

A Bayesian approach to estimating variance components within a multivariate generalizability theory framework

Zhehan Jiang¹  · William Skorupski¹

Published online: 12 December 2017
© Psychonomic Society, Inc. 2017

Abstract In many behavioral research areas, multivariate generalizability theory (mG theory) has been typically used to investigate the reliability of certain multidimensional assessments. However, traditional mG-theory estimation—namely, using frequentist approaches—has limits, leading researchers to fail to take full advantage of the information that mG theory can offer regarding the reliability of measurements. Alternatively, Bayesian methods provide more information than frequentist approaches can offer. This article presents instructional guidelines on how to implement mG-theory analyses in a Bayesian framework; in particular, **BUGS** code is presented to fit commonly seen designs from mG theory, including single-facet designs, two-facet crossed designs, and two-facet nested designs. In addition to concrete examples that are closely related to the selected designs and the corresponding **BUGS** code, a simulated dataset is provided to demonstrate the utility and advantages of the Bayesian approach. This article is intended to serve as a tutorial reference for applied researchers and methodologists conducting mG-theory studies.

Keywords Multivariate statistic · Generalizability theory · Measurement · Bayesian inference · Markov chain Monte Carlo · Reliability

In this article, the authors present **BUGS** code to fit several commonly seen designs from multivariate generalizability

theory (*mG theory*). Prior to delving into mG theory, it is necessary to introduce univariate generalizability theory (*G theory*).

G theory has been developed for several decades since Cronbach and his associates initially introduced it (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach, & Rajaratnam, 1965). Similar to well-known classical test theory (CTT), which decomposes observed variances into several components, G theory provides a framework for separating factors underlying various measurement designs. It is primarily used to inform users the quality of a certain measurement and what decisions can be made on the basis of the analyses. To be more specific, G theory offers a comprehensive understanding of different sources of error variance within a measurement system (Brennan, 2001a). Thus, for all studies that require measurement instruments, G theory is essential to inform researchers what reliability level the studies do possess.

For illustration, assume that there are scores from an administration of a 3rd-grade writing test. It's known that the writing performance scores do not reflect test takers' true abilities with perfect reliability, as raters and/or test formats will also affect scores. Rater and format effects influence observed scores with systematic variability, which eventually impede the validity of the scores. These effects must be understood before the writing ability levels of those test takers can be estimated such that fair decisions about those scores can be made. In addition to studies on educational testing, works across other fields have been presented under the framework of G theory—for example, student evaluations of instruction by Gillmore, Kane, and Naccarato (1978), teaching behavior measurement by Shavelson and Dempsey-Atwood (1976), psychotherapy process ratings' dependability by Wasserman, Levy, and Loken (2009), clinical child and adolescent research by Lakes and Hoyt (2009), and the Big Five Personality

✉ Zhehan Jiang
zjiang4@ku.edu

¹ Department of Educational Psychology, University of Kansas, Lawrence, Kansas 66045-3101, USA

Inventory score generalizability by Arterberry, Martens, Cadigan, and Rohrer (2014).

The concept of a “true” score from CTT partitions all sources of variance as systematic or random; theoretically, systematic variance is influenced by the differences of true scores, whereas all other variability is caused by random error. In G theory, the term “universe” score is used to define the idea of a true score, after accounting for systematic and random sources of error variance. That said, instead of limiting the scores to a certain admissible setting, G theory enables users to generalize scores by partitioning out rater effect and/or occasion effect. A facet in G theory is defined by a certain set of conditions. For example, test items and test admission occasions could be regarded as two different facets, while person ability, which is a facet as well, is treated as the universe of admissible observations for a given attribute; in other words, person ability is the objects of measurement (Nußbaum, 1984). Mathematical expressions for a typical one-facet fully crossed design are:

$$X_{pi} = \mu + \theta_p + \theta_i + \epsilon \quad (1)$$

$$\sigma(X)_{pi}^2 = \sigma_p^2 + \sigma_i^2 + \sigma_e^2 \quad (2)$$

Equation 1 shows that an observed score for person p at item i is made of grand mean μ , person effect θ_p , item effect θ_i , and error effect ϵ . Correspondingly, the relevant variance components are outlined in Eq. 2. A typical two-facet fully crossed design can be expressed as:

$$X_{phi} = \mu + \theta_p + \theta_i + \theta_h + \theta_{pi} + \theta_{ih} + \theta_{ph} + \epsilon \quad (3)$$

$$\sigma(X)_{phi}^2 = \sigma_p^2 + \sigma_i^2 + \sigma_h^2 + \sigma_{pi}^2 + \sigma_{ih}^2 + \sigma_{ph}^2 + \sigma_e^2 \quad (4)$$

In addition to person effect θ_p and item effect θ_i , Eq. 3 also contains rater effect θ_h and interaction terms of any two random components. Equation 4 provides variance components in accordance with Eq. 3. Once one obtains the variance components, generalizability coefficients (*G coefficients*; a G-theory-based analogue to the CTT reliability coefficient) may be calculated.

Estimating variance components as well as G coefficients is typically called a generalizability (G) study, whereas sequential research dealing with the practical application of a measurement procedure is named a decision (D) study; that is, a D study helps researchers use the results of a G study to design a measurement procedure that minimizes error at selected facet(s). Sound D studies can provide practitioners useful information that involves logistical and cost considerations as well as the change of G coefficients. For example, a D study could inform one that increasing the number of items from, say, 10 to 20, on the basis of findings from the corresponding G study, would increase the G coefficient from .8 to .9 (how much the G coefficient would increase would depend directly on the size of the estimated variance components and the number of facets in the study).

As the name indicates, mG theory is an extension of univariate G theory. Similarly, the G coefficient can be extended to the multivariate case (Joe & Woodward, 1976; Marcoulides, 1995; Woodward & Joe, 1973). Statistically, mG theory assumes that scores from the universe of levels of the fixed multivariate facets are not orthogonal, meaning that there are correlations within the facets. In contrast to univariate G theory, in which universe scores and error scores are estimated via variance components, analyzing measurement procedures via a multivariate approach “can provide information about facets that contribute to covariations among the multiple scores.” (Raykov & Marcoulides, 2006, p. 213). A simple example is that of the Medical College Admission Test (MCAT), which is composed of four subtests: biological sciences (biology MCAT), physical sciences (physical MCAT), verbal reasoning (verbal MCAT), and a writing sample (written MCAT). The subtest designs of MCAT lend themselves naturally to mG theory, in which the proficiencies measured by the various dimensions are presumed to be correlated. Thus, in addition to variance components that are estimated in univariate G theory, the multivariate version also accounts for covariance components among subtests. One could theoretically conduct four separate univariate G-theory analyses for these subtests, however, that assumes zero correlation between the subtests. Marcoulides (1990) points out that such a flawed assumption can potentially influence the accuracy of the estimated variance components.

Overall, mG theory for MCAT and other measures of this kind provides information at the **profile** level. The relation between univariate G theory and mG theory are akin to that of between the analysis of variance (ANOVA) and the multivariate analysis of variance (MANOVA): mG theory gives reliability of profile scores, not the reliability of individual variables. If the reliability of individual variables were also of interest, one could run a mG-theory analysis and follow it with individual univariate G-theory analyses, much like one might follow a MANOVA with separate ANOVAs for each dependent variable. That said, for instance, if physical MCAT is defined to be the attribute of interest, one can use univariate G theory to conduct further studies.

To summarize, when research design and/or data nature are multidimensionally structured, mG theory can be implemented to (1) calculate the reliability of difference scores, observable correlations, or universe score and error correlations (Brennan, 2001a), (2) obtain score profile reliability estimates (Brennan, 2001a), and (3) estimate composite scores with highest generalizability level (Shavelson & Webb, 1981). Most of the published works using mG theory demonstrate its application to testing structure. For example, Nußbaum (1984) implemented mG theory to study students’ ability to paint in watercolors; Wu and Tzou (2015) use mG theory to analyze standard setting data; Clauser and colleagues have published several mG-theory studies using National Board

of Medical Examiners' data (Clauser, Harik, & Margolis, 2006; Clauser, Margolis, & Swanson, 2002; Clauser, Swanson, & Harik, 2002). Details about relevant designs and mathematical expressions of mG theory are provided in following sections. Further information about theoretical concerns with subtest dimensionality, for example, the level of correlation between subtests for supporting the application of mG theory, can be found in the subscore literature (see Puhan, Sinharay, Haberman, & Larkin, 2008; Sinharay, 2010, 2013; Sinharay & Haberman, 2014, for details). Researchers can also use techniques provided in the present article to conduct simulation studies for addressing more in-depth subscore inquiries.

Multivariate generalizability theory designs

This article illustrates three genetic categories of mG-theory designs: (1) single-facet, (2) two-facet crossed, and (3) two-facet nested. Each design contains two subdesigns, shown in Table 1. These designs were selected because of their wide applicability in practice. All six subdesigns are elaborated by Brennan (2001a). The notations and symbols here match Brennan's delineation to label the random effects: “*p*” is the person effect, “*i*” is item effect, and “*h*” is the rater effect. The superscript filled circle designates that the effects are the same across dimensions, and the superscript empty circle designates that the effects are different across dimensions. Additional explanation for each design and corresponding **BUGS** code are provided in later sections.

