Journal of Educational Measurement
Summer 2007, Vol. 44, No. 2, pp. 131–155

Generalizability in Item Response Modeling

Derek C. Briggs
University of Colorado, Boulder
Mark Wilson
University of California, Berkeley

An approach called generalizability in item response modeling (GIRM) is introduced in this article. The GIRM approach essentially incorporates the sampling model of generalizability theory (GT) into the scaling model of item response theory (IRT) by making distributional assumptions about the relevant measurement facets. By specifying a random effects measurement model, and taking advantage of the flexibility of Markov Chain Monte Carlo (MCMC) estimation methods, it becomes possible to estimate GT variance components simultaneously with traditional IRT parameters. It is shown how GT and IRT can be linked together, in the context of a single-facet measurement design with binary items. Using both simulated and empirical data with the software WinBUGS, the GIRM approach is shown to produce results comparable to those from a standard GT analysis, while also producing results from a random effects IRT model.

As measurement models, item response theory (IRT) and generalizability theory (GT) seem, on the surface at least, incompatible. Brennan (2001a), for example, writes "Generalizability Theory is primarily a sampling model, whereas IRT is principally a scaling model." Nonetheless, because each approach can provide information fundamental to the design and analysis of measurement instruments, one might expect to see IRT and GT applied in tandem, both in large-scale testing and smaller scale efforts (see, for example, Bock, Brennan, & Muraki, 2002). In practice, the use of IRT alone seems considerably more prevalent in the measurement literature than the sequential use of both IRT and GT.

In this article we introduce an approach we call "Generalizability in Item Response Modeling" (GIRM). The GIRM approach essentially incorporates the sampling model of GT into the IRT "scaling model" by making distributional assumptions about the relevant measurement facets. Given these assumptions, and taking advantage of the flexibility of Markov Chain Monte Carlo (MCMC) estimation methods, it becomes possible to estimate GT variance components within the same framework, and simultaneously, with traditional IRT parameters. The underlying model of the GIRM approach comes from IRT, but the results from this model are used as the basis for a GT analysis. In essence, what we are calling the GIRM approach performs a GT analysis on a matrix of expected, rather than observed, item responses. It is our introduction of this matrix of expected item responses, and our approach to estimating a distribution of values for each element of this matrix, that are the principal new methodological contributions of our article. The objectives of this article are to (1) present the details of the GIRM approach, (2) show that it can produce results comparable to those of GT for a simple measurement design, and (3) test the robustness of

the approach to some violations of the principal assumptions that distinguish it from GT. Along the way we will highlight what we view as some interesting similarities and distinctions between GT and IRT.

There are five sections to this article. In the first two sections, we briefly present GT, and then a random effects IRT model that is consistent with GT sampling assumptions. In the third section we present the GIRM approach by way of showing how GT and IRT can be linked together in the context of a single-facet measurement design with binary items. In the fourth section, using simulated data and MCMC estimation procedures within the software WinBUGS, we demonstrate how the GIRM approach can produce results equivalent to those from a standard GT analysis, and we test the sensitivity of the approach to several key assumptions. In the fifth section, we apply both the GIRM approach and GT to an empirical dataset, and show how the use of results in the language of both GT and IRT can serve to strengthen interpretations about the design and analysis of a measurement instrument. In the last section, we discuss some of the apparent strengths and limitations of the GIRM approach.

Generalizability Theory

In classical test theory, a person's observed test score, X, is modeled linearly as the sum of a "true score" T and "error," ε: $X = T + \varepsilon$. GT has been described as a liberalization of classical test theory because it helps to differentiate, using ANOVA-like procedures, the multiple sources of error that comprise ε. In GT terminology, a potential source of measurement error is called a *facet*. For example, the facets of an achievement test might include items, raters, test forms, or test administrators. The conditions of each facet of any particular test are viewed as a random sample from the "universe of admissible observations." A person's expected score, taken as an average across all possible facet conditions, will generally not equal a person's observed score for the set of facet conditions sampled for use in any given measurement. The difference between the observed score and expected score can be explained in part by facet-based measurement error. The key question in GT is the degree to which a person, responding a certain way to a measurement instrument with randomly sampled facet conditions, would respond approximately the same way when faced with a different random sample of facet conditions.

A GT analysis has two stages. In the first stage, called a generalizability study (G Study), the relevant facets in a measurement design are defined along with the universe of admissible observations for these facets (i.e., the populations from which the facets are presumably sampled). A G Study proceeds by estimating variance components for a single condition of each facet main effect and facet interaction in the measurement design. In the second stage, called a Decision Study (D Study), it is determined how much an increase or a decrease in the number of conditions for a particular facet would decrease or increase the various types of error variance in the measurement procedure, and thus increase or decrease the generalizability (reliability). When a measurement procedure is multifaceted, the results of a D Study will suggest the optimal combination of facet conditions (i.e., number of items, raters, test forms, and test administrators) necessary to secure some minimum level of generalizability.

We illustrate the GT approach using a single-facet measurement design and notation that is for the most part consistent with that presented by Cronbach, Gleser, Nanda, and Rajaratnam (1972) and Brennan (2001a). Let X_{pi} be the dichotomous observed score of person p ($p = 1, \ldots, P$) on item i ($i = 1, \ldots, I$). This scenario is represented in a G Study as a $p \times i$ design, where p is the object of measurement and i is the facet. The strongest and most fundamental assumption of GT is that the object of measurement and measurement facets, persons and items, are sampled independently and at random from some population of persons and a "universe" of items. The observed score X_{pi}, based on the combination of any one-person p and any one-item i, is considered a random variable. We further assume that the design is balanced.[1] The grand mean across persons and items is defined as

$$\mu \equiv \underset{p}{E}\,\underset{i}{E}\, X_{pi}. \tag{1}$$

Next, we can define a person-specific mean as

$$\mu_p \equiv \underset{i}{E}\, X_{pi}, \tag{2}$$

and then the item-specific mean as

$$\mu_i \equiv \underset{p}{E}\, X_{pi}. \tag{3}$$

Given these definitions, a linear model for observed score X_{pi} can be written as

$$X_{pi} = \mu + v_p + v_i + v_{pi,e}, \tag{4}$$

where the successive terms to the right of the observed score X_{pi} are, respectively, μ, the grand mean; $v_p = \mu_p - \mu$, the person effect; $v_i = \mu_i - \mu$, the item effect; and $v_{pi,e} = X_{pi} - \mu_p - \mu_i + \mu$, the residual/interaction effect. Note that in this expression the concept of "error" is defined in terms of that which is left over (the residual) when person and item effects are subtracted from the grand mean. This term is sometimes described as the p by i interaction effect confounded with all remaining, unspecified sources of random error.

