Psychological Test and Assessment Modeling, Volume 60, 2018 (1), 53-80

Modeling rater effects using a combination of Generalizability Theory and IRT

Jinnie Choi[1] & *Mark R. Wilson*[2]

Abstract

Motivated by papers on approaches to combine generalizability theory (GT) and item response theory (IRT), we suggest an approach that extends previous research to more complex measurement situations, such as those with multiple human raters. The proposed model is a logistic mixed model that contains the variance components needed for the multivariate generalizability coefficients. Once properly set-up, we can estimate the model by straightforward maximum likelihood estimation. We illustrate the use of the proposed method with a real multidimensional polytomous item response data set from classroom assessment that involved multiple human raters in scoring.

Keywords: generalizability theory, item response theory, rater effect, generalized linear mixed model

[1] *Correspondence concerning this article should be addressed to:* Jinnie Choi, Research Scientist at Pearson, 221 River Street, Hoboken, NJ 07030; email: jinnie.choi@pearson.com.

[2] University of California, Berkeley

While item response theory (IRT; Lord, 1980; Rasch, 1960) and generalizability theory (GT; Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) share common goals in educational and psychological research in order to provide evidence of the quality of measurement, IRT and GT have evolved into two separate domains of knowledge and practice in psychometrics that rarely communicate with one another. In practice, it is often recommended that researchers and practitioners be able to use and understand both methods, and to distinguish the same term with different meanings (e.g., reliability) or different terms with similar meanings (e.g., unidimensional testlet design in IRT and p x (i : h) design in GT), neither of which is desirable or practical. The separate foundations and development of these two techniques have resulted in a wide gap between the two approaches and have hampered collaboration between those who specialize in each. Additionally, despite the theories' extensive applicability, IRT and GT are often applied to somewhat different areas of research and practice. For example, applications of GT are often found in studies on reliability and sampling variability of smaller-scale assessments. Meanwhile, IRT is, relatively speaking, more commonly and more widely employed, than GT for developing large-scale educational assessments, such as the Programme for International Student Assessment (PISA) and the ones currently used by the US National Center for Education Statistics (NCES), and the products of large testing companies such as Educational Testing Service (ETS). Moreover, most advanced applications of IRT and GT take only one approach, not both. Considering the advantages of the two theories, this limitation and bias in usage call for an alternative approach to promote a more efficient and unified way to deliver the information that can be provided by IRT and GT together.

Several researchers have undertaken efforts to find the solution to this separation. For example, the researchers either: (a) highlight the differences but suggest using both, consecutively (Linacre, 1993), (b) discuss the link between the models (Kolen & Harris, 1987; Patz, Junker, Johnson, & Mariano, 2002), or (c) propose a new approach to combine the two (Briggs & Wilson, 2007).

Linacre (1993) emphasized the difference between IRT and GT and suggested that decision-makers select either one or the other, or use both, based on the purpose of the analysis. Many researchers took this advice and used both the IRT and the GT models, for example, for performance assessments of English as Second Language students (Lynch & McNamara, 1998), for English assessment (MacMillan, 2000), for writing assessments of college sophomores (Sudweeks, Reeve, & Bradshaw, 2005), for problem-solving assessments (Smith & Kulikowich, 2004), and for clinical examinations (Iramaneerat, Yudkowsky, Myford, & Downing, 2008).

While Linacre's suggestion promoted the idea of combining the use of the models, the statistical notion of links between IRT and GT began to emerge when Kolen and Harris (1987) proposed a multivariate model based on a combination of IRT and GT. The model assumed that the true score in GT could be approximated by the proficiency estimate in IRT. Patz, Junker, Johnson, & Mariano (2002) proposed a new model that combines IRT and GT, namely, the hierarchical rater model (HRM), which they see as a standard generalizability theory model for rating data, with IRT distributions replacing the normal theory true score distributions that are usually implicit in inferential applications of the model. The proposed use of the model is open to other possible extensions, although it is currently conceptualized as being used for estimation of rater effects.

These efforts motivated a noteworthy advance in combining IRT and GT, namely, the Generalizability in Item Response Modeling (GIRM) approach by Briggs & Wilson (2007), and its extensions by Choi, Briggs, & Wilson (2009) and Chien (2008). The GIRM approach provides a method for estimating traditional IRT parameters, the GT-comparable variance components, and the generalizability coefficients, not with observed scores but with the "expected item response matrix" — EIRM. By estimating a crossed random effects IRT model within a Bayesian framework, the GIRM procedure constructs the EIRM upon which GT-comparable analysis can be conducted. The steps can be described as follows:

Step 1. The probability of a correct answer is modeled using a crossed random effects item response model that considers both person and item as random variables. The model parameters are estimated using the Markov chain Monte Carlo (MCMC) method with the Gibbs sampler.

Step 2. Using the estimates from Step 1, the probability of the correct answer for each examinee answering each item is predicted to build the EIRM.

Step 3. The variance components and generalizability coefficients are estimated based on the EIRM.

Estimation of the variance components and generalizability coefficients utilizes an approach described by Kolen and Harris (1987) that calculates marginal integrals for facet effects, interaction effect, and unexplained error using the prior distributions and the predicted probabilities of IRT model parameters.

The main findings of the Briggs & Wilson (2007) study were as follows:

- GIRM estimates are comparable to GT estimates in the simple *p x i* test design where there are person and item facets alone, and with binary data.
- GIRM easily deals with the missing data problem, a problem for earlier approaches, by using the expected response matrix.
- Because GIRM combines the output from IRT with the output from GT, GIRM provides more information than either approach in isolation.
- Although GIRM adds the IRT assumptions and distributional assumptions to the GT sampling assumptions, GIRM is robust to misspecification of item response function and prior distributions.

In the multidimensional extension of the same method, Choi, Briggs, and Wilson (2009) found that the difference between GIRM and traditional GT estimates is more noticeable, with GIRM producing more stable variance component estimates and generalizability coefficients than traditional GT. Noticeable patterns of differences included the following:

- GIRM item variance estimates were smaller and more stable than GT,

- GIRM error variance (*pi* + *e*) estimates were larger and more stable than GT residual error variance (*pie*) estimates, and
- GIRM generalizability coefficients were generally larger and more precise than GT generalizability coefficients.

With the testlet extension of the procedure by Chien (2008), (a) the estimates of the person, the testlet, the interaction between the item and testlet, and the residual error variance estimates were found to be comparable to traditional GT estimates when data are generated from IRT models. (b) For the dataset generated from GT models, the interaction and residual variance estimates were slightly larger while person variance estimates were slightly smaller than traditional GT estimates. (c) The person-testlet interaction variance estimates were slightly larger than the traditional GT estimates for all conditions. (d) When the sample size was small, the discrepancy between the estimated universe mean scores in GT and the expected data in GIRM increased. (e) MCMC standard errors were notably underestimated for all variance components.

The mixed results from the studies of GIRM and its extensions yielded interesting questions.

- What is the statistical nature of the EIRM? The main advantage of the GIRM procedure comes from this matrix, coupled with the MCMC estimation within a Bayesian framework. This is a notable departure from the analogous-ANOVA estimation of traditional GT that brings the following benefits: (a) the variance component estimates are non-negative and (b) the problems that arise from unbalanced designs and missing data are easily taken care of. However, the extension studies revealed that the EIRM does not theoretically guarantee the equivalence of GIRM and traditional GT estimates in more complicated test conditions.
- Then, what is the benefit of having the extra step that requires multiple sets of assumptions and true parameters for each stage?
- Are there other ways to deal with the negative variance estimate problem in traditional GT and the missing data problem, and still get comparable results?
- Among the different approaches, which procedure gives more correct estimates?

These questions led to the search for an alternative strategy that requires simple one-stage modeling, and possibly non-Bayesian estimation that produces GT-comparable results, while capturing the essence of having random person and item parameters and variance components. The following section describes a different approach, one within the GLLAMM framework, to combine GT and IRT. In the next section, we explain how the random person and item parameters are estimated using a Laplace approximation implemented in the lmer() function (Bates, Maechler, Bolker, & Walker, 2014) in the R Statistical Environment (R Development Core Team, 2017). After that, we demonstrate applications of our approach to classroom assessment data from the 2008-2009 Carbon Cycle project, which includes 1,371 students' responses to 19 items, rated by 8 raters.