Estimation

Although G-theory analyses have been conducted in different statistical software platforms such as **R** (R Development Core Team, 2017), **SAS** (SAS Institute, 1990), and **SPSS** (Norusis, 1990), current estimation practice for conducting mG-theory analyses primarily relies on **mGENOVA** software (Brennan, 2001b). The computation of **mGENOVA** is based on mean square error, which is an extension of a traditional MANOVA approach. Although Brennan (2001a) provides solutions to handling missing data, sparse matrices, and other problems

Table 1 Six commonly seen multivariate generalizability theory designs

Single-facet designs	$p^{\bullet \times} i^{\circ}$ $p^{\bullet \times} i^{\bullet}$
Two-facet crossed designs	$p^{\bullet \times} i^{\bullet \times} h^{\bullet}$ $p^{\bullet \times} i^{\bullet \times} h^{\circ}$
Two-facet nested designs	$p^{\bullet \times} (i^{\bullet} h^{\bullet})$ $p^{\bullet \times} (i^{\circ} h^{\circ})$

caused by violating model assumptions, a large number of these solutions are not available in **mGENOVA**. One of the most important criticisms of the traditional estimation approach is that an assessment of model-data fit is not provided (Gessaroli, 2003). If a proposed model—that is, mG theory in the present context—fits the data poorly, further analyses based on the estimation are not trustworthy. In addition, traditional MANOVA and mG theory do not provide standard errors for the estimation of variance and covariance components, potentially masking the precision of the estimates.

Using a Bayesian approach, however, can minimize the aforementioned problems. In this article we demonstrate a Bayesian approach to mG theory using Markov chain Monte Carlo (MCMC) techniques to estimate the parameters of interest. MCMC techniques are able to deal with complex models and sparse data matrices that many traditional methods cannot (see Muthén & Asparouhov, 2012, for example). Furthermore, users are allowed to incorporate prior information to make estimations more robust (Lynch, 2007, p. 2). For example, incorporating prior distributions of variance components allows users to restrict the estimates to permissible numeric space, and therefore avoid such issues as Heywood cases (negative variance estimates). If prior information is unavailable, *uninformative* priors may be specified; posterior estimation is then essentially equivalent to traditional estimation (Jeffreys, 1946). Note that *uninformative* prior distributions are used throughout the article. Another advantage of a Bayesian framework is that missing-data imputation can be accommodated simultaneously. The MCMC samples can be applied to missing values as part of the algorithm (Little & Rubin, 2014). Literature has shown that Bayesian methods are an effective way to impute missing values, and the process is consistently stable (Enders, 2010). Lastly, similar to confidence intervals from frequentist approaches, Bayesian methods provide credible intervals for each posterior estimate so that users can investigate the probable range of the parameters with a specified probability level (e.g., 95%). This feature circumvents the asymptotic assumptions, which may not be always reasonable (Jackman, 2009).

The following sections provide additional details on how to implement a Bayesian version of mG theory using **BUGS** software. In particular, the content (1) covers how to specify **BUGS** code, (2) provides specific information about popular mG-theory designs, and (3) sheds lights on the further applications of the proposed approach.

The BUGS language and software

The use of **BUGS** software to perform analytical tasks has grown in popularity, as it is straightforward to alter existing code to fit new variations of current models. For example, Curtis (2010) illustrates a family of item response theory (IRT) models that can be fit by modifying existing **BUGS**

code from simpler models. The **BUGS** software possesses multiple MCMC sampling methods customizable to different needs, including derivative-free adaptive rejection sampling (Gilks, 1992), Slice sampling (Neal, 1997), current-point Metropolis, and direct sampling using standard algorithms (Lunn, Spiegelhalter, Thomas, & Best, 2009). As a result, a wide range of statistical models can be appropriately estimated, including mG theory in the present article.

Three statistical packages are frequently used for running **BUGS** syntax: **WinBUGS** (Lunn, Thomas, Best, & Spiegelhalter, 2000; Spiegelhalter, Thomas, Best, & Lunn, 2003), **OpenBUGS** (Thomas, O’Hara, Ligges, & Sturtz, 2006), and **JAGS** (Plummer, 2003, 2010). Details about each of these software packages can be found from their websites. Although some differences exist among these options, they frequently produce similar results when the identical tasks are given. **OpenBUGS** is implemented in the present article, as it has been actively developed over the past decade.

BUGS code

We begin with an example of a single-facet design, in which the construction of corresponding statistical model is illustrated. In addition, we describe how input data are structured and loaded. Throughout the instructional guidelines, the same number of observations N_p and number of dimensions, V , are used in every design, whereas the levels of random effects, N_i and N_h , vary across designs.

$p^* \times i^0$ design

The first single-facet design is presented in Fig. 1: 100 persons take a test in which there are 24 items in total. The first subtest, or dimension $v1$, is composed of items $i1$ to $i11$, and the second subtest contains item $i12$ to $i24$. Mathematically, the design can be expressed as:

$$\begin{pmatrix} X_{1pi} \\ X_{2pi} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \theta_{1p} \\ \theta_{2p} \end{pmatrix} + \begin{pmatrix} \theta_{1i} \\ \theta_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \tag{5}$$

X represent observed data points: the numeric subscript indicates dimension v , where p and i again represent person and item, respectively. Equation 5 indicates that an observed score for an item i on subtest v is a function of the grand mean,

μ , the person effect, θ_p , the item effect, θ_i , and the residual effect, ϵ . The corresponding variances and covariances from Eq. 6 can be decomposed as:

$$\begin{bmatrix} \sigma_{1X_{pi}}^2 & \sigma_{21X_{pi}} \\ \sigma_{12X_{pi}} & \sigma_{2X_{pi}}^2 \end{bmatrix} = \begin{bmatrix} \sigma_{1p}^2 & \sigma_{1p2p} \\ \sigma_{1p2p} & \sigma_{2p}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1i}^2 & 0 \\ 0 & \sigma_{2i}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1e}^2 & 0 \\ 0 & \sigma_{2e}^2 \end{bmatrix}. \tag{6}$$

The left side on the equation is the observed covariance matrix, whereas the matrices on the right are covariances for person, p , item, i , and residual effects, e , respectively. Note that the off-diagonals for the item effects i and the residual effect e are zeros. This assumption matches the design, in that different subtests possess different items; there is no way to estimate covariance since the items are not crossed.

Table 2 contains **BUGS** code to fit the $p^* \times i^0$ design. The following list contains several comments to help clarify the code. A similar relevant list will follow each example below.

- When data are read directly from **BUGS**, they need to be specified as a matrix by using *structure* function so that the data can be loaded as seen in Fig. 1. Note that **BUGS** places data into matrices or arrays by row, which is different from R environment by default that places data into matrices by column (R Development Core Team, 2017).
- Certain observed variables need to be manually created in addition to the data. $Np = 100$ is the number of persons, or examinees in the current example; $Ni = 24$ is the total number of items; $V = 2$ is the number of subtests; $Ncol = 24$ is the number of columns of the input data, which in this case is equivalent to Ni ; *zero.vector* = $c(0,0)$ is used to fulfill the means of the random effects that are multivariate normally distributed; *grand.mu* = $c(50,60)$ is the grand means that users calculated by averaging data points of each subtest; *R* = *diagonal(1)* is used to specifying an unstructured uninformative prior distributions for the precision (inverse of covariance matrix), which is assumed to follow a Wishart distribution (Gelman, 2006).
- Other variables are created for model formatting so that the loaded data can be mapped to corresponding random effects. *i.format* = $c(1, 2, \dots, 24)$ is simply an identification vector for items; *iv.format* = $c(1, 1)$ is a vector to show the subtest to which the items belong.

	v1											v2													
p	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24	
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
...
100	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Fig. 1 Typical example of a $p^* \times i^0$ design

Table 2 Code for a $p^* \times i^0$ design model

```

model{
  for ( p in 1: Np ) {
    for ( col in 1: Ncol ){
      X [ p , col ] ~ dnorm ( mu [ p , col ], TauE [ iv.format [ col ] ] )
    }
  }
  #Random Effect Components
  for ( p in 1: Np ) {
    for ( col in 1: Ncol ){
      mu[p, col] <- grand.mu [ iv.format[ col ] ] + Peffect [ p,
iv.format [ col ] ] + Ieffect [ i.format [ col ] , iv.format [ col ] ]
    }
  }

  for ( p in 1 : Np ) {
    Peffect [ p , 1: V ] ~ dnorm ( zero.vector[ ], OmegaP[,] )
  }

  for ( i in 1: Ni ) {
    Ieffect [ i , iv.format[i] ] ~ dnorm ( 0 , iv.format[i] )
  }

  #Priors

  for ( v in 1: V ) {
    TauE [v] ~ dgamma (0.01, 0.01)
    Taul [v] ~ dgamma (0.01, 0.01)
    VarE[v]<-pow(TauE [v],-1)
    VarI[v]<-pow(Taul [v],-1)
  }

  OmegaP [1: V, 1:V] ~dwish(R[,],V)
  SigmaP[ 1:V, 1:V]<-inverse(OmegaP[,])
}