We assume[2] that all effects are uncorrelated. In formalizing this, we follow the convention presented by Brennan (2001a), letting a prime designate a different person or item:

$$E(v_p v_{p'}) = E(v_i v_{i'}) = E(v_{pi,e} v_{p'i,e}) = E(v_{pi,e} v_{pi',e}) = E(v_{pi,e} v_{p'i',e}) = 0, \tag{5}$$

$$E(v_p v_i) = E(v_p v_{pi,e}) = E(v_i v_{pi,e}) = 0. \tag{6}$$

The variance components for the person, item, and person by item effects are, respectively

$$\sigma^2(p) = \underset{p}{E}(\mu_p - \mu)^2, \tag{7}$$

$$\sigma^2(i) = \underset{i}{E}(\mu_i - \mu)^2, \text{ and} \tag{8}$$

$$\sigma^2(pi, e) = \underset{p}{E}\,\underset{i}{E}(X_{pi} - \mu_p - \mu_i + \mu)^2. \tag{9}$$

This leads to the decomposition of the total variance of observed scores as

$$\sigma^2(X_{pi}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(pi, e). \tag{10}$$

The culmination of the G Study stage is estimating values for these unobservable variance components using a series of expected mean square equations (Cornfield & Tukey, 1956; Searle, Casella, & McCulloch, 1992). In this case:

$$EMS(p) = \sigma^2(pi) + n_i\sigma^2(p), \tag{11}$$

$$EMS(i) = \sigma^2(pi) + n_p\sigma^2(i), \text{ and} \tag{12}$$

$$EMS(pi, e) = \sigma^2(pi, e). \tag{13}$$

In the equations above n_i and n_p represent the number of items and persons sampled for the G Study, respectively. We estimate $\hat{\sigma}^2(\cdot)$ in this system of three equations with three unknown values by replacing $EMS(\cdot)$ with the observed mean square errors $MS(\cdot)$, calculated from the observed scores in a person by item matrix, as illustrated in Figure 1.

Solving the system of equations that result from substituting the values of observed mean squares for the expected mean squares in Equations (11–13) leads to the G Study variance component estimates $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(pi, e)$.

		items				
		1	...	i	...	I
persons	1	X_{11}	...	...	...	X_{1I}
	...	...	...	...	...	...
	p	...	...	X_{pi}	...	...
	...	...	...	...	...	...
	P	X_{P1}	...	...	...	X_{PI}

FIGURE 1. *Observed response matrix in GT.*

In a subsequent D Study, the "decision" to be made is the number of sampled conditions of a facet (e.g., n'_I) to be included in the measurement instrument. The observed score variability of an instrument due to a particular facet decreases as more conditions of that facet are included as part of the design. For the $p \times I$ design (where an upper case I denotes that scores are now computed as an average over a set of items), as more items are included in the measurement instrument, the observed score variance associated with the item facet $\frac{\hat{\sigma}^2(I)}{n'_I}$ and the person by item interaction $\frac{\hat{\sigma}^2(pI,e)}{n'_I}$ decreases. If the aim is to generalize the test results from a sample of items to the larger universe of items they are intended to represent, then the less that observed score variance is due to the sampling instances of the item facet and its interactions, the higher the generalizability of the measurement procedure on the whole.

The generalizability of observed scores are summarized in a D Study by two indices, $\mathrm{E}\rho^2$ and Φ. For the $p \times I$ example,

$$E\hat{\rho}^2 = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \dfrac{\hat{\sigma}^2(pi,e)}{n'_I}} \tag{14}$$

$$\hat{\Phi} = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \dfrac{\hat{\sigma}^2(pi,e)}{n'_I} + \dfrac{\hat{\sigma}^2(i)}{n'_I}}. \tag{15}$$

In the GT literature, these two indices have been called the *generalizability coefficient* and *index of dependability*, respectively. We prefer to describe both indices as generalizability coefficients, where the principle distinction is whether observed scores are being used to make relative (i.e., norm-referenced) or absolute (i.e., criterion-referenced) decisions. For relative decisions (e.g., who were the top scorers on the test?), only the standing of persons relative to others persons is of interest, so only sources of error that interact with the object of measurement, known as *relative error variance*, are added to $\hat{\sigma}^2(p)$, in the denominator. For the $p \times i$ design, $E\hat{\rho}^2$ is equivalent to Cronbach's coefficient α. For absolute decisions (e.g., how many students passed the test?), the error variance of the item facet main effect (and all other main effects in more complex designs) is added to the denominator, and this comprises *absolute error variance*. It follows that the generalizability for absolute decisions will always be smaller than that for relative decisions.

We note in passing that the generalizability coefficients presented above are a function of both the quality of a particular measurement procedure, and the sample to which the instrument has been applied. If true variance in the sample is very small, then even if the quality of the procedure is high in an absolute sense (i.e., little measurement error), a generalizability coefficient will be low. Conversely, if true variance is large, then even if a procedure is very inaccurate (i.e., large measurement error), a generalizability coefficient will remain high. Because of this, the interpretation of these sorts of coefficients can sometimes be misleading, and it is often preferable,

or at least complementary, to report the square root of relative and absolute error variance as relative and absolute standard errors of measurement.

For the $p \times I$ example used here, it is the difference between relative and absolute error variance that distinguishes GT from CTT. That is, if score inferences are to be generalized in an absolute sense, it will matter a great deal to the object of measurement whether the measurement procedure, by chance, consisted of items that were very difficult to answer. On the other hand, for score inferences to be generalized only in a relative sense, it makes no difference from the perspective of GT whether the measurement procedure is based upon items that were extremely hard or easy by chance, as all objects of measurement are affected the same way. For more complex (i.e., multifaceted) measurement designs, the key benefits of GT are both the distinction between relative and absolute error variance, and the unique attribution of each measurement facet as a source of both absolute and relative error variance.

Item Response Theory as a Random Effects Model

A measurement procedure can be modeled using IRT for purposes that may well be complementary to a GT analysis. Though, as the name implies, IRT is typically intended for the analysis of how subjects respond when faced with a set of items, "items" are often defined quite flexibly to make IRT applicable to most of the same measurement procedures found in GT. The $p \times i$ design with dichotomous items could be modeled using IRT according to the item response function

$$P(\theta_p, \beta_i) = P(X_{pi} = 1 \mid \theta_p, \beta_i) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}. \tag{16}$$

This is an example of what is commonly known as the one parameter logistic, or Rasch model. The probability of a correct response to the observed item score X_{pi} is modeled as a nonlinear function of person proficiency, θ_p, and item difficulty, β_i. To be consistent with GT, we would need to interpret $P(\theta_p, \beta_i)$ as the probability that a randomly sampled person with ability θ_p will respond correctly to a randomly sampled item with difficulty β_i. Other item response functions with additional item parameters could be chosen to model the $p \times i$ design, but here we use the Rasch model because it is the simplest and most easily interpretable item response model.

In common with other item response models, the Rasch model assumes statistical independence across p, and conditional independence across i given θ. It is further typically assumed that the latent variable θ being measured is unidimensional. Now, when the Rasch model is estimated with θ_p and β_i as fixed effects, Brennan's characterization of IRT as primarily a scaling model is accurate. However, in most large-scale uses of item response models, an implicit sampling model for persons is incorporated through the use of marginal maximum likelihood estimation (cf. Holland, 1990). The assumption under marginal maximum likelihood is that the latent variable for each person is drawn at random from some population distribution of interest. This brings item response models much closer to the assumptions of GT. If we take the next step, and assume that person and item parameters are each drawn

at random from a population of interest, then the transition from scaling model to scaling and sampling model is conceptually complete. Random effects models with these sorts of assumptions are commonly used in Bayesian IRT applications.