The proposed model

This paper uses a generalized linear latent and mixed model (GLLAMM; Skrondal & Rabe-Hesketh, 2004; Rabe-Hesketh, Skrondal, & Pickles, 2004) approach as an alternative to existing efforts to combine GT and IRT. GLLAMM offers a flexible one-stage modeling framework for a combination of crossed random effects IRT models and GT variance components models. The model is relatively straight-forward to formulate and easily expandable to more complex measurement situations such as multidimensionality, polytomous data, and multiple raters. In this section, we describe how the model specifies a latent threshold parameter as a function of cross-classified person, item, and rater random effects and the variance components for each facet.

GLLAMM is an extended family of generalized linear mixed models (Breslow & Clayton, 1993; Fahrmeir & Tutz, 2001), which was developed in the spirit of synthesizing a wide variety of latent variable models used in different academic disciplines. This general model framework has three parts. The response model formulates the relationship between the latent variables and the observed responses via the linear predictor and link function, which accommodates various kinds of response types. The structural model specifies the relationship between the latent variables at several levels. Finally, the distribution of disturbances for the latent variables is specified. For more details, see Rabe-Hesketh et al. (2004) and Skrondal & Rabe-Hesketh (2004). In this section, GT and IRT are introduced as special case of GLLAMM. Then, the GLLAMM approach to combining GT and IRT is detailed.

GT in the GLLAMM framework

The GLLAMM framework for traditional GT models consists of the response model for continuous responses, and multiple levels of crossing between latent variables. A multifaceted measurement design with person, item and rater facets will be used for an example. First, suppose, for the moment, that there is a continuous observed score for person *j* on item *i* rated by rater *k* which is modeled as

$$y_{ijk} = \nu_{ijk} + \epsilon_{ijk}, \tag{1}$$

Where the error ϵ_{ijk} has variance σ and the linear predictor ν_{ijk} is defined as a three-way random effects model

$$\nu_{ijk} = \beta_0 + \eta_{1i}^{(2)} + \eta_{2j}^{(2)} + \eta_{3k}^{(2)}, \tag{2}$$

where β_0 is the grand mean in the universe of admissible observations. $\eta_{1i}^{(2)}$, $\eta_{2j}^{(2)}$, and $\eta_{3k}^{(2)}$ are interpreted as item, person, and rater effects, respectively. The (2) superscript denotes that the units of the variable vary at level 2. The subscript starts with a number identifier for latent variables and the alphabetical identifier for units. These effects are not considered nested but crossed because each person could have answered any item, each person's responses could have been rated by any rater, and each rater could have rated any item. The model with interactions between the random effects can be written as a reduced form multilevel model,

$$\nu_{ijk} = \beta_0 + \eta_{1i}^{(3)} + \eta_{2j}^{(3)} + \eta_{3k}^{(3)} + \eta_{4ij}^{(2)} + \eta_{5jk}^{(2)} + \eta_{6ik}^{(2)}, \tag{3}$$

assuming that the interaction effects are latent variables varying at level 2 and the cluster-specific main effects are varying at level 3.

In traditional GT, the latent variables are assumed to equal the disturbances without covariates or factor loadings. Thus, they are described as the random intercepts such that $\eta_{1i}^{(3)} = \zeta_{1i}^{(3)}$, $\eta_{2j}^{(3)} = \zeta_{2j}^{(3)}$, $\eta_{3k}^{(3)} = \zeta_{3k}^{(3)}$, $\eta_{1ij}^{(2)} = \zeta_{1ij}^{(2)}$, $\eta_{2jk}^{(2)} = \zeta_{2jk}^{(2)}$, and $\eta_{3ik}^{(2)} = \zeta_{3ik}^{(2)}$. The distribution of the disturbances can be specified as $\zeta_{1i}^{(3)} \sim N(0, \psi_1^{(3)})$, $\zeta_{2j}^{(3)} \sim N(0, \psi_2^{(3)})$, $\zeta_{3k}^{(3)} \sim N(0, \psi_3^{(3)})$, $\zeta_{4ij}^{(2)} \sim N(0, \psi_4^{(2)})$, $\zeta_{5jk}^{(2)} \sim N(0, \psi_5^{(2)})$, and $\zeta_{6ik}^{(2)} \sim N(0, \psi_6^{(2)})$.

The generalizability coefficient is defined as the ratio of the universe score variance to the sum of the universe score variance and relative error variance.

$$E(\hat{\rho}_J^2) = \frac{\widehat{Var}\left(\eta_{2j}^{(3)}\right)}{\widehat{Var}\left(\eta_{2j}^{(3)}\right) + \widehat{Var}\left(\eta_{4ij}^{(2)}\right) + \widehat{Var}\left(\eta_{5jk}^{(2)}\right) + \widehat{Var}(\epsilon_{ijk})} = \frac{\hat{\varphi}_2^{(3)}}{\hat{\varphi}_2^{(3)} + \hat{\varphi}_4^{(2)} + \hat{\varphi}_5^{(2)} + \hat{\sigma}} \tag{4}$$

The index of dependability is defined as the ratio of the universe score variance to the total variance that includes the universe score variance and absolute error variance.

$$\begin{aligned} \widehat{\Phi}_J &= \frac{\widehat{Var}(\eta_{2j}^{(3)})}{\widehat{Var}(\eta_{1i}^{(3)}) + \widehat{Var}\left(\eta_{2j}^{(3)}\right) + \widehat{Var}(\eta_{3k}^{(3)}) + \widehat{Var}\left(\eta_{4ij}^{(2)}\right) + \widehat{Var}\left(\eta_{5jk}^{(2)}\right) + \widehat{Var}(\eta_{6ik}^{(2)}) + \widehat{Var}(\epsilon_{ijk})} \\ &= \frac{\hat{\varphi}_2^{(3)}}{\hat{\varphi}_1^{(3)} + \hat{\varphi}_2^{(3)} + \hat{\varphi}_3^{(3)} + \hat{\varphi}_4^{(2)} + \hat{\varphi}_5^{(2)} + \hat{\varphi}_6^{(2)} + \hat{\sigma}} \end{aligned} \tag{5}$$

IRT in the GLLAMM framework

The GLLAMM framework for traditional IRT models requires a response model for categorical responses, two levels of nesting, and a latent variable for persons (Skrondal & Rabe-Hesketh, 2004). An important difference between IRT and GT is the type of response that is modeled. As item responses are categorical, a classical latent response model can be formulated as introduced by Pearson (1901). The underlying continuous response y_{ij}^* is modeled[3] as

$$y_{ij}^* = \nu_{ij} + \epsilon_{ij}. \tag{6}$$

[3] *Note* that for continuous responses such as the ones modeled in traditional GT, $y_{ij}^* = y_{ij}$.

ν_{ij} is the log odds of correct answers to items i for person j conditional on person ability η_j and ϵ_{ij} has a logistic distribution, ϵ_{ij}~logistic, that has mean 0 and variance $\frac{\pi^2}{3}$. This is the same as writing the model with a logit link function, $\text{logit}\big(P(y_{ij} = 1|\eta_j)\big) = \nu_{ij}$ for dichotomous responses. Other distributions such as probit are used in certain cases when it is more appropriate to assume 1 for the error variance of the latent variable and when it is not desired to interpret the coefficients in terms of odds ratios.

For dichotomous responses, the observed response y_{ij} is defined as $y_{ij} = 1$ if $y_{ij}^* > 0$, and $y_{ij} = 0$ otherwise.

$$\ln\left(\frac{\Pr(y_{ij}^* > 0|\eta_j)}{\Pr(y_{ij}^* \leq 0|\eta_j)}\right) = \nu_{ij} \tag{7}$$

The Rasch model (Rasch, 1960) or the one-parameter (1PL) model, has a random intercept for persons and a fixed parameter for items denoted by

$$\nu_{ij} = \eta_j - \beta_i \tag{8}$$

where η_j is the latent variable for person j, $\eta_j \sim N(0,1)$, and β_i is the fixed effect for item i. In the two-parameter logistic model, or 2PL model, a slope parameter or a factor loading is added for each item such that

$$\nu_{ij} = \lambda_i(\eta_j - \beta_i) \tag{9}$$

where λ_i represents item discrimination.