```

 $p^* \times i^0$ design

The second single-facet design is presented in Fig. 2: 100 persons take a test comprised of 12 items in total. These items are graded on the basis of two criteria (i.e., two dimensions). The design is mathematically identical to Eq. 5 in $p^* \times i^0$ design. However, in the present example, the variance and covariance matrices are different from Eq. 6:

$$\begin{bmatrix} \sigma_{1X_{pi}}^2 & \sigma_{21X_{pi}} \\ \sigma_{12X_{pi}} & \sigma_{2X_{pi}}^2 \end{bmatrix} = \begin{bmatrix} \sigma_{1p}^2 & \sigma_{1p2p} \\ \sigma_{1p2p} & \sigma_{2p}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1i}^2 & \sigma_{1i2i} \\ \sigma_{1i2i}^2 & \sigma_{2i}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1e}^2 & \sigma_{1e2e} \\ \sigma_{1e2e}^2 & \sigma_{2e}^2 \end{bmatrix}. \quad (7)$$

p	v1												v2											
	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
...
100	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Fig. 2 Typical example of a $p \times i$ design

The left side on the equation is the observed covariance matrix, whereas the matrices on the right side are covariance for person effects, p , item effects, i , and residual effects, e , respectively. Note that for each matrix both off-diagonal and diagonal elements are not fixed to zero. This setting matches the nature of a fully crossed design: all the effects are present in each dimension v .

Table 3 contains **BUGS** code to fit the $p \times i$ design. Similar to the first design, below are several comments helping clarify the code.

- Unlike in the $p \times i^0$ design, when data are read directly from **BUGS**, they need to be specified as an array instead of a matrix. Follow the order of $N_p \times N_i \times V$ array, the current data are formatted by commending: `structure(Data = c(X1p1i1, X1p1i2, ..., X1p100i12, X2p1i1, X2p1i2, ..., X2p100i12), Dim = c(100, 12, 2))`. The subscripts of x indicate the identification of criterion (or dimension), person, and item, respectively. It means that **BUGS** forms an array by loading $v1$ data matrix and $v2$ data matrix in sequential order.
- As we have shown previously, users need to manually create some of the observed variables in addition to the data. Np , V , `zero.vector`, `grand.mu`, and R are identical to those in $p \times i^0$ design, except $Ni = 12$ in the present example.
- There is no need to specify `iv.format` as the data loading process over an array already navigates the identification for each dimension. Meanwhile a new identification vector for items needs to be specified accordingly: `i.format =`

$c(1, 2, \dots, 12)$, since the present example only contains 12 items in total instead of 24.

$p \times i \times h$ design

The first two-facet design is presented in Fig. 3: 100 persons take a test that includes four items in total. These items are graded by three raters on the basis of two criteria (or dimensions). Each rater and each criterion are labeled as h and v , respectively. Mathematically, the design can be expressed as:

$$\begin{pmatrix} X_{1pih} \\ X_{2pih} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \theta_{1p} \\ \theta_{2p} \end{pmatrix} + \begin{pmatrix} \theta_{1i} \\ \theta_{2i} \end{pmatrix} + \begin{pmatrix} \theta_{1h} \\ \theta_{2h} \end{pmatrix} + \begin{pmatrix} \theta_{1pi} \\ \theta_{2pi} \end{pmatrix} + \begin{pmatrix} \theta_{1ph} \\ \theta_{2ph} \end{pmatrix} + \begin{pmatrix} \theta_{1ih} \\ \theta_{2ih} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \tag{8}$$

The notations and symbols in Eq. 8 are similar to the previous examples, but with more random effects than in the first two designs. X represents observed data points: the numeric subscript indicates dimension v , where p , i , and h represent the person, item, and rater, respectively. Equation 8 indicates that an observed score of an item i graded by rater h on the basis of criterion v is a function of the grand mean, μ , the person effect, θ_p , the item effect, θ_i , the rater effect, θ_h , the interaction effect between persons and items, θ_{pi} , the interaction effect between persons and raters, θ_{ph} , the interaction effect between items and raters, θ_{ih} , and finally the residual effect, ϵ . Correspondingly, the variance and covariance from Eq. 8 can be decomposed as:

$$\begin{bmatrix} \sigma_1^2 X_{pi} & \sigma_{21} X_{pi} \\ \sigma_{12} X_{pi} & \sigma_2^2 X_{pi} \end{bmatrix} = \begin{bmatrix} \sigma_{1p}^2 & \sigma_{1p2p} \\ \sigma_{1p2p} & \sigma_{2p}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1i}^2 & \sigma_{1i2i} \\ \sigma_{1i2i} & \sigma_{2i}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1h}^2 & \sigma_{1h2h} \\ \sigma_{1h2h} & \sigma_{2h}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1pi}^2 & \sigma_{1pi2pi} \\ \sigma_{1pi2pi} & \sigma_{2pi}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1ph}^2 & \sigma_{1ph2ph} \\ \sigma_{1ph2ph} & \sigma_{2ph}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1ih}^2 & \sigma_{1ih2ih} \\ \sigma_{1ih2ih} & \sigma_{2ih}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1e}^2 & \sigma_{1e2e} \\ \sigma_{1e2e} & \sigma_{2e}^2 \end{bmatrix}. \tag{9}$$

Since the $p \times i \times h$ design is fully crossed, each element of all matrices in Eq. 9 needs to be estimated as an extension of the $p \times i$ design. Note that the total the number of the estimates is 21 in the present

example: Generally, models with a large number of parameters require a larger sample size, but a Bayesian framework can use informative prior distributions to circumvent that challenging requirement (Dunson, 2001).

Table 3 Code for a $p^* \times i^*$ design model

```

model {
  for ( i in 1 : Ni ) {
    for ( p in 1:Np ){
      X [ p , i , 1 : V ] ~ dnorm ( mu [ p , i , 1 : V ] , OmegaE [ , ] )
    }
  }
  #Random Effect Components
  for ( i in 1 : Ni ) {
    for ( p in 1 : Np ) {
      for ( v in 1 : V ) {
        mu [ p , i , v ] <- grand.mu [ v ] + Peffect [ p , v ] +
Ieffect [ i.format [ i ] , v ]
      }
    }
  }

  for ( p in 1 : Np ) {
    Peffect [ p , 1 : V ] ~ dnorm ( zero.vector [ ] , OmegaP [ , ] )
  }

  for ( i in 1 : Ni ) {
    Ieffect [ i , 1 : V ] ~ dnorm ( zero.vector [ ] , OmegaI [ , ] )
  }

  # Priors for All the Random Effects

  OmegaE [ 1 : V , 1 : V ] ~ dwish ( R [ , ] , V )
  SigmaE [ 1 : V , 1 : V ] <- inverse(OmegaE [ , ] )

  OmegaP [ 1 : V , 1 : V ] ~ dwish ( R [ , ] , V )
  SigmaP [ 1 : V , 1 : V ] <- inverse(OmegaP [ , ] )

  OmegaI [ 1 : V , 1 : V ] ~ dwish ( R [ , ] , V )
  SigmaI [ 1 : V , 1 : V ] <- inverse(OmegaI [ , ] )
}

```

	v1												v2											
	h1				h2				h3				h1				h2				h3			
p	i1	i2	i3	i4																				
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
...
100	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Fig. 3 Typical example of a $p^* \times i^* \times h^*$ design

Table 4 contains the **BUGS** code to fit the $p \times i \times h$ design. Below are several comments to help clarify the code.

- Similar to $p \times i$ design, when data are read directly from **BUGS**, they need to be specified as an array instead of a matrix. Again **BUGS** fills data matrix by matrix and thus in the example $v1$ data points and $v2$ data points are loaded in order.
- Np , V , $zero.vector$, $grand.mu$, and R are identical to what were specified in $p \times i$ design. $Ni = 4$ and $Nh = 3$ indicate that there are four items and three raters. $Ncol = 24$ is the number of columns of the input data, but in this case it is not equivalent to Ni anymore.
- The formatting variables are straightforward; each matches the corresponding header from Fig. 3. $i.format = c(1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4)$ and $h.format = c(1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3)$. Similar to $p \times i$ design, the formatting variables do not have to go through both dimensions as the formats of $v1$ and $v2$ are identical.