Beyond the assumption of conditional independence, in a random effects IRT model assumptions must be invoked to specify prior distributions $f(\theta)$ and $g(\beta)$ for the random person and item parameters, θ_p and β_i. Often (and this is the approach we take in what follows), normal densities are chosen for θ_p and β_i, but others would be possible. For identification purposes, the first moment of either $f(\theta)$ or $g(\beta)$ is constrained to equal 0. Parameter estimation in a random effects IRT proceeds by defining the likelihood function for the $P \times I$ matrix of person by item responses as

$$L(\mathbf{X} \mid \theta, \beta) = \prod_{p=1}^{P} \prod_{i=1}^{I} P(\theta_p, \beta_i)^{X_{pi}} [1 - P(\theta_p, \beta_i)]^{1-X_{pi}}, \tag{17}$$

where θ and β are vectors of person and item parameters, and $\mathbf{X}$ is the observed data matrix. The joint density function

$$P(\mathbf{X}, \theta, \beta) = L(\mathbf{X} \mid \theta, \beta) f(\theta) g(\beta) \tag{18}$$

results from combining the "scaling model" defined in Equation (17) with the "sampling model" defined by the prior distributions $f(\theta)$ and $g(\beta)$. Our interest is in (a) the posterior distribution of θ for each person, conditional on the difficulty of the test items in the item population and the observed response vector; and (b) the posterior distribution of β for each item, conditional on the ability of persons in the target population and the observed response vector. Namely,

$$P(\theta \mid \beta, \mathbf{X}) = \frac{L(\mathbf{X} \mid \theta, \beta) f(\theta) g(\beta)}{\int\limits_{\theta} L(\mathbf{X} \mid \theta, \beta) f(\theta) g(\beta) d\theta}, \text{ and} \tag{19}$$

$$P(\beta \mid \theta, \mathbf{X}) = \frac{L(\mathbf{X} \mid \theta, \beta) f(\theta) g(\beta)}{\int\limits_{\beta} L(\mathbf{X} \mid \theta, \beta) f(\theta) g(\beta) d\beta}. \tag{20}$$

The conditional distributions in Equations (19) and (20) can be estimated using MCMC techniques. It is from these distributions that values are drawn to characterize $P(\theta_p, \beta_i)$ for each person-by-item combination in the Rasch model item response function. More specifically, $P(\theta \mid \beta, \mathbf{X})$ and $P(\beta \mid \theta, \mathbf{X})$ are estimated using the Metropolis–Hastings algorithm within the Gibbs sampler. Once these posterior distributions for θ_p and β_i have been estimated, we can summarize these distributions by their mean (e.g., the "expected a posteriori"), median, and standard deviation.[3]

In the random effects IRT context, the impact of measurement error can be quantified in a way that is analogous to the CTT concept of reliability by comparing, for

each person and item, the variance in the posterior distribution to the variance in the prior distribution, where each variance represents a quantification of posterior and prior measurement uncertainty. For an instrument that measures person and item parameters with great precision, these ratios (which will range from 0 to 1), should be quite small, although once again, as in GT, this will also depend upon the characteristics of the person and item samples. The smaller the ratio, the more "reliable" the scores produced by the instrument. We can produce an estimate for the *marginal* reliability (Adams, 2006; Green, Bock, Humphreys, Linn, & Reckase, 1984; Mislevy, Beaton, Kaplan, & Sheehan, 1992) of person and item parameter estimates by taking the average of the posterior to prior distribution variance ratios over persons and items, respectively. We subtract these from 1 to put them on the same scale as a CTT-based reliability coefficient, hence

$$R_p = 1 - \frac{\bar{\sigma}_p^2}{\text{var}(\theta)}, \text{ and} \tag{21}$$

$$R_i = 1 - \frac{\bar{\sigma}_i^2}{\text{var}(\beta)}. \tag{22}$$

In Equations (21) and (22) the terms $\bar{\sigma}_p^2$ and $\bar{\sigma}_i^2$ represent the variance in posterior distributions averaged over persons and items respectively, while the terms var(θ) and var(β) represent the variance in prior distribution for the population of persons and items. Note that in IRT models where items are considered fixed effects, the concept of reliability/generalizability with respect to items is meaningless. However, in a fully random effects model, the object of measurement can be thought of as either persons or items, so reliability can be expressed either with respect to persons or items. The same is true, of course, in GT if the object of measurement is items; generalizability coefficients can be estimated with respect to items instead of persons.

Relative to GT, there are two limitations to the IRT marginal reliability coefficients R_p and R_i. First, because they are not expressed in closed form as a function of sample sizes, we cannot apply Spearman–Brown adjustments as in a GT D Study to predict how the values will change as person or item sample size changes. Second, the terms to the right of the equal sign in Equations (21) and (22) will not change, even when we have a measurement procedure that is multifaceted. This makes it difficult to disentangle the contribution that distinct facets make to an increase or decrease in reliability. It would seem then, that there are some useful concepts within GT, which currently have no clear analog in IRT.

From IRT to GT

An insight that may not be readily apparent is that both the GT model used for a G Study and IRT models can be conceptualized as instances of multilevel statistical models (Goldstein, 1995). This is a point that has been made nicely by Verhelst and Verstralen (2001) and Patz et al. (2002). (For in-depth presentations of just IRT

models from a multilevel perspective see Raudenbush & Bryk, 2002, pp. 365–371; Van den Noortgate & Paek, 2004, pp. 167–187). Given the conceptual similarity between GT and IRT as multilevel random effects models, it should come as no surprise that we can derive, given certain assumptions, the variance component and generalizability coefficient estimates central to GT from within an IRT framework. In this section we show how this can be accomplished.

The Kolen and Harris Link

A foundation for the link between IRT parameters and GT variance components was supplied in an unpublished conference article by Kolen and Harris (1987). Kolen and Harris started with an item response function as in Equation (16) and then defined the following parameters:

$$\mu = \int_{\theta}\int_{\beta} P(\theta_p, \beta_i) f(\beta) g(\theta) d\beta d\theta, \tag{23}$$

$$\pi(\theta) = \int_{\beta} P(\theta_p, \beta_i) f(\beta) d\beta - \mu, \tag{24}$$

$$\iota(\beta) = \int_{\theta} P(\theta_p, \beta_i) g(\theta) d\theta - \mu, \text{ and} \tag{25}$$

$$\nu(\theta, \beta) = P(\theta_p, \beta_i) - \pi(\theta) - \iota(\beta) - \mu. \tag{26}$$

In comparing the IRT-based parameters represented in Equations (24–26) with the GT parameters represented in Equations (1–3), we can interpret μ as the mean person-item response equivalent to the GT grand mean, $\pi(\theta)$ as the person-specific mean item response equivalent to the GT person effect ν_p, and $\iota(\beta)$ as the item-specific person response equivalent to the GT item effect ν_i . Note that the person-item interaction effect $\nu\ (\theta, \beta)$ in Equation (26) is not directly comparable to the residual effect $\nu_{pi,e}$ in Equation (4). This is because $\nu\ (\theta, \beta)$ is conceptually distinct from other unmeasured sources of error in the IRT formulation.