For polytomous items, let C be the number of categories for an item. Assume that the category score is defined as c = 1, …, C-1, also representing the steps between the scores. In the polytomous case, each category score y_{icj} for the category score c is modeled with a separate linear predictor ν_{icj}. Depending on data and intended interpretation, one can specify the model differently. For example, using the sequential stage continuation ratio logit scheme, y_{icj} takes the value of 1 if $y_{icj}^* > c$ and 0 if $y_{icj}^* = c$ for category c. Then

$$\ln\left(\frac{\Pr(y_{icj}^* > c|\eta_j)}{\Pr(y_{icj}^* = c|\eta_j)}\right) = \nu_{icj}. \tag{10}$$

Using the adjacent category logit scheme (Agresti & Kateri, 2011), y_{icj} takes the value of 1 if $y_{icj}^* = c$ and 0 if $y_{icj}^* = c - 1$ for category c. The adjacent category logit specification is widely used in polytomous item response models such as the rating scale model (Andrich, 1978) and the partial credit model (Masters, 1982):

$$\ln\left(\frac{\Pr(y_{icj}^* = c|\eta_j)}{\Pr(y_{icj}^* = c-1|\eta_j)}\right) = \nu_{icj}. \tag{11}$$

In cumulative models for ordered categories, y_{icj} takes the value of 1 if $y^*_{icj} > c$ and 0 if $y^*_{icj} \leq c$ for category *c*. The graded response model (Samejima, 1969) is specified using cumulative probabilities and threshold parameters. When a logit link is used, the model is specified as,

$$\ln\left(\frac{\Pr(y^*_{icj} > c|\eta_j)}{\Pr(y^*_{icj} \leq c|\eta_j)}\right) = \nu_{icj}. \tag{12}$$

In the case of the partial credit model, the linear predictor is specified as

$$\nu_{icj} = c\eta_j - \beta_{ic} \tag{13}$$

with β_{ic} representing the *c*th step difficulty for item *i*. The graded response model uses a set of ordered threshold parameters κ_c such that

$$\nu_{icj} = \kappa_c\eta_j - \beta_{ic} \tag{14}$$

where κ_c can be viewed as the factor loadings for each step.

Zheng & Rabe-Hesketh (2007) presented the Rasch, 2PL, partial credit model and rating scale model using the additional parameters for covariates for latent variables and other parameters, so that the structure of item loading and scoring is more explicit.

Theoretical link and justification of combining GT and IRT using GLLAMM

The combination of GT and IRT ideas become simpler when GT and IRT features are expressed in the same GLLAMM language. The key elements of the combined model include: the latent response specification, a logit link, and a linear predictor specified as a crossed random effects model.

For dichotomous responses, the underlying continuous response to the *i*th item of the *j*th person rated by the *k*th rater is modeled using a classical latent response model:

$$y^*_{ijk} = \nu_{ijk} + \epsilon_{ijk}, \;\; \epsilon_{ijk} \sim logistic(\mu, \frac{\pi^2}{3}), \tag{15}$$

where ν_{ijk} designates the true score for every possible pair of units *i*, *j*, and *k*, or the expected responses. The observed response y_{ijk} is modeled as a threshold that takes the value of 1 if $y^*_{ijk} > 0$ and 0 otherwise.

The linear predictor ν_{ijk} is defined as a crossed random effects model with or without interaction. For simplicity, a model without interaction is presented here as

$$\nu_{ijk} = \beta_0 + \eta^{(2)}_{1i} + \eta^{(2)}_{2j} + \eta^{(2)}_{3k} \tag{16}$$

where β_0 is the average logit of the probability of response 1 averaging over all persons, items, and raters. Note that the effects for persons and items are not considered nested but crossed because each person could have answered each item. As above, the (2) superscript denotes that the units of the variable vary at level 2. $\eta_{1i}^{(2)}$ is the first latent variable that varies among items (i = 1,…,I) at level 2, and $\eta_{2j}^{(2)}$ is the second latent variable at level 2 that varies among persons (j = 1, …, J). $\eta_{3k}^{(2)}$ is the third latent variable at level 2 that varies among raters (k = 1, …, K). The interpretations of $\eta_{1i}^{(2)}$, $\eta_{2j}^{(2)}$, and $\eta_{3k}^{(2)}$ are item easiness, person ability, and rater leniency, respectively. If the addition signs are switched to subtraction signs for $\eta_{1i}^{(2)}$ and $\eta_{3k}^{(2)}$, the interpretations are also reversed as item difficulty and rater severity.

The latent variables are assumed to equal the disturbances, $\eta_{1i}^{(2)} = \zeta_{1i}^{(2)}$, $\eta_{2j}^{(2)} = \zeta_{2j}^{(2)}$ and $\eta_{3k}^{(2)} = \zeta_{3k}^{(2)}$, which are specified as $\zeta_{1i}^{(2)} \sim N(0, \psi_1^{(2)})$, $\zeta_{2j}^{(2)} \sim N(0, \psi_2^{(2)})$, and $\zeta_{3k}^{(2)} \sim N(0, \psi_3^{(2)})$, corresponding to the assumptions of traditional GT.

In the case of person-specific unobserved heterogeneity, the model is specified as

$$\nu_{ijk} = \beta_0 + \eta_{1i}^{(2)} + \textstyle\sum_{d=1}^{D} \lambda_{id}\eta_{2jd}^{(2)} + \eta_{3k}^{(2)}, \;\; \zeta_{2jd}^{(2)} \sim MVN(\mathbf{0}, \mathbf{\Psi}_2^{(2)}). \tag{17}$$

with the number of dimensions D, the item factor loadings λ_{id}, and a covariance matrix $\mathbf{\Psi}$. For the Rasch model, λ_{id} is 1 if the ith item maps onto the dth dimension, 0 otherwise.

The continuation ratio approach is used for polytomous data, following Tutz's (1990) parameterization in his sequential stage modeling (De Boeck, Bakker, Zwitser, Nivard, Hofman, Tuerlinckx, & Partchev, 2011). y_{icjk} takes the value of 1 if $y_{icjk}^{*} > c$ and 0 if $y_{icjk}^{*} = c$, where c (c = 1, …, C–1) denotes the category score and C denotes the number of score categories including the score 0. The linear predictors ν_{icjk} for unidimensional and multidimensional cases are specified as

$$\nu_{icjk} = \beta_0 + \eta_{1ic}^{(2)} + \eta_{2j}^{(2)} + \eta_{3k}^{(2)}, \tag{18}$$

$$\nu_{icjdk} = \beta_0 + \eta_{1ic}^{(2)} + \sum_{d=1}^{D} \lambda_{id}\eta_{2jd}^{(2)} + \eta_{3k}^{(2)}, \;\; \zeta_{2jd}^{(2)} \sim MVN(\mathbf{0}, \mathbf{\Psi}_2^{(2)}). \tag{19}$$

Using the variance components estimates, the generalizability coefficient $E(\hat{\rho}_J^2)$ for the person estimates is calculated. In GT terms, $E(\hat{\rho}_J^2)$ is the ratio of the universe score variance to the sum of itself plus the relative error variance. The universe score variance is defined as the variance of all the scores in the population of all the persons, items, and raters. The relative error variance means the measurement error variance relevant to the relative rank order between persons. The variance of ϵ_{ijk} is included in the denominator to take into account the variance of the underlying logit. Additionally, using the variance component for items and raters in the model, we calculate the generalizability coefficient for measurement of item easiness, $E(\hat{\rho}_I^2)$, and rater leniency, $E(\hat{\rho}_K^2)$, in the same manner:

$$E(\hat{\rho}_J^2) = \frac{\widehat{Var}(\eta_{2j}^{(2)})}{\widehat{Var}(\eta_{2j}^{(2)}) + \widehat{Var}(\epsilon_{ijk})} = \frac{\hat{\varphi}_2^{(2)}}{\hat{\varphi}_2^{(2)} + \frac{\pi^2}{3}} \tag{20}$$

$$E(\hat{\rho}_I^2) = \frac{\widehat{Var}(\eta_{1i}^{(2)})}{\widehat{Var}(\eta_{1i}^{(2)}) + \widehat{Var}(\epsilon_{ijk})} = \frac{\hat{\varphi}_1^{(2)}}{\hat{\varphi}_1^{(2)} + \frac{\pi^2}{3}} \tag{21}$$

$$E(\hat{\rho}_K^2) = \frac{\widehat{Var}(\eta_{3k}^{(2)})}{\widehat{Var}(\eta_{3k}^{(2)}) + \widehat{Var}(\epsilon_{ijk})} = \frac{\hat{\varphi}_3^{(2)}}{\hat{\varphi}_3^{(2)} + \frac{\pi^2}{3}} \tag{22}$$