Table 4 Code for a $p \times i \times h$ design model

```

model {
  for ( p in 1: Np) {
    for ( col in 1: Ncol) {
      X [ p, col, 1: V ] ~ dnorm ( mu [ p , col , 1:V ] , OmegaE [, ] )
    }
  }
  #Random Effect Components
  for ( p in 1: Np) {
    for ( col in 1: Ncol) {
      for ( v in 1:V) {
        mu [ p, col , v ] <- grand.mu [v] + Peffect [p,v] + Ieffect
[ i.format[col] , v ] + Heffect[ h.format[col], v ] + Pieffect [ p, i.format[col] , v ]
+ PHeffect [ p, h.format[col], v ] + IHeffect[ i.format[col],
h.format[col], v]
      }
    }
  }

  for ( p in 1: Np){
    Peffect [ p, 1:V ] ~ dnorm (zero.vector[], OmegaP[,])
  }

  for ( i in 1:Ni) {
    Ieffect [ i, 1:V ] ~ dnorm (zero.vector [], OmegaI[,])
  }

  for ( h in 1:Nh){
    Heffect[ h, 1:V ] ~ dnorm(zero.vector [], OmegaH[,])
  }

  for (p in 1:Np){
    for ( i in 1:Ni){
      Pieffect[ p, i, 1:V ] ~ dnorm (zero.vector[], OmegaPI[,])
    }
  }
}

```

Table 4 (continued)

```

    }
  }

  for (p in 1:Np){
    for (h in 1:Nh){
      PHeffect[p, h, 1:V] ~dmnorm (zero.vector[], OmegaPH[,])
    }
  }

  for (i in 1:Ni){
    for (h in 1:Nh){
      IHeffect[i, h, 1:V] ~dmnorm (zero.vector[], OmegaIH[,])
    }
  }

  OmegaP [1: V, 1:V] ~dwish(R[,],V)
  SigmaP [1:V, 1:V]<-inverse(OmegaP[,])
  OmegaI [1: V, 1:V] ~dwish(R[,],V)
  SigmaI [1:V, 1:V]<-inverse(OmegaI[,])
  OmegaH [1: V, 1:V] ~dwish(R[,],V)
  SigmaH [1:V, 1:V]<-inverse(OmegaH[,])
  OmegaPI [1: V, 1:V] ~dwish(R[,],V)
  SigmaPI [1:V, 1:V]<-inverse(OmegaPI[,])
  OmegaPH [1: V, 1:V] ~dwish(R[,],V)
  SigmaPH [1:V, 1:V]<-inverse(OmegaPH[,])
  OmegaIH [1: V, 1:V] ~dwish(R[,],V)
  SigmaIH [1:V, 1:V]<-inverse(OmegaIH[,])
  OmegaE [1: V, 1:V] ~dwish(R[,],V)
  SigmaE [1:V, 1:V]<-inverse(OmegaE[,])
}

```

$p \times i \times h^0$ design

The second two-facet design is presented in Fig. 4: 100 persons take a test with three items in total. These items are

graded by eight raters on the basis of two different criteria (or dimensions): The first three raters focus on the first criterion, whereas the remaining five raters follow the second. Each rater and each criterion are labeled as h and v ,

p	$v1$												$v2$														
	$h1$			$h2$			$h3$			$h4$			$h5$			$h6$			$h7$			$h8$					
	$i1$	$i2$	$i3$																								
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
...
100	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Fig. 4 Typical example of a $p \times i \times h^0$ design

respectively. Mathematically, the design is identical to Eq. 8 (the $p \bullet \times i \bullet \times h \bullet$ design). However, in the present example, the

variance and covariance matrices are different from those in Eq. 9:

$$\begin{bmatrix} \sigma_1^2 X_{pi} & \sigma_{21} X_{pi} \\ \sigma_{12} X_{pi} & \sigma_2^2 X_{pi} \end{bmatrix} = \begin{bmatrix} \sigma_{1p}^2 & \sigma_{1p2p} \\ \sigma_{1p2p} & \sigma_{2p}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1i}^2 & \sigma_{1i2i}^2 \\ \sigma_{1i2i}^2 & \sigma_{2i}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1h}^2 & 0 \\ 0 & \sigma_{2h}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{2pi}^2 & \sigma_{1pi2pi} \\ \sigma_{1pi2pi} & \sigma_{2pi}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1ph}^2 & 0 \\ 0 & \sigma_{2ph}^2 \end{bmatrix} \\ + \begin{bmatrix} \sigma_{1ih}^2 & 0 \\ 0 & \sigma_{2ih}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1e}^2 & 0 \\ 0 & \sigma_{2e}^2 \end{bmatrix}. \tag{10}$$

In Eq. 10, several matrices' off-diagonal elements are fixed at zero; this pattern, again, matches the nature of a non-fully-crossed design. A straightforward way to view this equation, in line with Fig. 4, is that the rater effects h are not defined identically across grading criteria v . Thus, matrices related to h do not have covariance components including the residual effect.

Table 5 contains **BUGS** code to fit the $p \bullet \times i \bullet \times h^0$ design. Comments are presented again to explain the code; they are essentially an extension of those for the $p \bullet \times i^0$ design.

- The data are entered as a matrix, as Fig. 4 demonstrates.
- $Np, V, zero.vector, grand.mu, Ncol,$ and R are identical to those in the previous designs, whereas some new and different variables are required in the present example: Ni is now 3, since there are three items only; $Nh.v1 = 3$ and $Nh.v2 = 5$ represent the numbers of raters within each dimension v .
- There are more formatting variables than in previous designs, due to the complexity of the model: $i.format = c(1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3)$ is the identification vector of items matching the data columns; $iv.format = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2)$ indicates that the first nine columns of the input matrix belong to dimension $v1$ and the rest 15 columns are connected to dimension $v2$; $ih.format = c(1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8)$ contains the structure for items i and rater h effects, for example, item responses from the first three columns of the input data are associated with Rater 1 ($h1$); $hv.format = c(1, 1, 1, 2, 2, 2, 2, 2)$ connects the dimensions v for rater effects h .

$p \bullet \times (i^0: h \bullet)$ design

The first two-facet nested design is presented in Fig. 5: 100 persons take a test containing 24 items in total. The first subtest, or dimension $v1$, is composed of items $i1$ to $i12$, and the second subtest contains items $i13$ to $i24$. Across the two subtests $v1$ and $v2$, the same three raters grade the item responses; for example,

the first rater ($h1$) grades items $i1$ to $i3$ belonging to subtest $v1$, as well as items $i13$ to $i15$ that are under subtest $v2$. Mathematically, the design can be expressed as:

$$\begin{pmatrix} X_{1pih} \\ X_{2pih} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \theta_{1p} \\ \theta_{2p} \end{pmatrix} + \begin{pmatrix} \theta_{1h} \\ \theta_{2h} \end{pmatrix} + \begin{pmatrix} \theta_{1i:h} \\ \theta_{2i:h} \end{pmatrix} \\ + \begin{pmatrix} \theta_{1ph} \\ \theta_{2ph} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \tag{11}$$

Note that in Eq. 11, item effects no longer exist, since items are now nested within raters. Correspondingly, $\theta_{1i:h}$ and $\theta_{2i:h}$ represent these nested random effects. The variance and covariance matrices can be formatted as:

$$\begin{bmatrix} \sigma_1^2 X_{pi} & \sigma_{21} X_{pi} \\ \sigma_{12} X_{pi} & \sigma_2^2 X_{pi} \end{bmatrix} = \begin{bmatrix} \sigma_{1p}^2 & \sigma_{1p2p} \\ \sigma_{1p2p} & \sigma_{2p}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1h}^2 & 0 \\ 0 & \sigma_{2h}^2 \end{bmatrix} \\ + \begin{bmatrix} \sigma_{1i:h}^2 & 0 \\ 0 & \sigma_{2i:h}^2 \end{bmatrix} \\ + \begin{bmatrix} \sigma_{1ph}^2 & \sigma_{1ph2ph}^2 \\ \sigma_{1ph2ph}^2 & \sigma_{2ph}^2 \end{bmatrix} \\ + \begin{bmatrix} \sigma_{1e}^2 & 0 \\ 0 & \sigma_{2e}^2 \end{bmatrix}. \tag{12}$$

As one can see, as compared with the aforementioned crossed designs, the nested design has fewer parameters to be estimated. It can be considered that (1) item effects and the interaction effects between item i and rater h are bundled, and (2) the interaction effects between person p and item i are now entangled with residual effects.

Table 6 contains **BUGS** code for fitting the $p \bullet \times (i^0: h \bullet)$ design. As with the previous designs, comments are outlined below.

- The data are entered as a matrix, as Fig. 5 shows.
- $Np, V, zero.vector, grand.mu, Ncol,$ and R are identical to those variables in the previous designs. The only exception is that $Nh = 3$ due to the fact that the present example has three raters in total.