The terms in Equations (24–26) have expectations and covariances of zero. It follows that,

$$\begin{aligned}\int_{\theta} \pi(\theta) g(\theta) d\theta &= \int_{\beta} \iota(\beta) f(\beta) d\beta = \int_{\theta} \nu(\theta, \beta) g(\theta) d\theta \\ &= \int_{\beta} \nu(\theta, \beta) f(\beta) d\beta = \int_{\theta}\int_{\beta} \nu(\theta, \beta) f(\beta) g(\theta) d\beta d\theta = 0.\end{aligned} \tag{27}$$

The variance components for Equations (24–26) can now be defined as

$$\sigma^2(\pi) = \int_\theta \pi^2(\theta) g(\theta) d(\theta)\,; \tag{28}$$

$$\sigma^2(\iota) = \int_\beta \iota^2(\beta) f(\beta) d(\beta);\ \text{and} \tag{29}$$

$$\sigma^2(\nu) = \int_\theta \int_\beta \nu^2(\theta, \beta) f(\beta) g(\theta) d\beta d\theta\,. \tag{30}$$

To create an IRT-based analog to the GT linear equation represented by Equation (4), Kolen and Harris re-wrote Equation (26) as

$$P(\theta_p, \beta_i) = \mu + \pi(\theta) + \iota(\beta) + \nu(\theta, \beta), \tag{31}$$

where, as we noted in Section 3, $P(\theta_p, \beta_i) = E(X_{pi})$. Next, they defined the random error term $e_{pi} = e(\theta_p, \beta_i)$ as the error component for a random examinee p of ability θ_p responding to a random item i of difficulty β_i. So the expected value of an item response can be translated into the observed score metric directly comparable to Equation (4) as

$$X_{pi} = E(X_{pi}) + e_{pi} = \mu + \pi(\theta) + \iota(\beta) + \nu(\theta, \beta) + e_{pi}. \tag{32}$$

In parallel with GT, the concept of error is defined as the residual left over when the observed score is subtracted from the expected score, but in this case the expected score has been decomposed to include a distinct person-item interaction effect. Of course, the truth of Equation (32) depends on the truth of the assumptions on which it is based, including that the item response function specified in Equation (16) holds. Under the assumption that the observed variable X_{pi} has a Bernoulli distribution, the variance of e_{pi} will be $P(\theta_p, \beta_i)[1 - P(\theta_p, \beta_i)]$. Taken over all students and all items,

$$\sigma^2(e) = \int_\theta \int_\beta P(\theta_p, \beta_i)\left[1 - P(\theta_p, \beta_i)\right] f(\beta) g(\theta) d\beta d\theta. \tag{33}$$

This leads to the decomposition of the total variance of observed scores as

$$\sigma^2(X_{pi}) = \sigma^2(\pi) + \sigma^2(\iota) + \sigma^2(\nu) + \sigma^2(e). \tag{34}$$

This equation is directly comparable to the GT decomposition of total variance, where $\sigma^2(\pi)$, is the equivalent of $\sigma^2(p)$, $\sigma^2(\iota)$ is the equivalent of $\sigma^2(i)$, and $\sigma^2(\nu) + \sigma^2(e)$ is the equivalent of $\sigma^2(pi, e)$.

The Expected Response Matrix

A key contribution of this article is to suggest an approach for estimating the variance components in Equation (34) concurrently with the parameters typical of the IRT formulation for a $p \times i$ design. We begin by pointing out that if the item response model parameters θ_p and β_i were known for all persons and items, we could use these parameters to generate an expected response matrix. Such a matrix would take the form shown in Figure 2.

Given dichotomous items and the Rasch model item response function from Equation (16), in each cell of the matrix $E(X_{pi}) = P(\theta_p, \beta_i)$. Now, if such an expected response matrix could be computed, then estimating scalar and vector values for $\pi(\theta)$, $\iota(\beta)$, and $\nu(\theta, \beta)$ would be straightforward. For example, to estimate a vector of person effects $\hat{\pi}(\theta)$, integration over β could be approximated by taking the average across the columns β_i for each value of θ_p and then subtracting the mean person-item response, $\hat{\mu}$, which itself is estimated by taking the average over all columns and rows of the expected response matrix. The same approach would be taken to estimate the variance components $\sigma^2(\pi)$, $\sigma^2(\iota)$, $\sigma^2(\nu)$, and $\sigma^2(e)$. To arrive at an estimate of $\sigma^2(\pi)$, integration of $\pi(\theta)^2$ over θ is approximated by taking the average of $\hat{\pi}(\theta)^2$ over the rows θ_p of the expected response matrix. That these parameter estimates will be unbiased (i.e., BQUE, see Searle et al., 1992) depends upon the assumption that both persons and items have been sampled in the way that we have specified with our prior distributions.

In practice θ_p and β_i are unknown and must themselves be estimated. This can be done by estimating posterior distributions for θ_p and β_i, as in Equations (19) and (20), using MCMC estimation of a random effects item response model. For each step m of the Markov Chain, a new estimate of θ_p and β_i is generated: $\hat{\theta}_p^{(m)}$ and $\hat{\beta}_i^{(m)}$. These estimates are then used to compute the elements of the expected response matrix, which are in turn used to produce the variance component estimates $\hat{\sigma}^2(\pi)^{(m)}$, $\hat{\sigma}^2(\iota)^{(m)}$, $\hat{\sigma}^2(\nu)^{(m)}$, and $\hat{\sigma}^2(e)^{(m)}$. With each new step of the Markov Chain, a new expected response matrix and set of variance components are estimated. The process culminates in a posterior distribution for each variance component, where

item parameters

person parameters	β_1	...	β_i	...	β_I
θ_1	$E(X_{11})$	...	...	...	$E(X_{1I})$
...	...	...	...	...	...
θ_p	...	...	$E(X_{pi})$	...	...
...	...	...	...	...	...
θ_P	$E(X_{P1})$	...	...	...	$E(X_{PI})$

FIGURE 2. *Expected response matrix in GIRM.*

each distribution can be summarized with respect to its central tendency and spread. It is our hypothesis that the means of these posterior distributions will lead to the same point estimates as would be produced under GT. The uncertainty of these estimates would usually be quantified by the standard deviations of the posterior distributions. As we discuss later, these standard deviations produced under the GIRM approach will be likely to be biased downward.

The GIRM Approach with Simulated Data

Comparing GIRM and GT

Of primary interest is in how GIRM estimates of variance components and generalizability coefficients compare to those that would be produced by analyzing the simulated data directly using GT.[4] As an initial test of the GIRM approach, a 500 by 5 person by item matrix of simulated item responses was replicated 100 times according to the Rasch model with the parameters θ_p and β_i drawn independently from standard normal distributions.[5] Both GT and GIRM procedures were applied to these 100 data sets. For each replication, GIRM estimates were produced from relevant item and person parameter posterior distributions using the software WinBUGS.[6] Posterior distributions were based on MCMC estimation using three chains with 10,000 iterations after a burn-in of 1,000 iterations. GT estimates were computed using standard formulas in the *R* statistical programming environment.