The index of dependability $\widehat{\Phi}$ is the ratio of the universe score variance to the sum of itself plus the absolute error variance. Absolute error variance focuses on the measurement error variance of a person that is attributed by the measurement facets regardless of how other people do on the test. Thus, $\widehat{\Phi}$ accounts for the variance related to another random facet, for example, items. The denominator also includes the variance of the underlying logit. The same logic can be extended to calculation of the indices of dependability for item easiness and rater leniency:

$$\widehat{\Phi_J} = \frac{\hat{\varphi}_2^{(2)}}{\hat{\varphi}_1^{(2)} + \hat{\varphi}_2^{(2)} + \hat{\varphi}_3^{(2)} + \frac{\pi^2}{3}} \tag{23}$$

$$\widehat{\Phi_I} = \frac{\hat{\varphi}_1^{(2)}}{\hat{\varphi}_1^{(2)} + \hat{\varphi}_2^{(2)} + \hat{\varphi}_3^{(2)} + \frac{\pi^2}{3}} \tag{24}$$

$$\widehat{\Phi_K} = \frac{\hat{\varphi}_3^{(2)}}{\hat{\varphi}_1^{(2)} + \hat{\varphi}_2^{(2)} + \hat{\varphi}_3^{(2)} + \frac{\pi^2}{3}} \tag{25}$$

In the multidimensional and/or polytomous case, we use the dimension- and category-specific variance component estimates along with the number of items in each dimension and with the number of persons who got each category score as weights to calculate the composite generalizability coefficient and the index of dependability (Brennan, 2001; Choi, Briggs & Wilson, 2009).

$$E(\hat{\rho}_d^2) = \frac{\hat{\varphi}_d^{(2)}}{\hat{\varphi}_d^{(2)} + \frac{\pi^2}{3}} \tag{26}$$

$$E(\hat{\rho}_c^2) = \frac{\hat{\varphi}_c^{(2)}}{\hat{\varphi}_c^{(2)} + \frac{\pi^2}{3}} \tag{27}$$

$$\widehat{\Phi_d} = \frac{\hat{\varphi}_d^{(2)}}{\hat{\varphi}_c^{(2)} + \hat{\varphi}_d^{(2)} + \hat{\varphi}_3^{(2)} + \frac{\pi^2}{3}} \tag{28}$$

$$\widehat{\Phi_c} = \frac{\hat{\varphi}_c^{(2)}}{\hat{\varphi}_c^{(2)} + \hat{\varphi}_d^{(2)} + \hat{\varphi}_3^{(2)} + \frac{\pi^2}{3}} \tag{29}$$

where $\hat{\varphi}_d^{(2)}$ is the composite of universe score variance on the person side, and $\hat{\varphi}_c^{(2)}$ is the composite of universe score variance on the item side. Formally, these are defined as

$$\hat{\varphi}_d^{(2)} = \sum_d w_d^2 \hat{\varphi}_{2d}^{(2)} + \sum\sum_{d' \neq d} w_d^2 w_{d'}^2 \hat{\varphi}_{2dd'}^{(2)} \tag{30}$$

$$\hat{\varphi}_c^{(2)} = \sum_c w_c^2 \hat{\varphi}_{1c}^{(2)} + \sum\sum_{c' \neq c} w_c^2 w_{c'}^2 \hat{\varphi}_{1cc'}^{(2)} \tag{31}$$

where the weights $w_d = \frac{n_{id}}{n_I}$ for d = 1,…D and $w_c = \frac{n_{jc}}{n_J}$ for c = 1,…C–1, n_I is the total number of items over all dimensions, n_{id} is the number of items in dimension d, n_J is the total number of items over all dimensions, and n_{jc} is the number of persons who got each category score.

Estimation

For generalized linear mixed models with crossed random effects, the likelihood of the data given the random variables needs to be integrated over the latent distribution. Since the high-dimensional likelihood function does not have a closed form in general, there are several approaches to approximating the maximum likelihood. The Laplacian approximation evaluates the unscaled conditional density at the conditional mode and is optimized with respect to the fixed effects and the disturbances. It is equivalent to the adaptive Gaussian quadrature with one node and is most accurate when the integrand of the likelihood is proportional to a normal density. Thus, a large cluster size corresponds to close-to-normal posterior density of the random variables, which then again leads to better approximation and less bias in estimates, especially for person parameter estimation (Cho & Rabe-Hesketh, 2011; De Boeck et al., 2011; Joe, 2008; Pinheiro & Bates, 1995; Skrondal & Rabe-Hesketh, 2004).

Specifically, the model is fitted using the computational method implemented in the **lme4** package (Bates, 2010). Given the response vector $\mathcal{Y}$, the q-dimensional random effect vector $\mathcal{B}$, the variance-component parameter vector θ, the scale parameter σ for which it is assumed that $\sigma > 0$, and a multivariate Gaussian random variable $\mathcal{U}$ such that $\mathcal{B} = \Lambda_\theta \mathcal{U}$ where a covariance matrix Λ_θ satisfies

$$Var(\mathcal{B}) = \Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^{\intercal}, \tag{32}$$

the joint density function of $f_{\mathcal{U},\mathcal{Y}}(\boldsymbol{u}, \boldsymbol{y})$ is evaluated at the observed vector $\boldsymbol{y}_{obs}$. The continuous conditional density $f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y}_{obs})$ can be expressed as a function of an unnormalized conditional density $h(\boldsymbol{u})$, of which integral $\int f(\boldsymbol{u})d\boldsymbol{u}$ is the same as the likelihood that needs to be evaluated for our model fitting.

Since the integral does not have a closed form for the kinds of mixed model we are interested in, it is evaluated using the Laplace approximation that utilizes the Cholesky factor $\boldsymbol{L}_\theta$ and the conditional mode $\widetilde{\boldsymbol{u}}$. The conditional mode of u given $\mathcal{Y} = \boldsymbol{y}_{obs}$ is defined as a maximizer of the conditional density and a minimizer of a penalized residual sum of squares (PRSS) criterion or a function of the parameters given the data,

$$r^2_{\theta,\beta} = \min_{\boldsymbol{u}} \|\boldsymbol{y}_{obs} - \mu\|^2 + \|\boldsymbol{u}\|^2, \tag{33}$$

where μ is the mean of the conditional density. The Cholesky factor $\boldsymbol{L}_\theta$ is defined as the sparse lower triangular $q \times q$ matrix with positive diagonal elements such that

$$\boldsymbol{L}_\theta \boldsymbol{L}_\theta^{\intercal} = \boldsymbol{\Lambda}_\theta^{\intercal} \boldsymbol{Z}^{\intercal} \boldsymbol{Z} \boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q. \tag{34}$$

The sparse triangular matrix $\boldsymbol{L}_\theta$ can be efficiently evaluated even with large data sets by the fill-reducing permutation that reduces the number of non-zeros in the factor. After evaluating $\boldsymbol{L}_\theta$ and solving for $\widetilde{\boldsymbol{u}}$, the likelihood can be conveniently expressed as a function of σ, $\boldsymbol{L}_{\theta,\beta}$, and $r^2_{\theta,\beta}$. On a deviance scale, the Laplace approximation of the likelihood is given as

$$d(\theta, \beta, \sigma|y_{obs}) = -2\log(L(\theta, \beta, \sigma|y_{obs})) \approx= n \cdot \log(2\pi\sigma^2) + 2 \cdot \log|\boldsymbol{L}_{\theta,\beta}| + \frac{r^2_{\theta,\beta}}{\sigma^2} \tag{35}$$

and the parameter estimates are the values at which this deviance is minimized.

Currently, xtmelogit in Stata and the lmer() function in R are as available as general statistical packages that have the capacity to estimate one or more cross-classified random variables using the Laplacian approximation, while the lmer() function is significantly more efficient than xtmelogit with regard to the calculation time. Therefore, we chose the lmer() function to estimate our model parameters.