Table 5 (continued)

```

    for ( ncol in 1 :Ncol ){
        IHeffect[ ncol] ~ dnorm (0, TauIH[iv.format[ncol]])
    }

#Priors

    for ( v in 1:V) {
        TauE[v]~dgamma(0.01,0.01)
        VarE[v]<-pow(TauE[v],-1)
    }

    OmegaP [1: V, 1:V] ~dwish(R[,],V)
    SigmaP[ 1:V, 1:V]<-inverse(OmegaP[,])

    Omegal [1: V, 1:V] ~dwish( R [ , ],V)
    Sigmal[ 1:V, 1:V]<-inverse(Omegal[,])

    for ( v in 1:V) {
        TauH[v]~dgamma(0.01,0.1)
        VarH[v]<-pow(TauH[v] ,-1)
    }

    OmegaPI [1: V, 1:V] ~dwish(R[,],V)
    SigmaPI[ 1:V, 1:V]<-inverse(OmegaPI[,])

    for ( v in 1:V) {
        TauPH[v]~dgamma(0.01,0.01)
        VarPH[v]<-pow(TauPH[v] ,-1)
    }

    for ( v in 1:V) {
        TauIH[v]~dgamma(0.01,0.01)
        VarIH[v]<-pow(TauIH[v],-1)
    }

}

```

 $p^{\bullet} \times (i^{\circ}: h^{\circ})$ design

The last example discussed in this section is the two-facet nested design, shown in Fig. 6. The scenario is similar to the one for the $p^{\bullet} \times (i^{\circ}: h^{\bullet})$ design, except that five raters are nested within different the subtests. That is, $p^{\bullet} \times (i^{\circ}: h^{\circ})$ has

two nesting layers: (1) items are nested in raters, and (2) raters are nested in subtests. For example, Rater 1 and Rater 2 are assigned to grade items belonging to the first subtest, $v1$, whereas Raters 3–5 are assigned to grade items belonging to the second subtest, $v2$. Under the raters level, rater $h1$ grades items $i1$ to $i3$, rater $h2$ grades items $i6$ to $i12$, rater $h3$ grades

p	v1												v2											
	h1			h2			h3			h1			h2			h3								
	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
...
100	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Fig. 5 Typical example of a $p^* \times (i^0, h^*)$ design

items $i13$ to $i15$, rater $h4$ grades items $i16$ to $i19$, and rater $h5$ grades the rest of the items. The mathematical expression is

identical to that in Eq. 11, but the variance and covariance components are different:

$$\begin{bmatrix} \sigma_{1^2}^2 X_{pi} & \sigma_{21} X_{pi} \\ \sigma_{12} X_{pi} & \sigma_{2^2}^2 X_{pi} \end{bmatrix} = \begin{bmatrix} \sigma_{1p}^2 & \sigma_{1p2p} \\ \sigma_{1p2p} & \sigma_{2p}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1h}^2 & 0 \\ 0 & \sigma_{2h}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1i:h}^2 & 0 \\ 0 & \sigma_{2i:h}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1ph}^2 & 0 \\ 0 & \sigma_{2ph}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{1e}^2 & 0 \\ 0 & \sigma_{2e}^2 \end{bmatrix}. \tag{11}$$

Table 6 Code for a $p^* \times (i^0, h^*)$ design model

```

model{
  for ( p in 1:Np) {
    for (col in 1:Ncol) {
      X[p, col] ~ dnorm (mu[p,col], TauE[iv.format[col]] )
    }
  }

  #Random Effect Components
  for (p in 1:Np) {
    for (col in 1:Ncol) {
      mu[p, col] <-grand.mu[iv.format[col]] + Peffect[p,
iv.format[col]] + Heffect[ih.format[col], iv.format[col]] +
IHeffect[i.format[col],ih.format[col],iv.format[col]] + PHeffect[p,
ih.format[col], iv.format[col]]
    }
  }

  for ( p in 1:Np) {
    Peffect [ p, 1:V] ~dmnorm( zero.vector[], OmegaP[, ] )
  }

  for ( h in 1:Nh){
    Heffect[ h, 1:V] ~dmnorm(zero.vector [], OmegaH[, ])
  }

  for ( col in 1:Ncol){
    for (v in 1:V){
      IHeffect[i.format[col],ih.format[col],v] ~dnorm (0, TauIH[v])
    }
  }
}

```

Table 6 (continued)

```

    }
  }

  for (p in 1:Np){
    for (h in 1:Nh){
      PHeffect[p, h, 1:V] ~dmnorm (zero.vector[], OmegaPH[,])
    }
  }

#Priors
  for (v in 1:V) {
    TauE[v]~dgamma(0.01,0.01)
    VarE[v]<-pow(TauE[v],-1)
  }

  OmegaP [1: V, 1:V] ~dwish(R[,],V)
  SigmaP[ 1:V, 1:V]<-inverse(OmegaP[,])

  OmegaH [1: V, 1:V] ~dwish(R[,],V)
  SigmaH[ 1:V, 1:V]<-inverse(OmegaH[,])

  for (v in 1:V) {
    TauIH[v]~dgamma(0.01,0.01)
    VarIH[v]<-pow(TauIH[v],-1)
  }

  OmegaPH [1: V, 1:V] ~dwish(R[,],V)
  SigmaPH[ 1:V, 1:V]<-inverse(OmegaPH[,])
}

```

From Eq. 13, one can tell that only a full matrix (person effects) is estimated. Thus, among all the two-facet designs presented in this article, the present example has the fewest parameter to estimate.

Table 7 contains **BUGS** code for fitting the $p \times (i^0: h^0)$ design. Comments about the code are similar to those for the $p \times (i^0: h^*)$ design, except the parts that reflect that the raters are now nested in substests.

- The data are entered as a matrix, as Fig. 6 shows.
- Np , V , $zero.vector$, $grand.mu$, $Ncol$, and R are identical to those variables in the previous designs. The only exception is $Nh = 5$, since the present example has three raters in total.
- As we mentioned earlier, $i.format = c(1, 2, \dots, 24)$ is not necessary but still specified in order to maintain

p	v1												v2											
	h1					h2					h3			h4			h5							
	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
...
100	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Fig. 6 Typical example of a $p \times (i^0: h^0)$ design

consistency. $iv.format = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2)$ is a vector to show to which subtest an item belongs: Items $i1$ to $i15$ are under subtest $v1$ and the rest items are nested within subtest $v2$. $ih.format = c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5)$ indicates the relations between items and raters, for example, items $i20$ to $i24$ are graded by rater $h5$. Finally, $hv.format = c(1, 1, 1, 2, 2, 2, 2, 2)$ represents the nested structure of raters and subtests.

Example

In this section, the **BUGS** code for the $p \times i$ design is used to analyze a simulated dataset. This is achieved for the sake of

the tutorial, rather than for providing a comprehensive simulation study with multiple replications. Here, a single data set was generated for demonstration. To be specific, the dataset follows a one-facet fully crossed design whose number of fixed effects is 3. Equation 14 provides the data generation process:

$$\begin{pmatrix} X_{1pi} \\ X_{2pi} \\ X_{3pi} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \theta_{1p} \\ \theta_{2p} \\ \theta_{3p} \end{pmatrix} + \begin{pmatrix} \theta_{1i} \\ \theta_{2i} \\ \theta_{3i} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}. \quad (14)$$

Note that Eq. 14 is simply an extension of Eq. 5. Correspondingly, the *true* parameters used to generate the dataset are expressed in Eq. 15:

Table 7 Code for a $p \times (i^o: h^o)$ design model

```

model{
  for ( p in 1:Np) {
    for (col in 1:Ncol) {
      X[p, col] ~ dnorm (mu[p,col], TauE[iv.format[col]] )
    }
  }

  #Random Effect Components
  for (p in 1:Np) {
    for (col in 1:Ncol) {
      mu[p, col] <-grand.mu[iv.format[col]] + Peffect[p,
iv.format[col]] + Heffect [ ih.format[col]] + IHeffect [i.format[col], ih.format[col],
iv.format[col]] + PHeffect[p, ih.format[col]]
    }
  }

  for ( p in 1:Np) {
    Peffect [ p, 1:V] ~dmnorm( zero.vector[], OmegaP[,] )
  }

  for ( h in 1:Nh){
    Heffect[h] ~dnorm (0, TauH[hv.format[h]])
  }

  for ( col in 1:Ncol){
    for (v in 1:V){
      IHeffect[i.format[col],ih.format[col],v] ~dnorm (0, TauIH[v])
    }
  }
}

```

Table 7 (continued)

```

for ( p in 1:Np){
  for ( h in 1:Nh){
    PHeffect[p, h] ~dnorm(0, TauPH[hv.format[h]])
  }
}

#Priors
for ( v in 1:V) {
  TauE[v]~dgamma(0.01,0.01)
  VarE[v]<-pow(TauE[v],-1)
}

OmegaP [1: V, 1:V] ~dwish(R[,],V)
SigmaP[ 1:V, 1:V]<-inverse(OmegaP[,])

for ( v in 1:V) {
  TauH[v]~dgamma(0.01,0.01)
  VarH[v]<-pow(TauH[v],-1)
}

for ( v in 1:V) {
  TauIH[v]~dgamma(0.01,0.01)
  VarIH[v]<-pow(TauIH[v],-1)
}

for ( v in 1:V) {
  TauPH[v]~dgamma(0.01,0.01)
  VarPH[v]<-pow(TauPH[v],-1)
}
}