Table 1 compares the mean and standard deviation of the 100 posterior distribution means in each replication of the GIRM approach to the mean of the 100 point estimates in each replication of the GT approach. When both GIRM and GT are applied to our simulated data matrix, the results are very similar. The estimated generalizability coefficient for relative decisions is .474 under the GIRM approach, and .471 under the GT approach. The estimated generalizability coefficient for absolute decisions is .439 under the GIRM approach, and .430 under the GT approach. As we would expect, the generalizability of scores from a test consisting of five dichotomous items is quite low. The relationship between the generalizability coefficients can be examined graphically in Figures 3 and 4. As is evident from these scatterplots,

TABLE 1
Simulated Variance Component Estimates in GIRM and GT

Variance Components/G Coefficients	GIRM Mean and *SD* of Posterior Mean across 100 Reps	GT Mean and *SD* of Point Estimate across 100 Reps
$\hat{\sigma}^2(p)$	.033 (.006)	.033 (.006)
$\hat{\sigma}^2(i)$	.028 (.016)	.034 (.020)
$\hat{\sigma}^2(pi)$	.002 (.001)	NA
$\hat{\sigma}^2(e)$	.181 (.014)	NA
$\hat{\sigma}^2(pi, e)$	.002 + .181	.183 (.013)
$E\hat{\rho}^2$	.474 (.037)	.471 (.037)
$\hat{\Phi}$	.439 (.048)	.430 (.052)

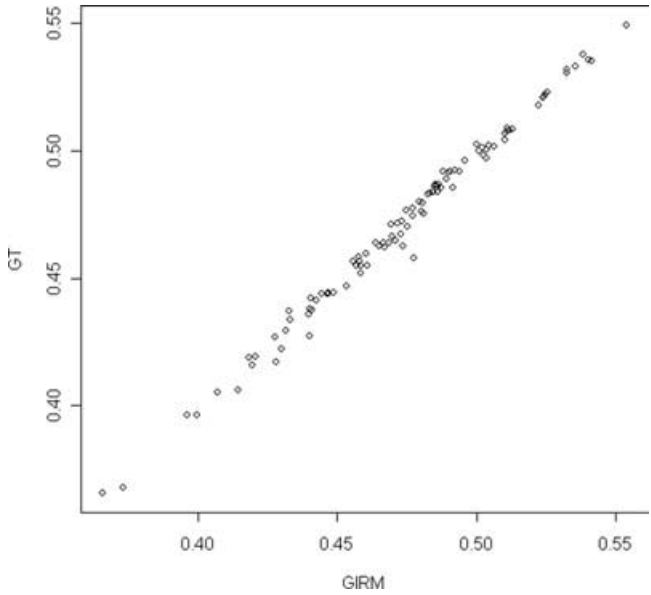

FIGURE 3. *Simulated estimates of generalizability for relative decisions ($E\hat{\rho}^2$).*

the indices of $E\hat{\rho}^2$ and $\hat{\Phi}$ under the GIRM and GT approaches have nearly a perfect linear association, with correlations of about 1.

One useful feature of the GIRM approach is apparent in the results summarized above: we are able to estimate separate variance components for the person by item interaction, $\hat{\sigma}^2(pi)$ and for error, $\hat{\sigma}^2(e)$. In GT, these components are confounded within the estimate $\hat{\sigma}^2(pi, e)$. The distinction is illustrated below in Figure 5, and is attributable to the added assumption under GIRM about the distribution of the error term, e_{pi} (i.e., as embodied in the assumption of a particular item response model).

The Uncertainty of Variance Component Estimates

The variance components in both GT and GIRM are estimated with some uncertainty, and it is desirable to quantify this uncertainty. Many approaches for estimating the uncertainty in GT variance component estimates have been proposed, and this continues to be an active area of research (cf., Betebenner, 1998; Brennan, 2001a, Chapter 6; Gao & Brennan, 2001; Wiley, 2001). In the GIRM approach, posterior distributions are estimated for all variance components and generalizability coefficients, with the means of these distributions reported as point estimates comparable to those produced using the most frequent ANOVA-like procedures in GT. The uncertainty in posterior distributions is typically summarized using the posterior standard deviations, and these are provided in Table 2 above. So, for example, when the GIRM approach was applied to a single simulation of the 500 person by 5 item data matrix we can put a 95% credibility interval of about $\pm.07$ around the estimated generalizability (.43) of a five-item instrument for relative decisions.

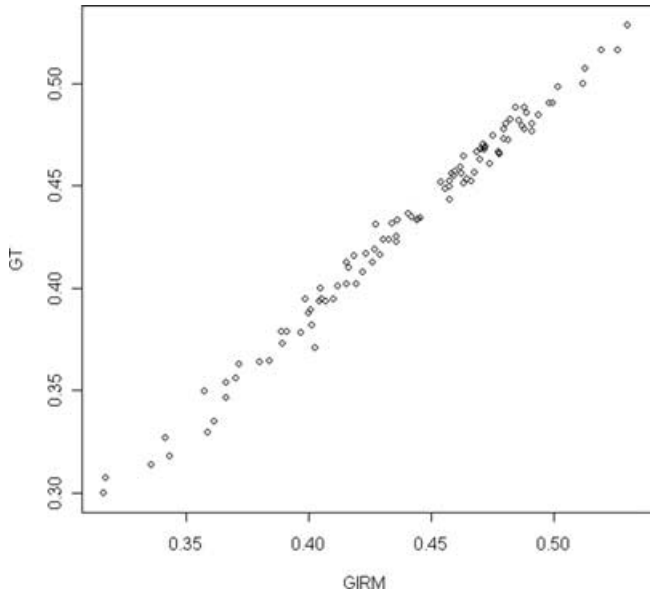

FIGURE 4. *Simulated estimates of generalizability for absolute decisions ($\hat{\Phi}$).*

Unfortunately, the standard deviations for posterior distributions of the variance components and generalizability coefficients in the GIRM approach are very likely to be biased downward. One cause of this bias has been described in some detail by Wiley (2001) in the context of bootstrap procedures for the estimation of variance component standard errors. The problem arises from the fact that across steps of the Markov Chain, estimates of variance components are not computed independently, which leads to a violation of the assumption of random effects. For example, take the variance component $\hat{\sigma}^2(i)$ estimated from the simulated data above. After an iteration of the Markov Chain, the GIRM approach would use 500 values of $\hat{\theta}_p$ and 5 values of $\hat{\beta}_i$ to generate an expected response matrix. From this, $\hat{\sigma}^2(i)$ is effectively