In addition to the IRT random person and item variables, we also parameterize and estimate the variance components using lmer(). This direction also corresponds to the recommendation by Robinson (1991), Gelman, Carlin, Stern, & Rubin (2003), and Gelman (2005) to treat the hierarchical regression coefficients as random variables (and thus 'predicted') and the variance components as parameters (and thus 'estimated'). In traditional GT, since it depends on the analogous-ANOVA variance decomposition procedure based on the design of the existing data, there are known limitations such as negative variance estimates. ANOVA's main advantage is the ease of variance component estimation, but it is mostly applied to balanced designs. With proper reconstruction of data, lmer() easily estimates the variance components of incomplete data, which, we argue, would serve as a significant improvement of the problems in traditional GT. In addition, there has been no clearly best estimation method for variance decomposition of incomplete data and unbalanced mixed designs. Even though the resulting estimates have been proved to be unbiased, other properties of estimates are generally unknown (Khuri, 2000; Khuri & Sahai, 1985; Skrondal & Rabe-Hesketh, 2004). It is thus useful to know that for unbalanced multistrata ANOVA, lmer() is preferred to estimate variance components rather than the aov() and Anova() functions, which are also currently available in R for general ANOVA.

The key to estimating the generalizability coefficients lies in using proper variance component estimates for diverse measurement situations. We take the variance component estimates from lmer() and use the calculation methods from Brennan (2001), which includes the most comprehensive set of calculation methods for these coefficients in measurement situations that match a variety of complex ANOVA-like designs. Recall that the classical definition of reliability is the proportion of the total variance of the measurements that is due to the true score variance. We take this definition to calculate the generalizability coefficient $E\left(\hat{\rho}_J^2\right)$ for person measurement. In the same manner, we calculate the generalizability coefficient for measurement of item easiness and rater leniency as specified in Equations (20) to (22). In the multidimensional and/or polytomous case, the dimension-specific and category-specific variance components are estimated as specified in Equations (26) to (29).

Through extensive simulation studies (Choi, 2013), the accuracy of the results from the proposed approach in various measurement conditions was evaluated. In conclusion, the simulation results suggested that the proposed approach gives overall accurate generalizability coefficients. While more students and skewness in person distributions showed a significant interaction effect on the accuracy of the generalizability coefficients, the effect sizes were all very small. The next section presents the datasets and design of an empirical study.

The example data

The illustrative data set has three features that illuminate the utility of the proposed method: multidimensionality, polytomous responses, and multiple raters. The data was collected by researchers from Michigan State University and the University of California, Berkeley for the Carbon Cycle project, which was supported by the National Science Foundation: *Developing a Research-based Learning Progression on Carbon-Transforming Processes in Socio-Ecological Systems* (NSF 0815993). The participants included U.S. students from the state of Michigan in grades 4 through 12 during the 2008–2009 school year. After the data were cleaned, the data consisted of 869 students, including 190 elementary students, 346 middle school students, and 333 high school students. The 19 items in the data set represented six latent ability domains and were polytomously scored into four categories by 8 raters. The numbers of items designed to measure each of the dimensions were 3, 3, 3, 2, 3, 3, and 5, respectively. However, not every item was designed to measure all six domains, not every item was rated by all 8 raters, not every person answered all items, not every item had four category scores, and so on. That is, the data was unbalanced and incomplete. The reshaping of the data, based on Tutz's (1990) sequential stage continuation ratio logit, resulted in a response vector with a length of 18,695. A unidimensional model and a multidimensional model for polytomous data with a rater facet were fitted to this data set.

The models fitted to the Carbon Cycle 2008-2009 empirical data sets were (a) a unidimensional model (without raters), UD, (b) a unidimensional model (with raters), UDR, (c) a multidimensional model (without raters), MD, and (d) a multidimensional model (with raters), MDR. The composite person and item variance components are the weighted averages based on the number of items per each dimension and the number of persons who scored each category, respectively. The results are summarized in Table 1.

The person, item, and rater variance component estimates stay relatively stable across the four models. Overall, adding the rater effect to the unidimensional model (UD to UDR) and to the multidimensional model (MD to MDR) did not result in noticeable changes in the person and item variance component estimates and generalizability coefficients. The person generalizability coefficients decreased about 0.02 on average: from 0.340 and 0.249 to 0.315 and 0.224 for the unidimensional case, and from 0.376 and 0.286 to 0.352 and 0.261 for the multidimensional case. The item generalizability coefficients changed on average less than 0.005: from 0.358 and 0.269 to 0.360 and 0.274 for the unidimensional case, and from 0.337 and 0.241 to 0.335 and 0.243 for the multidimensional case.

Table 1:

Estimated Variance Components and Generalizability Coefficients

	Model			
	UD	UDR	MD	MDR
	Est	Est	Est	Est
Person	1.696	1.509	1.981(c)	1.783(c)
Dim 1			2.803	2.678
Dim 2			2.044	1.785
Dim 3			1.105	1.062
Dim 4			2.066	1.977
Dim 5			2.321	1.978
Dim 6			1.623	1.386
Item	1.836(c)	1.845(c)	1.669(c)	1.658(c)
Step 1	2.799	2.780	2.470	2.446
Step 2	0.639	0.659	0.541	0.559
Step 3	2.089	2.123	2.046	2.015
Rater		0.093		0.092
Error	3.287	3.287	3.287	3.287
AIC	12,168	12,130	12,177	12,140
BIC	12,246	12,216	12,451	12,422
Dev.	12,148	12,108	12,107	12,068
GCP	0.340	0.315	0.376	0.352
GCI	0.358	0.360	0.337	0.335
GCR		0.027		0.027
IDP	0.249	0.224	0.286	0.261
IDI	0.269	0.274	0.241	0.243
IDR		0.014		0.013

Notes. 1. The item and rater parameters are interpreted as easiness and leniency, respectively. 2. The (c) marks are for the weighted or composite variance component estimates.

Similarly, adding the multiple dimensions did not produce significantly different generalizability coefficients for person, item, and rater facets. Compared to the unidimensional results, the person-side generalizability coefficients were slightly greater (about a 0.035 increase on average) while the item-side generalizability coefficients were slightly smaller (about a 0.025 decrease on average). Including the multiple person dimensions in the model only slightly improved the deviances: from 12,148 to 12,107 for the models without the rater effect and from 12,108 to 12,068 for the models with the rater effect.

However, the AIC and the BIC showed that the multidimensional models did not fit better than the unidimensional models did. Thus, we can select the simpler unidimensional model when summarizing the generalizability coefficients for the Carbon Cycle data analysis. The best fit was found for the unidimensional model with the rater effect (UDR), where the generalizability coefficient and the index of dependability of the person measurements (0.315 and 0.224) were about 0.05 logit less than those of the item measurements (0.360 and 0.274). The rater variance component was very small (0.093) and the resulting generalizability coefficients for the raters were also very small (0.027 and 0.014).

Previous research on the multidimensionality of the same data using multidimensional item response models also reported high latent correlations between the dimensions, as shown in Table 2 (Choi, Lee, & Draney, 2009). All empirical results lead to the conclusion that the person dimensions are statistically indistinguishable.

We use the results from the polytomous multidimensional model with the rater effect to illustrate the advantages of using the proposed approach. While the model fit was worse than the unidimensional model, we purposefully chose this model because our goal here is to demonstrate the extendibility of the proposed approach to more complex test conditions. We can not only estimate random variance components of the persons, items, and raters but also estimate the individual person, item, and rater estimates via the proposed method. Table 3 shows examples of those predicted individual estimates. When the intercepts and group means are added, these estimates are comparable to the traditional item response theory person ability, item step difficulty (easiness with reversed sign), and rater severity (leniency with reversed sign) estimates on the same logit scale. For example, the fixed effects estimates for the grand mean and the dimension 2 were 1.998 and 0.177, respectively. Thus, the estimated ability for the dimension 2 of the student S00001 is 1.998 + 0.177 – 2.623 = –0.448 logit.

Table 2:

The correlation between six person dimensions for the 2008-2009 Carbon Cycle data

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
Dimension 2	0.994				
Dimension 3	0.989	0.999			
Dimension 4	0.857	0.901	0.911		
Dimension 5	0.962	0.984	0.988	0.904	
Dimension 6	0.992	1.000	1.000	0.911	0.983

Next, Figure 1 and Figure 2 present the precision in predicting the random person ability, item difficulty, and rater severity effects. First, in Figure 1, the students and items are ordered from left to right according to increasing standard normal quantiles. The dots are the conditional modes of the random effects, and the lines indicate the 95% prediction intervals. The prediction interval is obtained from the conditional standard deviation, which is a measure of the dispersion of the parameter estimates given the data (Bates et al., 2015). The x axis is the standard normal quantiles for students and items and the y axis is the logit scale. The patterns show that about 40% to 50% of the person random effects contain zero in their prediction interval while most of the item random effects do not. This means that for the students it is more probable that their ability estimate is close to the mean than for the items. This is reasonable because the cluster sizes for person estimation are much smaller than those for items.