```

$$\begin{bmatrix} \sigma_1^2 X_{pi} & \sigma_{12} X_{pi} & \sigma_{13} X_{pi} \\ \sigma_{12} X_{pi} & \sigma_2^2 X_{pi} & \sigma_{23} X_{pi} \\ \sigma_{13} X_{pi} & \sigma_{23} X_{pi} & \sigma_3^2 X_{pi} \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 3 & 2 \\ 3 & 5 & 1 \\ 2 & 1 & 4.5 \end{bmatrix} + \begin{bmatrix} 6.5 & 3.3 & 4.2 \\ 3.3 & 8.5 & 2.4 \\ 4.2 & 2.4 & 4.5 \end{bmatrix} + \begin{bmatrix} 4.1 & 3.5 & 2.2 \\ 3.5 & 5.1 & 2.1 \\ 2.2 & 2.1 & 2.3 \end{bmatrix}. \quad (15)$$

After obtaining the variance and covariance estimates, the generalizability (reliability) coefficient in mG theory is also calculated. The G coefficient in mG theory is defined as:

$$\rho_{\delta}^2 = \frac{\mathbf{a}' \Sigma_p \mathbf{a}}{\mathbf{a}' \Sigma_p \mathbf{a} + \mathbf{a}' \Sigma_{\delta} \mathbf{a} / n_i}. \quad (16)$$

Σ_p is the person effect covariance matrix, whereas Σ_{δ} is the error covariance matrix. In the equation, n_i is the number of items. The \mathbf{a} vector is a weighting scheme; that is, it defines the importance levels of each subtest. There are various approaches to the estimation of weights, which are beyond the

scope of this tutorial (see Brennan, 2001a; Marcoulides, 1994; and Srinivasan & Shocker, 1973, for details). It is evident that different weight schemes can lead to different G coefficients. One can use $\frac{1}{V}$ to deploy an equal-weight scheme if no pre-assumptions are made, where V is the number of the subtest dimensions. Throughout the article, the equal weight scheme is applied throughout the simulation study and real data analysis. Here in the example, the true ρ_{δ}^2 is .904.

The R environment was used to perform the data generation. Using the same terminology from the $p \times i$ design example, the number of persons and the number of items were set to 200 and 10, respectively. In addition, to generate the observed data points, random missingness was purposely created to demonstrate the utility of the Bayesian approach: Missingness was randomly assigned to 10% of the data set. For the Bayesian approach, (1) initial values were automatically generated by **OpenBUGS**, (2) the number of iterations was set to 175,000, (3) the burn-in was set to 8,000, and (4) the thinning lag was set to 60. In addition to the **BUGS** estimation, **mGENOVA** was also used, for the purpose of comparison. This data set and the syntax of both **BUGS** and **mGENOVA** are downloadable.¹ The results are provided in Table 8.

Figure 7 contains trace plots of several selected estimates in the model. All of the plots, including those not shown here due to space limitations, do not show any systematic pattern. This indicates that the MCMC chain reached stationarity (Cowles & Carlin, 1996), which leads to a conclusion that the model converged well, such that further analyses could be initiated. As one can see from Table 8, **BUGS** and **mGENOVA** provided similar point estimates, despite some bias (possibly due to a single replication of data generation). This bias was largely caused by sampling errors; primarily, the insufficiency of the number of the random effects was problematic. In particular, the item effects i had ten levels only, and thus the point estimates deviated more. **mGENOVA** is not able to provide confidence intervals, whereas **BUGS** simultaneously forms credible intervals along with the individual point estimates. Within these 95% credible intervals, the true parameters used to generate the data set were fully covered.

Meanwhile, in the missing-data situation, **BUGS** provided results similar to those in the full situation, but **mGENOVA** was not able to handle the missing-data problems. In addition to variance component estimates, the G coefficient (reliability) can be calculated on fly, and thus form a posterior distribution. The deviance information criterion (DIC), a model fit index, is also provided by **BUGS**, so that model comparisons can become available.

Discussion

The value of the proposed approach method primarily lies in providing uncertainty to estimates that other methods do not have. In practice, practitioners often rely on the point estimate when using G coefficients to make decisions, but it can be problematic if the estimate lacks precision. In the previous example section, the 95% credible interval of the G coefficient ranges from .873 to .916. If one uses .88 as a criterion, he/she may think the test is reliable because the point estimate is .896, which is higher than the criterion. On the other hand, if he/she realizes that uncertainty could pull the G coefficient down below the criterion with a good chance, the decision could be different from the view that only focuses on point estimates.

In addition to the G study, sequential D studies can also be conducted to demonstrate the utility of Bayesian methods. To illustrate, the number of items (or the levels of the item facet) was shifted by a predefined range, by imagining that the test could range from as few as four to as many as 14 items (thus, the possible length could be the ten observed items, plus this range: $[-6, -5, -4, -3, -2, -1, 0, +1, +2, +3, +4]$). Figure 8 shows the posterior distributions of G coefficients as defined in Eq. 15. Assume that .90 is the G-coefficient cut score. One can see that decreasing the number of items to nine would still yield a posterior distribution that would cover the cut score within its 95% credible interval, but reducing to eight or even fewer items would fail to meet this requirement. Similarly, if one would like to ensure that, with a 97.5% chance, the G coefficient would not be lower than the cut score, increasing the number of items to 12 or above would become necessary. Overall, the Bayesian approach provides more information for researchers to make scientific decisions about sequential designs and follow-up practices, where traditional approaches are not capable of doing the same.

In the example section, we used plots to verify the status of convergence. Relying on visualizations only, although they are straightforward and useful, cannot guarantee that the MCMC chain has converged. To provide supplementary supports for monitoring the MCMC process, researchers have proposed various ways to quantify the convergence: For example, (1) Geweke (1991) compared means obtained from different Markov chains, (2) Raftery and Lewis (1992) estimated the minimum chain length needed to estimate a percentile to a certain precision level, (3) Gelman and Rubin's (1992) \hat{R} compares variances between chains, and (4) Brooks and Gelman (1998) modified \hat{R} to \hat{R}_c so that it could account for sampling variability. Here we briefly discuss \hat{R} and its modified version, since they have been widely used in practice (Woodard, 2007).

Suppose there are M chains; after sufficient truncations, the length of each chain is N . For a given estimate β , let $\{\beta_{mt}\}_{t=1}^N$

¹ The simulated dataset and mGENOVA syntax can be downloaded via: https://s3-us-west-1.amazonaws.com/zjiang4/data_code.zip.

Table 8 Estimation of the simulated dataset

name	True Value	Nonmissing Situation				10% Missing Situation		
		mGENOVA	BUGS Estimate	2.50%	97.50%	BUGS Estimate	2.50%	97.50%
σ_{1p}^2	4	3.28	3.28	2.64	4.07	3.37	2.82	3.76
σ_{1p2p}	3	2.83	2.82	2.15	3.61	2.68	1.92	3.76
σ_{1p3p}	2	1.78	1.76	1.20	2.41	1.70	1.08	1.75
σ_{2p}^2	5	4.46	4.45	3.58	5.52	4.14	3.84	6.33
σ_{2p3p}	1	1.42	1.41	0.81	2.10	1.17	0.66	1.57
σ_{3p}^2	4.5	3.66	3.66	2.97	4.48	4.48	3.32	4.50
σ_{1i}^2	6.5	8.77	8.92	3.64	21.29	9.45	3.24	21.89
σ_{1i2i}	3.3	5.86	5.89	1.39	16.08	5.90	1.27	16.62
σ_{1i3i}	4.2	4.04	4.05	1.48	9.93	3.59	2.22	9.03
σ_{2i}^2	8.5	8.17	8.37	3.35	19.97	8.79	3.31	20.17
σ_{2i3i}	2.4	3.57	3.61	1.20	8.93	3.21	1.25	8.71
σ_{3i}^2	4.5	2.26	2.39	0.98	5.75	2.61	0.85	6.25
σ_{1e}^2	4.1	4.10	4.10	3.84	4.39	4.74	4.88	4.06
σ_{1e2e}	3.5	3.45	3.45	3.20	3.72	4.27	3.16	3.83
σ_{1e3e}	2.2	2.20	2.20	2.03	2.39	2.01	2.70	2.35
σ_{2e}^2	5.1	5.01	5.02	4.71	5.36	5.44	3.82	5.79
σ_{2e3e}	2.1	2.06	2.06	1.87	2.25	2.90	1.40	2.55
σ_{3e}^2	2.3	2.34	2.34	2.19	2.50	2.61	3.26	2.62

*The DIC is 22,199 for the non-missing-data situation, and 22,231 for the 10% missing-data situation

be the m th chain, $\hat{\beta}_m$ be the sample posterior mean, and $\hat{\sigma}_m$ be the sample variance of the chain. Akin to the idea of analysis of variance, \hat{R} can be calculated by obtaining between-chains and within-chain variances. Given that $\hat{\beta}$ is the average of $\sum_{m=1}^M \hat{\beta}_m^2$, mathematically it can be expressed as:

$$\hat{R} = \sqrt{\frac{\frac{N-1}{N} \sigma_{wc} + \frac{M+1}{MN} \sigma_{bc}}{\sigma_{wc}}}$$

where $\sigma_{wc} = \frac{N}{M-1} \sum_{m=1}^M (\hat{\beta}_m - \hat{\beta})^2$ and $\sigma_{bc} = \frac{1}{M} \sum_{m=1}^M \hat{\sigma}_m^2$.