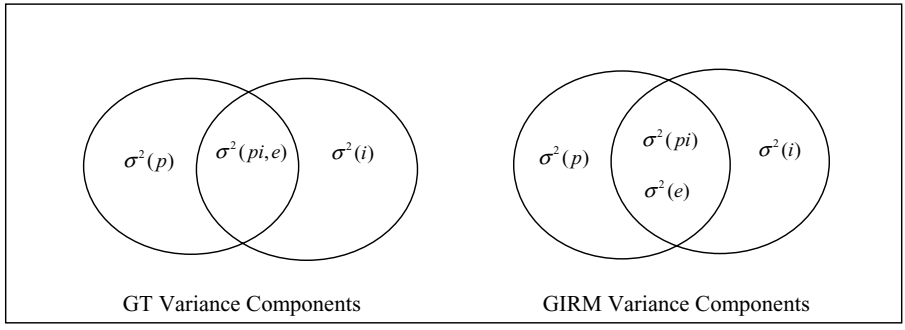

FIGURE 5. *Differences in variance component decomposition in GT and GIRM.*

TABLE 2
Sensitivity of GIRM Estimates to Normal Distributional Assumptions

			Generating Distribution for Item Difficulty					
			Normal		Uniform		Beta	
			GIRM	GT	GIRM	GT	GIRM	GT
Generating Distribution for Person Ability	Normal	$\hat{\sigma}^2(p)$	.029	.029	.030	.030	.036	.036
		$\hat{\sigma}^2(i)$	.043	.045	.055	.058	.005	.004
		$\hat{\sigma}^2(pi)$	.002		.003		.000	
		$\hat{\sigma}^2(e)$	.174		.162		.201	
		$\hat{\sigma}^2(pi, e)$	.176	.177	.165	.166	.201	.202
		$E\hat{\rho}^2$	.767	.766	.784	.782	.782	.781
		$\hat{\Phi}$	.726	.723	.731	.727	.778	.778
	Uniform	$\hat{\sigma}^2(p)$	.044	.045	.037	.037	.053	.054
		$\hat{\sigma}^2(i)$	.036	.037	.046	.047	.006	.006
		$\hat{\sigma}^2(pi)$	.003		.003		.001	
		$\hat{\sigma}^2(e)$	.167		.162		.182	
		$\hat{\sigma}^2(pi, e)$	.170	.170	.165	.166	.183	.183
		$E\hat{\rho}^2$	.839	.840	.818	.818	.854	.854
		$\hat{\Phi}$	.812	.812	.779	.778	.850	.850
	Gamma	$\hat{\sigma}^2(p)$	.018	.017	.023	.022	.029	.028
		$\hat{\sigma}^2(i)$	.035	.036	.042	.043	.007	.006
		$\hat{\sigma}^2(pi)$	.002		.003		.000	
		$\hat{\sigma}^2(e)$	.168		.171		.212	
		$\hat{\sigma}^2(pi, e)$	.170	.172	.174	.175	.212	.214
		$E\hat{\rho}^2$	.672	.665	.721	.714	.728	.724
		$\hat{\Phi}$	.630	.621	.676	.667	.722	.719

Note: GIRM estimates based on specification of normal prior distributions for person and item difficulty parameters.

computed by averaging across the rows of the expected response matrix. To quantify the variability in $\hat{\sigma}^2(i)$ we would want to allow the columns of the expected response matrix to vary while holding the rows constant, when in fact, for each iteration of the Markov Chain both columns and rows vary together. The problem described above applies equally to the GT approach, but will typically be less visible because GT, as implemented using the software GENOVA, provides estimates for standard errors using a closed-form formula under the assumption that score effects are normally distributed.

One clever approach that would at least partially adjust for the underestimation of posterior uncertainty was proposed by an anonymous reviewer of this article. In this approach, instead of estimating variance components by applying GT to the matrix of expected responses at each step of the MCMC chain, one would use the IRT parameters at each step of the chain to simulate a new matrix of item responses, and then apply GT to this predictively simulated matrix. In other words,

for the GIRM approach described here, GT is applied to $E[X_{pi}|\theta_p^{(m)}, \beta_i^{(m)}]$. In the xsalternative approach, GT would be applied to $X_{pi}^{(m)} \sim Rasch(\theta_p, \beta_i)$. So in this case the difference between GT and GIRM would be that the former operates on the observed X_{pi}, while the latter operates on the posterior predictive $X_{pi}^{(m)}$. Implementing this and other potential adjustments to the way that uncertainty is quantified in the GIRM approach is outside the scope of this current article, but clearly such adjustments would be possible, and this is something we plan to explore in a subsequent study.

We make no claims at this point that adjustments to the estimation of standard errors within the GIRM approach necessarily provide any clear advantage relative to other approaches being taken. We also note that the appealing feature of obtaining posterior distributions for variance components is not unique to GIRM, but a feature of taking a Bayesian estimation approach using MCMC methods. An application of MCMC methods to the estimation of GT variance components on the basis of observed scores has previously been presented by Mao, Shin, and Brennan (2005).

Sensitivity of GIRM Results to Distributional Assumptions

The comparisons above between the GIRM and GT approaches were based on an ideal scenario in which the distributions of the θ_p and β_i parameters governing the item response simulation matched the specification of prior distributions for the GIRM item response function. To test the sensitivity of the GIRM approach to the specification of prior distributions, we now hold fixed the prior specifications of the GIRM approach while varying the θ_p and β_i distributions from which item parameters were initially generated. The fixed and varying conditions of the simulations are as follows:

FIXED CONDITIONS

1. 500 respondents.
2. 20 items.[7]
3. One parameter item response function.
4. Prior distributions drawn from standard normal distribution.

VARYING CONDITIONS

1. Distribution from which β_i is drawn before simulating item responses.
   a. Normal (0, 1) [same as that assumed by GIRM]
   b. Uniform (−2, 2)
   c. Beta (.5, .5)
2. Distribution from which θ_p is drawn before simulating item responses.
   a. Normal (0, 1) [same as that assumed by GIRM]
   b. Uniform (−2, 2)
   c. Gamma (1)

We consider three distributional conditions, respectively, for β_i and θ_p. This leads to the simulation of nine different 500 by 20 matrices of item responses, to which we

apply both the GIRM and GT approaches. Table 2 compares the resulting variance component and generalizability coefficient estimates.

The comparison in the upper left corner of Table 2 represents the case where the distributional assumptions of the GIRM approach match the way item responses were actually generated. The GIRM and GT results are virtually identical. The three comparisons in the middle to upper left portions of Table 2 (normal/uniform, uniform/normal, and uniform/uniform) are all examples where we might expect the misspecification of GIRM prior distributions to be fairly mild, as the normal distribution might not be a bad approximation for a uniform distribution. Here we again find no noticeable difference in the results produced under either the GIRM or GT approaches. Of greatest interest are the comparisons in Table 2 under the beta column and the gamma row, as these skewed distributions can be expected to be poorly approximated by a normal distribution. However, as the results indicate, the estimated variance component and generalizability coefficients remain consistent with those that are estimated under GT. While the misspecification of prior distributions as normal does affect the ability of the GIRM approach to correctly estimate parameters for β_i and θ_p (results not shown here), it does not seem to have an effect on the ability to the GIRM approach to produce estimates consistent with those produced by GT for variance components and generalizability coefficients.

Sensitivity of GIRM Results to Specification of Item Response Function

In the simulations above, there has been a perfect match between the item response function used to simulate item responses and the item response function used to estimate parameters under the GIRM approach. In both cases, a Rasch model was used. To test the sensitivity of mis-specifying the item response function in GIRM, we simulated data according to a 3PL model, and then applied the GIRM approach using the Rasch model and normal prior distributions for both item and person parameters. In simulating a 20 by 500 matrix of item response according to the 3PL model, discrimination parameters were sampled from a normal distribution with a mean of .9 and an *SD* of .2, location parameters were sampled from a normal distribution with mean of 0 and an *SD* of 1, and guessing parameters were fixed at .25.

The results from this simulation (not shown here) provide preliminary evidence that, at least in the context of this simple $p \times i$ design, the GIRM-based estimates of variance components appear to be robust to misspecification of the underlying item response function. Consistent with the findings from our previous simulations, when applied to data simulated using the 3PL model, both GIRM and GT produced virtually identical variance component and generalizability coefficient estimates.

The GIRM Approach with Empirical Data

We now present the GIRM approach in the context of an empirical data set with a $p \times i$ measurement design. The data are taken from a survey instrument known as the Colorado Learning Attitudes about Science Survey (CLASS), which was administered to 349 undergraduate students enrolled in an introductory physics course at the University of Colorado. The intent of the CLASS instrument is to measure student beliefs about learning physics. In taking the CLASS, students are given 36

items in the form of statements such as "A significant problem in learning physics is being able to memorize all the information I need to know." Students select responses to these items along a five-category Likert scale from strongly disagree to strongly agree. These responses are subsequently scored dichotomously according to whether the response is in the same direction as that which would be expected from a practicing physicist (i.e., an "expert"). For details on the design and development of the CLASS, see Adams, Perkins, Dubson, Finkelstein, and Wieman (2005).