Figure 2 depicts the pattern of the 95% prediction intervals on the person and item random effects ordered differently according to increasing estimated values for the first level (e.g., dimension 1 for persons, step 1 for items). By doing so, we can discern whether the patterns of the dimensions are similar to each other. The x axis is the logit scale and the y axis is the persons or the items, respectively. The graph confirms that the person dimensions are highly correlated with the dimension 1 except for dimension 4, as shown by the previous results of the latent correlation from a multidimensional IRT analysis shown in Table 2.

Table 3:

An example set of predicted person, item, and rater effects estimates

Student ID	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
S00001	-3.178	-2.623	-2.029	-2.257	-2.679	-2.318
S00002	-0.211	-0.161	-0.102	-0.017	-0.138	-0.136
S00003	-1.141	-0.925	-0.688	-0.647	-0.900	-0.808
S00004	-0.202	-0.191	-0.187	-0.377	-0.159	-0.170
S00005	-1.578	-1.307	-1.019	-1.166	-1.237	-1.144
S00006	-1.441	-1.144	-0.811	-0.583	-1.182	-1.002
S00007	-1.885	-1.564	-1.223	-1.418	-1.643	-1.389
S00008	-1.405	-1.179	-0.944	-1.186	-1.151	-1.040
S00009	-2.677	-2.186	-1.651	-1.665	-2.222	-1.924
S00010	-2.593	-2.128	-1.625	-1.718	-2.197	-1.880
S00011	-0.977	-0.821	-0.658	-0.831	-0.776	-0.721
S00012	-1.597	-1.289	-0.951	-0.855	-1.193	-1.118
S00013	-1.441	-1.210	-0.968	-1.215	-1.353	-1.087
S00014	-1.277	-1.076	-0.867	-1.113	-1.048	-0.949
S00015	-0.853	-0.663	-0.446	-0.207	-0.571	-0.563
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Note. All students shown here are at elementary school level, hence the low estimates.

Item ID	Step 1	Step 2	Step 3
i01	0.949	0.092	0.960
i02	-2.519	-1.355	-0.278
i03	-0.842	0.268	0.302
i04	-0.715	-0.903	0.061
i05	1.307	0.308	0.755
i06	1.128	0.901	0.215
i07	-2.832	-0.916	-3.245
i08	-0.772	-0.166	0.521
i09	1.255	0.558	0.672
i10	-1.354	-0.802	0.157
i11	-0.148	-0.001	0.726
i12	0.212	-0.588	1.121
i13	-0.198	-0.092	2.055
i14	0.020	0.321	-0.271
i15	1.759	0.611	1.592
i16	-0.155	-0.166	2.304
i17	-1.167	0.250	-2.576
i18	-1.998	-1.529	-2.813
i19	1.803	0.971	0.557

Rater ID	Estimates
r1	-0.014
r2	0.200
r3	0.014
r4	0.270
r5	-0.005
r6	-0.240
r7	-0.551
r8	0.182

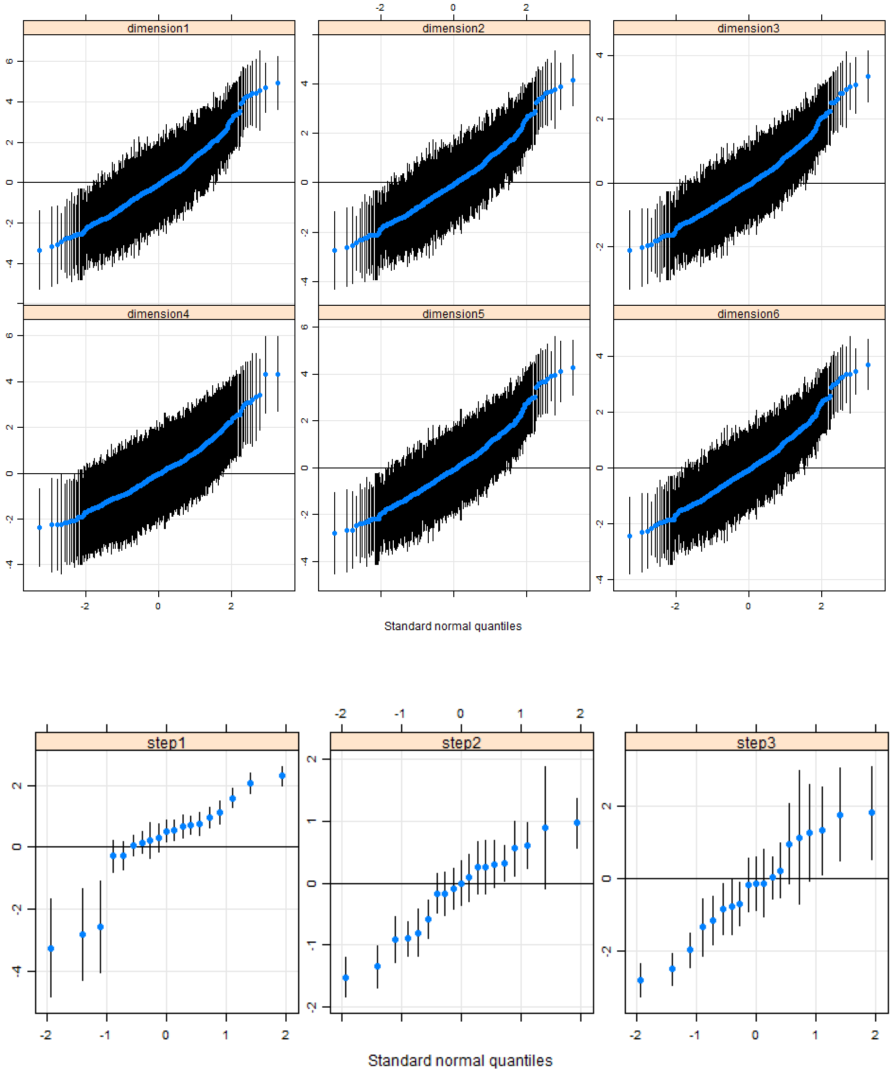

Figure 1:

95% prediction intervals on the person and item random effects compared to the standard normal quantiles of students and items