Note that σ_{wc} and σ_{bc} represent the between-chain and within-chain variances, respectively. According to Gelman and Rubin (1992), if the convergence has been reached, \hat{R} would be close to 1. Furthermore, correcting for the influence sampling variability through MCMC process, Brooks and Gelman (1998) propose \hat{R}_c by multiplying \hat{R} with $\sqrt{\frac{\hat{d}+3}{\hat{d}+1}}$,

where \hat{d} , the degrees of freedom estimate of a Student t distribution, can be estimated by:

$$d \approx \frac{2(\hat{V})}{\hat{var}(\hat{V})}, \text{ where } \hat{V} = \frac{2(\hat{V})}{\hat{var}(\hat{V})} \sigma_{wc} + \frac{2(\hat{V})}{\hat{var}(\hat{V})} \sigma_{bc}.$$

In practice, model parameters are considered converged if $\hat{R}_c < 1.2$, and models overall are considered converged if $\hat{R}_c < 1.2$ for all model parameters.

DIC values are automatically calculated and available from **BUGS** analyses (except for a few highly complex models). Although it is not discussed in detail here, DIC is beneficial when one is examining whether a model fits data better than an alternative model. For instance, if the data in the example section are fitted with a univariate G-theory model, DIC would be able to indicate the inappropriateness of the wrong model selection when compared with the correctly specified model.

In addition to the capacity to handle missing-data problems, the Bayesian approach to G theory has other beneficial features. Traditional methods can produce negative estimated variance components due to sampling errors or model misspecifications, in which theoretically the sample space is

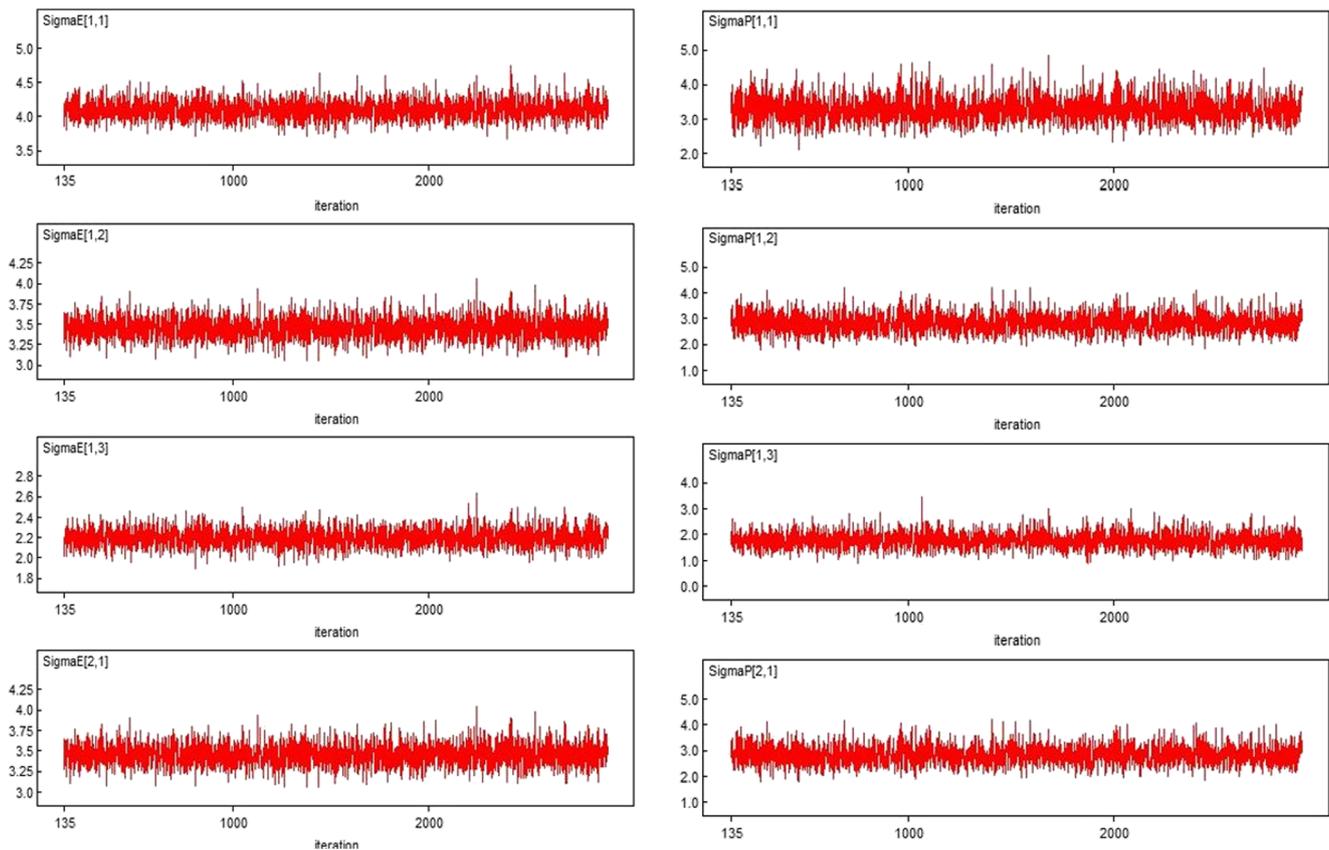


Fig. 7 MCMC trace plots. This figure illustrates the MCMC process for several of the estimate samples. The trace plots for the rest of the estimates follow a pattern similar to the one presented here

truncated at zero to positive infinity (Searle, 1971). Using bounded priors in the Bayesian approach can avoid the problem: It can set a lower bound of zero on the estimates such that values in impermissible numeric space can be prohibited (Box & Tiao, 1973; Fyans, 1977). For example, one can use a uniform prior distribution $(0, \infty)$ to define the sampling space of variance estimates such that the draws of the posterior distributions are bounded above zero. This “brute force” practice, however, is always improper (i.e., it has infinite total mass) and therefore yields unsatisfactory results in many situations. Instead, Gelman, Jakulin, Pittau, and Su (2008) suggest using weakly informative priors to solve the problem; That is, specifying the priors that are not only proper for the concern of analytical convenience but also contain little prior knowledge as fully subjective priors would. To illustrate, setting half-Cauchy with scale 25 as prior distributions of variance parameters in hierarchical linear models is more accurate than using uniform $(0, \infty)$, although both distributions are not informative (see Gelman, 2006, for details). In the present article, inverse Wishart distributions have been used in several models to define the prior information of the covariance matrices. The reason for using inverse Wishart distributions is that if a sample is randomly drawn from a multivariate normal distribution

with covariance matrix Σ , mathematically the sample covariance matrix is proved to follow a Wishart distribution. Specifying a Wishart distribution with the identity matrix (mean) and a positive scalar degrees of freedom parameter V (dispersion), as outlined in Tables 2, 3, 4, 5, 6 and 7, is uninformative and does not guarantee nonnegative estimates. In addition, smaller degrees of freedom, as discussed in the present article, would lead to larger sampling variability in some situations. Chung, Gelman, Rabe-Hesketh, Liu, and Dorie (2015) proposed a class of Wishart priors whose (1) degrees of freedom are set equal to the number of varying coefficients plus 2 and (2) scale matrix is set to the identity matrix multiplied by a value that is large relative to the scale of the problem; this solution assures the posterior distributions of covariance matrices strictly positive definite as well as less errors for the fixed coefficient estimates.