Both GT and the GIRM approach were applied to the CLASS data. Consistent with the GIRM approach taken in our simulations, we specified a Rasch item response model with standard normal item and person parameter distributions, and MCMC estimation with a burn-in of 1,000 and a chain of length of 10,000. There were 50 observations (less than 1%) in the 349 $\times$ 36 data matrix that were missing. This had no impact on the GIRM approach, as variance component estimation is based upon the expected, rather than the observed response matrix. The presence of missing data did require some adjustment to the GT approach; namely, an unbalanced random effects design had to be specified, with the software urGENOVA (Brennan, 2001b) used in place of GENOVA (for details see Brennan, 2001a, pp. 215–247).

The results from the GIRM and traditional GT analysis are presented in Table 3 and Figure 6. The results shown in Table 3 are expressed in the language of GT, and include the variance components and generalizability coefficients associated with the CLASS instrument. Consistent with the results from our simulations, when the observed responses from the CLASS data were analyzed using traditional GT, the estimated variance components were virtually identical. The results shown in Figure 6 are expressed in the language of IRT, and constitute what Wilson (2005) calls a "Wright Map," with estimates of student ability and item difficulty placed on a common logit scale. The Wright Map provides a graphical presentation of the probabilistic relationship between the location of student beliefs along a latent continuum relative to the location (i.e., difficulty) of the items to which they have responded.

Here are some statements about the CLASS instrument that can be made based on the GT results from Table 3:

- When total variance for a single person by item score combination X_{pi} is decomposed into person, item, person by item, and random error components, we see that the largest proportion of this variability, 77%, can be attributed to the residual term. About 10% of total variability can be attributed to item variability. The proportion of variability attributable to person by item interactions (.4%) appears to be negligible. Note that this result would not otherwise be obtained as part of an IRT analysis.
- If student mean scores for the CLASS instrument are computed from the 36 items analyzed here, the respective generalizability coefficients for relative and absolute decisions based on these observed scores are .85 and .83. This suggests that these scores have fairly high generalizability over the full universe of items that could have been used for the instrument. On the other hand, if the instrument designers were to use only half as many items for their survey

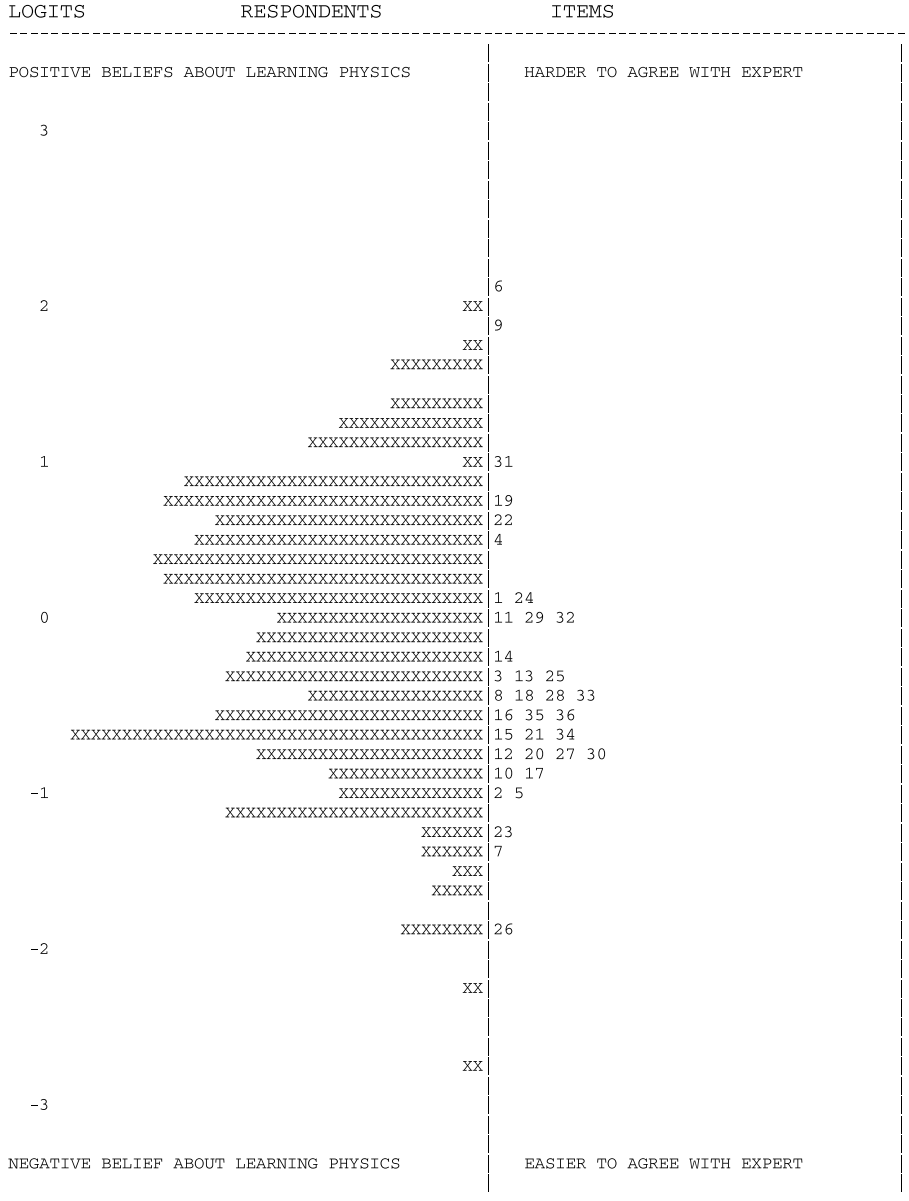

FIGURE 6. *A Wright map for the CLASS data.*

(perhaps because of time constraints), then these generalizability coefficients would decrease to .74 and .71.

Here are some statements about the CLASS instrument that can be made based on the IRT results from Figure 6:

TABLE 3
Variance Component Estimates in GIRM and GT for CLASS Instrument

Variance Components/G Coefficients	GIRM Mean of Posterior Distribution	GT Point Estimate
$\hat{\sigma}^2(p)$	.030	.031
$\hat{\sigma}^2(i)$	.025	.025
$\hat{\sigma}^2(pi)$	.001	NA
$\hat{\sigma}^2(e)$	.190	NA
$\hat{\sigma}^2(pi, e)$	.001 + .190	.192
$E\hat{\rho}^2$	.851	*
$\hat{\Phi}$	.834	*

*Estimates for these terms are not available as part of urGENOVA, but are otherwise derivable (cf., Brennan, 2001a, pp. 227–240).

- It is relatively easy for students taking the CLASS to agree with the expert belief about learning physics on the majority of the survey items. That is, 25 out of 36 are closely clustered together with item difficulties between 0 and −1 logits. On the other hand, there are three items (6, 9, and 31) that are clearly quite hard for students to agree with, and three items (23, 7, and 26) that are clearly quite easy to agree with. An important question to ask would be whether this result had been hypothesized by the designers of the CLASS instrument.
- A large part of the distribution of student attitudes is between 0 and 2 logits, yet there are relatively few items located in this range. Consequently, the beliefs for students in this part of the distribution will be measured less precisely than for students between 0 and −1 logits.