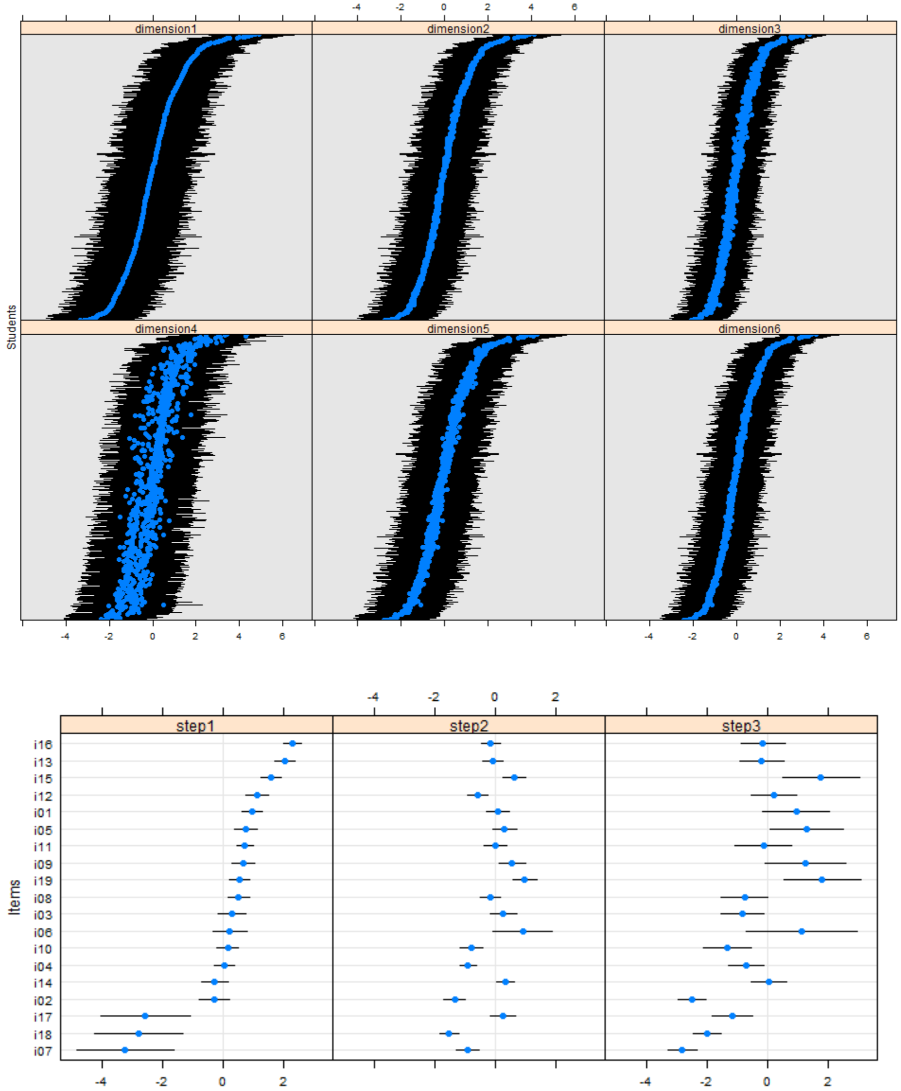

Figure 2:

95% prediction intervals on the person and item effects ordered based on increasing estimated values for the first levels (e.g., dimension 1 and step 1)

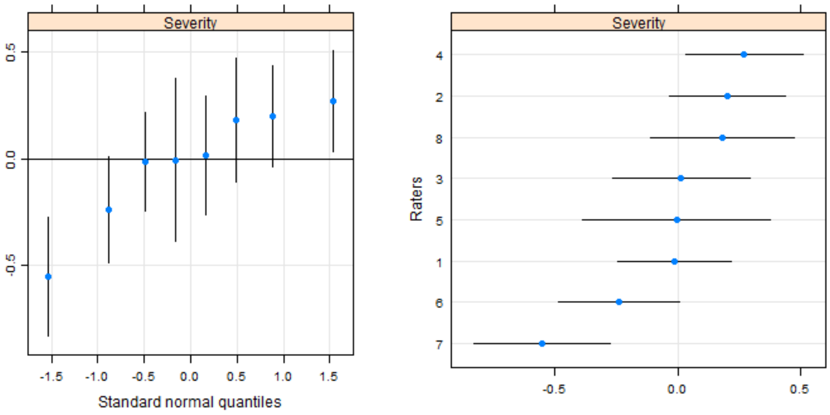

Figure 3:

95% prediction intervals on the rater effects ordered based on 1) the standard normal quantiles and 2) increasing estimated values

On the other hand, the item step estimates are not showing much correlation among the steps. What is more interesting in the item graphs is that we can observe for some items (e.g., the first three items from the bottom: i7, i17 and i18), achieving the second and the third steps was relatively easier than for other items. The greater imprecision (i.e., longer bars crossing the estimate) for the first step is caused by the small number of responses at that level of performance, compared to the responses at higher levels of performances which showed greater precision (i.e., shorter bars).

The 95% prediction intervals for the rater random effects are shown in Figure 3. In the graph on the left, the x axis is the standard normal quantiles and the y axis is the logit scale. Overall, the rater estimates are quite close to each other and to zero, as we should expect, since the raters went through training and screening procedures. Unlike the person and item estimates, the y scale ranges narrowly between -0.5 and 0.5 logits. Only two rater estimates do not have prediction intervals that contain zero. In the graph on the right, the x axis is the logit scale and the y axis is the persons or the items. The rater 7 was the most lenient: when other variables were controlled, it was easier for the students to get the scores on items when rated by the rater 7.

The benefit of having the set of predicted person, item, and rater effects in Table 3 is explicitly shown through Figure 4, a modified version of a Wright Map (Wilson, 2005). The person ability estimates for each dimension are calculated as sums of the estimated intercept, the cluster mean for each dimension (except the reference dimension 1), and the estimated person random effects. Likewise, the item difficulty estimates for each step are calculated as sums of the estimated intercept, the cluster mean for each step (except for the reference step 1), and the estimated item random effects. The rater severity estimates were calculated as the means of the rater severity for the three thresholds. One can compare the location of the distributions of the person, item, and rater parameter estimates on the same logit scale, which gives insight into how easy it was for the students to get the target score on the items by the raters.

The person ability scores are displayed as density curves on the logit scale. Dimensions 3 and 4 have relatively small variances and the shape of dimension 4 density is slightly off from the symmetric normal distribution. The ability distributions of the six dimensions shared similar locations and dispersions, again validating the finding that the multidimensional model does not fit the data better than the unidimensional one. The result corresponds with the lower latent correlation for dimension 4 with other dimensions as well as the fuzzy dots in the 95% prediction interval graph. Most likely, the reason why dimension 4 behaves as an oddity is the very low number of items (2) that measures it.

Next, the three groups of points represent item difficulties. The location of the points can be interpreted as where the students on average have a 50% probability to get a particular score as compared to a score below that. Since the data had four possible scores (0,1,2,3), three steps or thresholds exist between the four scores. As the Wright Map shows, most of the students had more than 50% chance to achieve the first thresholds of all items, except for the items 13 to 16. In other words, for most of the students, getting the score 1 versus 0 was relatively easy. The second thresholds of the items were reasonably located near the means of the person distributions, meaning that on average students were able to get the score 2 on most items with about a 50% probability of success. Getting the highest score was generally difficult for the students, particularly for the items 1, 5, 6, 9, and 15.

Last, the raters were generally non-separable from each other except for the rater 6 and 7 who were on average more lenient than others in giving scores (compared to the score below) for the items to the students. The small variance component, the small resulting generalizability coefficient, and non-separable individual rater effects that we found in this analysis suggest that the rater training sessions were highly effective. The eight raters were indeed graduate research assistants who were involved in every stage of the research process — thus for this group of raters it makes sense that they showed consistent ratings.

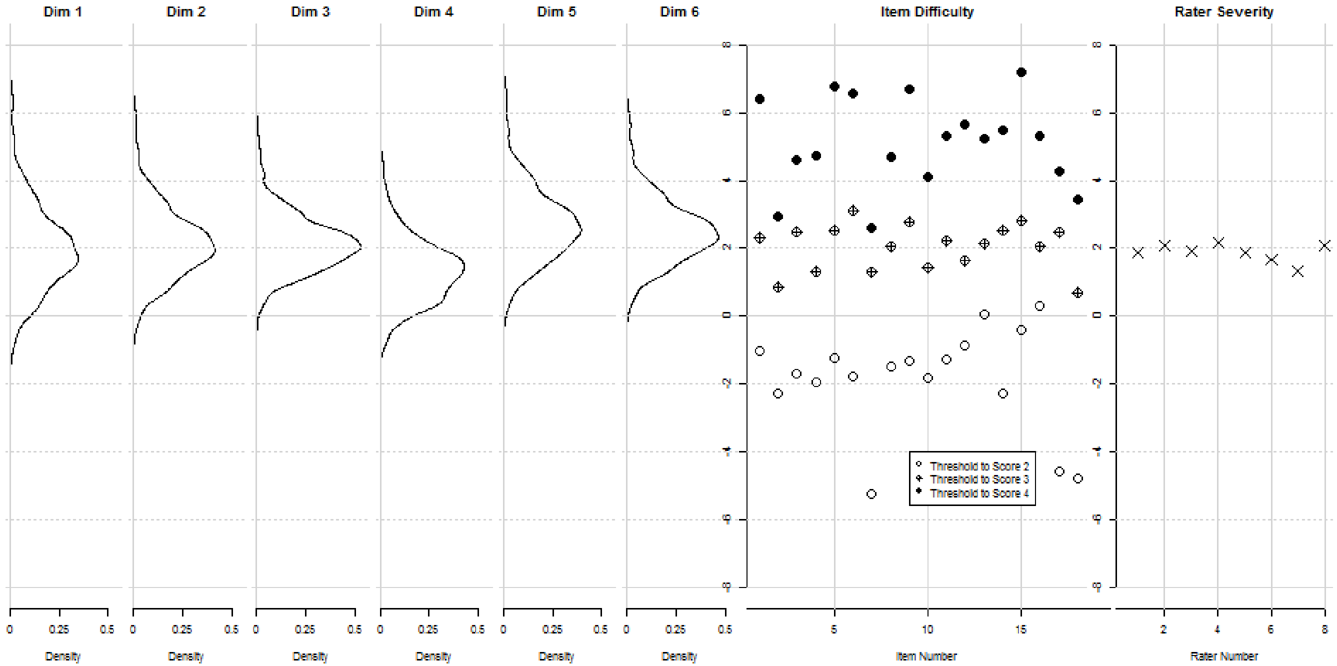

Figure 4:

Wright Map of person ability, item difficulty, and rater severity estimates

Discussion

In this study, we have suggested an approach for combining GT and IRT. We recognize that IRT models can be written in terms of a latent continuous response and that a classic IRT model can be modeled directly using a simple GT design with items as a fixed facet. The resulting logistic mixed models extend a classic IRT model by treating items as a random facet and/or considering other facets such as raters. The advantage of the proposed approach is that it allows a straightforward maximum likelihood estimation of individual random effects as well as the variance components needed for the generalizability coefficients.

