generalizability theory with complex design structures. *Educational and Psychological Measurement, 82*(4), 617–642.

Broatch, J., Green, J., & Karl, A. T. (2018). RealVAMS: An R package for fitting a multivariate value-added model (VAM). *The R Journal, 10*(1), 22.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling, 1*(2), 103–124.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ..., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32.

Cronbach, L. J. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.

DeCarlo, L. T. (2024). Fused SDT/IRT models for mixed-format exams. *Educational and Psychological Measurement, 2024*, 00131644241235333.

Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Educational Measurement: Issues and Practice, 36*(1), 5–13.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics, 40*(5), 530–543.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement, 15*(1), 1–13.

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

Jiang, Z., & Raymond, M. (2018). The use of multivariate generalizability theory to evaluate the quality of subscores. *Applied Psychological Measurement, 42*(8), 595–612.

Jiang, Z., & Skorupski, W. (2018). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior Research Methods, 50*, 2193–2214.

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour, 5*(10), 1282–1291.

Lakes, K. D., & Hoyt, W. T. (2009). Applications of generalizability theory to clinical child and adolescent psychology research. *Journal of Clinical Child & Adolescent Psychology, 38*(1), 144–165.

Lee, W. C., Kim, S. Y., Choi, J., & Kang, Y. (2020). IRT approaches to modeling scores on mixed-format tests. *Journal of Educational Measurement, 57*(2), 230–254.

Marsman, M., & Wagenmakers, E. J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology, 14*(5), 545–555.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Rand.

McNeish, D. M. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using M plus. *Journal of Educational and Behavioral Statistics, 41*(1), 27–56.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335.

Park, S., & Beretvas, S. N. (2020). The multivariate multiple-membership random-effect model: An introduction and evaluation. *Behavior Research Methods, 52*, 1254–1270.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121–129.

Raymond, M. R., & Jiang, Z. (2020). Indices of subscore utility for individuals and subgroups based on multivariate generalizability theory. *Educational and Psychological Measurement, 80*(1), 67–90.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163–184.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., ..., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers, 1*(1), 1–26.

Serang, S., Zhang, Z., Helm, J., Steele, J. S., & Grimm, K. J. (2015). Evaluation of a Bayesian approach to estimating nonlinear mixed-effects mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 202–215.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46*(4), 553–611.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Shi, D., & Tong, X. (2017). The impact of prior information on Bayesian latent basis growth model estimation. *SAGE Open, 7*(3), 2158244017727039.

Sinharay, S. (2015). Assessment of person fit for mixed-format tests. *Journal of Educational and Behavioral Statistics, 40*(4), 343–365.

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 131–161.

Sung, L., Hayden, J., Greenberg, M. L., Koren, G., Feldman, B. M., & Tomlinson, G. A. (2005). Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *Journal of Clinical Epidemiology, 58*(3), 261–268.

van der Ark, L. A., Jorgensen, T. D., & ten Hove, D., et al. (2023). Factors affecting efficiency of interrater reliability estimates from planned missing data designs on a fixed budget. In M. Wiberg (Ed.), *Quantitative psychology* (pp. 1–15). Springer. https://doi.org/10.1007/978-3-031-27781-8_1

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods, 24*(2), 153–178. https://doi.org/10.1037/met0000177

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ..., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35–57.

Wasserman, R. H., Levy, K. N., & Loken, E. (2009). Generalizability theory in psychotherapy research: The impact of multiple sources of variance on the dependability of psychotherapy process ratings. *Psychotherapy Research, 19*(4–5), 397–408.

Wei, J., Cai, Y., & Tu, D. (2023). A mixed sequential IRT model for mixed-format items. *Applied Psychological Measurement, 47*(4), 259–274.

Open Practices Statement The datasets and code in the current study are include in the appendix and available on OSF (https://osf.io/wud3x/), and are available from the corresponding author on reasonable request. Requests to access the material should be directed to JO, ouyangjinying@bjmu.edu.cn.