Despite the fact that weakly informative priors can be used to constrain the numeric space of posterior distributions, they do not necessarily “quantify one’s prior beliefs about the likely values for the unknowns independently of the data from the current study” (Dunson, 2001). Using informative priors instead, one can formally incorporate information that is not presented in the present data. This practice can be

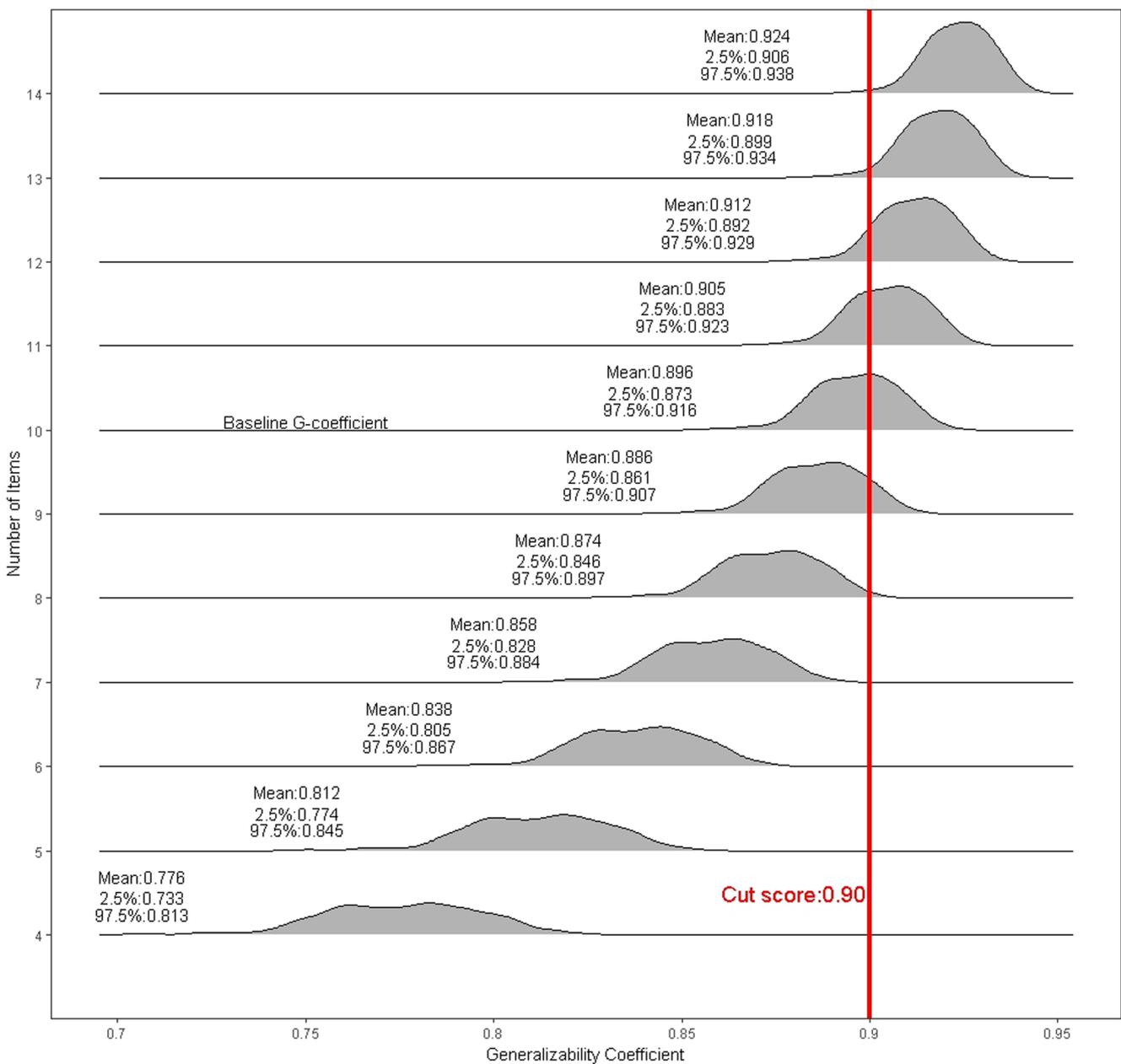


Fig. 8 Generalizability posterior distributions across levels of item facet

controversial because subjective judgments, especially those of uneducated guesses or unverified beliefs, can impede the integrity of the study. To handle the criticism, one reasonable strategy is adopting empirical Bayes methods in which the prior distribution is estimated from the data, or from previous analyses. For instance, given that mG theory variance and covariance components can be estimated via minimizing mean square error (Brennan, 2001a) and some non-fully-crossed designs can be estimated via factor analysis approach (Woehr, Putka, & Bowler, 2012), researchers can specify informative priors based on the results obtained from the frequentist methods.

Summary

In this article, several examples of **BUGS** code are provided to fit many of the common mG-theory designs found in the literature. Examples are provided along with theoretical benefits of estimating these variance components in a fully Bayesian framework. Note that other designs are mentioned in Brennan's (2001a) generalizability theory book that are not presented in the **BUGS** code section. However, it should be easy for readers to extend the current code to those designs not covered in this article.

References

- 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.
- Box, G. E., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. New York, NY: Wiley.
- Brennan, R. L. (2001a). *Generalizability theory*. New York, NY: Springer.
- Brennan, R. L. (2001b). *Manual for mGENOVA*. Iowa City, IA: Iowa Testing Programs, University of Iowa.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.
- Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics*, *40*, 136–157.
- Clauser, B. E., Harik, P., & Margolis, M. J. (2006). A multivariate generalizability analysis of data from a performance assessment of physicians' clinical skills. *Journal of Educational Measurement*, *43*, 173–191.
- Clauser, B. E., Margolis, M., & Swanson, D. B. (2002). An examination of the contribution of computer-based case simulations to the USMLE step 3 examination. *Academic Medicine*, *77*, 80–82.
- Clauser, B. E., Swanson, D. B., & Harik, P. (2002). Multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement*, *39*, 269–290.
- Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*, 883–904.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles*. New York, NY: Wiley.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137–163.
- Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*, *36*, 1–34.
- R Development Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from www.Rproject.org/
- Dunson, D. B. (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, *153*, 1222–1226.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- Fyans, L. J., Jr. (1977). *A new multiple level approach to cross-cultural psychological research*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–533.
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 1360–1383.
- Gelman, A., & Rubin, D. B. (1992). A single series from the Gibbs sampler provides a false sense of security. *Bayesian statistics*, *4*, 625–631.
- Gessaroli, M. E. (2003). *Addressing generalizability theory via structural modelling: Interesting relationships and practical implications*. Paper presented at the Annual meeting of the National Council on Measurement in Education, Philadelphia, PA.
- Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments* (Vol. 196). Minneapolis, MN, USA: Federal Reserve Bank of Minneapolis, Research Department.
- Gilks, W. R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In J.M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 641–649). Oxford, UK: Oxford University Press.
- 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*, 1–13.
- Gleser, G., Cronbach, L., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, *30*, 395–418.
- Jackman, S. (2009). *Bayesian analysis for the social sciences*. New York, NY: Wiley.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, *186*, 453–461. <https://doi.org/10.1098/rspa.1946.0056>
- Joe, G. W., & Woodward, J. A. (1976). Some developments in multivariate generalizability. *Psychometrika*, *41*, 205–217.
- 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*, 144–165.
- Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. New York, NY: Wiley.
- Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049–3067.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337. <https://doi.org/10.1023/A:1008929526011>
- Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer.
- Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports*, *66* (2), 379–386.
- Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement*, *54*, 3–7.
- Marcoulides, G. A. (1995). Designing measurement studies under budget constraints: Controlling error of measurement and power. *Educational and Psychological Measurement*, *55* (3), 423–428.
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. <https://doi.org/10.1037/a0026802>
- Neal, R. M. (1997). *Markov chain Monte Carlo methods based on 'slicing' the density function*. (Technical Report No. 9722). Toronto, ON, Canada: University of Toronto, Department of Statistics and Department of Computer Science.
- Norusis, M. J. (1990). *SPSS: Statistical data analysis*. SPSS.
- Nußbaum, A. (1984). Multivariate generalizability theory in educational measurement: An empirical study. *Applied Psychological Measurement*, *8*, 219–230.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, p. 125). Vienna, Austria: Technische Universität Wien. Retrieved from <https://www.r-project.org/conferences/DSC-2003/Proceedings/>
- Plummer, M. (2010). *JAGS Version 2.2.0 manual*. Available from <http://mcmcjags.sourceforge.net>.

- Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2008). *Comparison of subscores based on classical test theory methods* (ETS Research Report No. RR-08-54). Princeton, NJ: Educational Testing Service.
- Raftery, A. E., & Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In J.M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 763–773). Oxford, UK: Oxford University Press.
- Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modelling approach to scale reliability evaluation. *International Journal of Testing, 6*, 81–95.
- SAS Institute. (1990). *SAS/STAT user's guide: Version 6* (Vol. 2). Cary, NC: Author.
- Searle, S. R. (1971). *Linear models*. New York, NY: Wiley.
- Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46*, 553–611.
- Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133–166.
- Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*, 150–174.
- Sinharay, S. (2013). A note on assessing the added value of subscores. *Educational Measurement: Issues and Practice, 32*, 38–42.
- Sinharay, S., & Haberman, S. J. (2014). An empirical investigation of population invariance in the value of subscores. *International Journal of Testing, 14*, 22–48.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS user manual*. Cambridge, UK: MRC Biostatistics Unit.
- Srinivasan, V., & Shocker, A. D. (1973). Estimating the weights for multiple attributes in a composite criterion using pairwise judgments. *Psychometrika, 38*, 473–493.
- Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News, 6*, 12–17.
- 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*, 397–408.
- Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait–multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods, 15*, 134–161.
- Woodard, D. B. (2007). Detecting poor convergence of posterior samplers due to multimodality (Discussion Paper 2008-05). Duke University, Department of Statistical Science, Durham, NC.
- Woodward, J. A., & Joe, G. W. (1973). Maximizing the coefficient of generalizability in multi-facet decision studies. *Psychometrika, 38*, 173–181.
- Wu, Y. F., & Tzou, H. (2015). A multivariate generalizability theory approach to standard setting. *Applied Psychological Measurement, 39*, 507–524.