In addition, application of the proposed approach was illustrated using a moderately large-scale education data set. The results demonstrated another advantage of the proposed approach: its flexibility with respect to incorporating extra complications in measurement situations (e.g., multidimensionality, polytomous responses) and explanatory variables (e.g., rater facet). The variance components and the generalizability coefficients were presented. Also, predicted individual random effects were presented by taking advantage of the usefulness of a modified Wright Map.

The results motivate further research on the following. In suggesting an approach to combine GT and IRT, the robustness of the generalizability coefficient estimates may not necessarily become a concern. However, the effects of (a) the discrete nature of data (e.g., more than two categories), (b) the violation of normality assumptions, and (c) more complex designs (e.g., person by item by rater design, multidimensionality), on the estimation accuracy of the variance components and the generalizability coefficients, should be examined and reported.

The proposed approach can be generalized to other measurement situations, both simpler and more complex ones. A simpler example is a balanced design with no missing data, or a design where the facets are nested. A more complex example is an unbalanced design with more than three crossed facets. For example, in addition to person, item, and rater facets, one could include an occasion facet that involves repeated measurement. Such attempts may offer an even closer connection between existing GT designs and IRT models. Currently, research is underway to extend the proposed approach to such alternative designs. It is in our hopes that the results from these studies will provide a more comprehensive basis to understand and evaluate methodological advantages and disadvantages of the existing and proposed approaches.

In the meantime, whether the proposed method is extendable to different designs, such as nested designs or designs with more than three facets, partly depends on the estimation methods chosen. Until recently, estimation of crossed random effects models has been limited to a relatively small number of random effects, or facets, and their levels. Even though the flexibility of the proposed approach allows a straightforward extension of the models to those situations, questions remain regarding how to estimate the variance components in the models with an increased number of crossed random facets. Moreover, incorporating other item parameters such as discrimination differences or guessing in IRT models may add more challenges in estimation and interpretation. It will be interesting to investigate what advanced and/or alternative estimation methods might be needed in extending the approaches to combine GT and IRT.

Lastly, an interesting topic for further studies exists around understanding the interaction effect between raters and persons (i.e., ratees). For example, a rater's rating of a person's response can differ systematically based on the characteristics of the response that the person gave. An interaction can also exist between the persons' group membership and the raters. For example, some raters might rate female students' responses differently than male students' responses. Or raters might differ their ratings systematically between groups of students, not knowing which group each student belongs to. Jin & Wang (2017) recently discussed a mixture facets model to account for differential rater functioning — the interaction between the ratees' unknown group membership and raters. In the proposed approach the interaction between individual raters and individual persons can be either included or not included, although it was not the focus of this study to fully explore this topic. As interactions between these facets can occur in real testing situations, it will be worthwhile to further explore how best we can model this effect.

Conclusion

The integrated modeling approach provides advantages by combining GT and IRT analyses. The logistic mixed model allows for a straightforward and effective maximum likelihood estimation of individual random effects for IRT analysis as well as the variance components needed for GT analysis. Through the Laplacian approximation implemented in the lmer() function in R, it estimates more than one cross-classified random effect efficiently with regard to the calculation time; and it estimates the variance components of incomplete data from the unbalanced mixed design relatively easily, without producing negative variance estimates. The findings from the sample data analysis showed that the proposed approach can be extended to more complicated test conditions (e.g., multidimensionality, polytomous responses, multiple raters) and produces individual estimates for persons, items, and rater random effects as well as the generalizability coefficients for person, item, and rater facets.

References

Agresti, A., & Kateri, M. (2011). Categorical data analysis. In *The International Encyclopedia of Statistical Science* (pp. 206-208). Berlin: Springer.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R.* URL http://lme4.r-forge.r-project.org/book.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint arXiv:1406.5823.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88,* 9–25.

Briggs, D. C., & Wilson, M. R. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131–155.

Chien, Y. M. (2008). *An investigation of testlet-based item response models with a random facets design in generalizability theory.* Doctoral dissertation. The University of Iowa.

Cho, S. J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis, 55*(1), 12-25.

Choi, J. (2013). *Advances in combining Generalizability Theory and Item Response Theory*. Doctoral dissertation. University of California, Berkeley: Berkeley, CA.

Choi, J., Briggs, D. C., & Wilson, M. R. (2009, April). *Multidimensional extension of the generalizability in item response modeling (GIRM)*. Paper presented at the 2009 National Council on Measurement in Education Annual Meeting, San Diego, CA.

Choi, J., Lee, Y.-S. & Draney, K. (2009). *Principle-based and process-based multidimensionality and rater effects in validation of the carbon cycle learning progression*. Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1–28.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modeling based on generalized linear models* (2nd ed.). New York, NY: Springer.

Gałecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R: A step-by-step approach*. New York, NY: Springer.

Gelman, A. (2005). Analysis of variance — why it is more important than ever. *Annals of Statistics, 33*(1), 1-53.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Iramaneerat, C., Yudkowsky, R., Myford, C., & Downing, S. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education, 13*(4), 479–493.

Jin, K. Y., & Wang, W. C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivariate Behavioral Research, 52*(3), 391–402.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis, 52*, 5066–5074.

Khuri, A. I. (2000). Designs for variance components estimation: Past and present. *International Statistical Review, 68*(3), 311-322.

Khuri, A. I., & Sahai, H. (1985). Variance components analysis: A selective literature survey. *International Statistical Review/Revue Internationale de Statistique*, 279-300.

Kolen, M., & Harris, D. (1987, April). *A multivariate test theory model based on item response theory and generalizability theory*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Linacre, J. M. (1993, April). *Generalizability theory and many-facet Rasch measurement*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *Journal of Experimental Education, 68,* 167–190.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4), 341–384.

Pearson, K. (1901). Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal, 6*(2), 566.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4*(1), 12-35.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R Development Core Team (2011). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org/.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167-190.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmarks Pedagogiske Institut.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 15-32.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Smith, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem solving skills assessment. *Educational and Psychological Measurement, 64,* 617–639.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facets Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239–261.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wright, B. D., & Stone, M. A. (1979). *Best test design*. Chicago, IL: MESA Press,

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: Generalised item response modelling software [Computer software and manual]. Camberwell, Australia: ACER Press.

Zheng, X., & Rabe-Hesketh, S. (2007). Estimating parameters of dichotomous and ordinal item response models with gllamm. *Stata Journal, 7*(3), 313-333.

Appendix

We provide generic examples of lmer() syntax for logistic mixed models with crossed random effects. In the syntax, we assume that the name of dataset is ‘data’. Syntax for simpler models have been also provided for comparisons. We recommend Gałecki & Burzykowski (2013) for details in specifying linear mixed effects models using R. Please contact authors for further assistance with model specification.

(a) Main effects for persons, items, and raters

```
R> lmer(y ~ personid + itemid + raterid, data=data,
   family=binomial)
```

(b) Main effects with random intercepts for persons, items, and raters

```
R> lmer(y ~ (1|personid) + (1|itemid) + (1|raterid),
   data=data, family=binomial)
```

(c) Crossed random effects with random intercepts for persons, items, raters and their interactions

```
R> lmer(y ~ (1|personid) + (1|itemid) + (1|raterid) +
   personid:itemid + personid:raterid + per
   sonid:itemid:raterid , data=data, family=binomial)
```