In this empirical example, two more features of the GIRM approach come into clearer focus. First, the approach does not require any sort of reformulation in the presence of missing data. While GT can handle designs with missing data, changes in the underlying model used to estimate the relevant G and D Study variance components can become increasingly complex. Second, the GIRM approach provides an analyst with estimates for both GT and IRT parameters. This enables the analyst to interpret a measurement procedure both in terms of the sampling component typically associated with GT, and the scaling component typically associated with IRT.

It is worth noting that in this particular example, where only a single rather than a multifacet design was specified, some of the unique information available through a GT analysis may not be as evident. GT is particularly useful when one is speculating about the optimal design for a measurement procedure in terms of the quantity of facet conditions. IRT results can be quite complementary from the standpoint of making decisions about optimal measurement design. IRT may not readily communicate the same sort of information about the optimal *quantity* of facet conditions, but it can be used to provide useful information about the *quality* of the facet conditions.

Discussion

Our aim in this article has been to introduce the GIRM approach as what we hope will be a further step toward building a crosswalk between IRT and GT. We

have shown that this approach, situated within an item response modeling framework and the context of a simple measurement design, will lead to the same estimates of variance components and generalizability as would be reached under the more traditional GT approach. We have provided some evidence that the approach appears to be robust to misspecification of distributional assumptions and the form of the item response function.

We have pointed out three advantages to the GIRM approach:

1. Although there is no model-independent decomposition of *pi* and *e*, conditional on assumptions about the distribution of the error term, variance components for error and facet interaction effects can be estimated separately.
2. Because estimation of variance components under the GIRM approach is done as a function of expected rather than observed responses, all measurement designs can be treated as if they were complete and balanced.
3. The results from both a GT and IRT analysis are available within a single modeling framework.

It may well be that it is the last of these advantages that is the most practically important. We can at this point only speculate about one other potential advantage of the GIRM approach that merits further exploration. A wide variety of measurement designs can be easily incorporated into item response functions in the GIRM approach, and we suspect that many of these designs cannot be so easily expressed in GT notation. For example, it does not appear that standard GT notation and variance component estimation can accommodate a design in which a group of raters score both a common set of items along with a unique set specific to each rater. But this is something that can be set up relatively easily using IRT (cf. Wilson & Hoskens, 2001). If our suspicions are correct, then there will be a class of measurement designs for which GIRM can provide answers that would elude GT (although, of course, advances in GT may render such judgments untrue in the future).

This is, of course, the critical question we have yet to address: when would one expect the GIRM approach to give results that differ from those that could be readily obtained through the application of traditional GT? The integration of IRT and GT that we have proposed here mainly addresses the estimation of G Study variance components for a single-facet design. Many other GT topics such as multifacet and multivariate designs have not been addressed in detail. Hence, a next step would be to extend and compare the GIRM approach in the context of more complex measurement designs, in particular to designs where the use of GT may not be workable. These designs are likely to be multifaceted and unbalanced with polytomous item responses. Indeed, previous research by Verhelst and Verstralen (2001) and Patz et al. (2002) in the context of repeated ratings on performance assessments has led to a number of useful insights with respect to the connections between IRT, GT, and the broader framework of hierarchical random effects models, insights we have built upon in this article.

The GIRM approach comes with clear limitations, some of which we have alluded to throughout this article. A technical limitation of the GIRM approach is the lack of specialized software. The software WinBUGS 1.4 was used to estimate posterior

distributions for GT parameters. While writing WinBUGS code is fairly straightforward for a simple item response function and dichotomous data, it can become more difficult when using more complex item response models and polytomous data. In addition, running the GIRM approach through WinBUGS can be time-consuming. Estimating posterior distributions for variance components for a 500 person by 20 item matrix took approximately four hours using a standard desktop computer.[8] All this should come as little surprise, since the WinBUGS software is relatively new, and was not designed with IRT or GT extensions in mind. As WinBUGS becomes more developed, and/or as specialized MCMC estimation routines are written in flexible programs such as *R*, there should be fewer technical limitations to the implementation of the GIRM approach. Another limitation we have discussed is that the variance component posterior distributions produced in the GIRM approach as currently implemented using WinBUGS will be biased, tending to underestimate spread in the distributions. Finally, the GIRM approach requires a stronger set of assumptions than those required for GT.

From a theoretical standpoint, the assumptions of GIRM are not for the faint of heart. GIRM shares with GT the assumption that persons and facets of measurement are sampled independently from populations of persons and facets. The justifications offered for such assumptions made in applied contexts are often not very compelling, if they are offered at all. For instance, in the example with the CLASS instrument presented above, the IRT results in Figure 6 indicated that were the instrument to be revised it might be sensible to include more difficult items so as to obtain more precise estimates for respondents with more positive beliefs about physics. But if the items for the revised instrument were selected in this way, they would clearly not represent a random (or exchangeable) sample from the universe of possible items, however loosely the latter is conceptualized. It should be clear from this example that there are instances when random sampling assumptions are not appropriate, and where the notion of generalizability, at least as it is conceptualized in GT, is not compatible with IRT. But this is a topic for another article.

One justification for the potentially dubious sampling assumptions in GIRM (and these include the assumptions common to GT), is that these assumptions are no worse than the sorts of assumptions in classical test theory or other inferential statistical models (Brennan, 2001a, pp. 171–174). Unfortunately, such justifications do little to reassure us that the GIRM and GT approaches are robust to violations of their various independence assumptions. Research that explores this issue would go a long way toward determining whether the GIRM approach presented here is a sensible model in applied contexts.

In addition to the sampling assumptions of GT, GIRM adds the standard assumptions of IRT, and distributional assumptions about measurement facets. The simulations described in this article suggest that with respect to the estimation of GT variance components, the GIRM approach is robust to the misspecification of prior distributions and the parametric form of the item response function. Future research should continue to rigorously examine these findings, and also explore the sensitivity of the GIRM approach to other sorts of misspecifications, such as violations of the assumptions of unidimensionality and local independence.

Acknowledgments

The authors would like to acknowledge Chiang Liu and Laik Teh Woo for their research assistance on this project. We thank Ed Wiley and Rianne Jansen for helpful comments in the preparation of this manuscript. We thank especially the three anonymous reviewers of this manuscript for their many insightful comments.

Notes

[1]This assumption is made to simplify the comparison with the GIRM approach that follows. Both GT and GIRM can also be applied to unbalanced designs.

[2]Brennan (2001a, p. 24) points out that, for the product of most score effects, this "assumption" is really just a consequence of how score effects have been defined as random effects in the model.

[3]A detailed explication of this approach is beyond the scope of this article. For details on Bayesian analytical methods and MCMC estimation, see Gelman et al. (2004); for details on the use of Metropolis Hastings within Gibbs for IRT models, see Patz & Junker (1999).

[4]GT can be relatively easily implemented using the software GENOVA (Crick & Brennan, 1983).

[5]Our initial simulation only included five items to accomodate the computer processing demands of 100 replications of the GIRM approach.

[6]The GIRM approach is implemented using the software WinBUGS (Spiegelhalter et al., 2004) invoked out of the R programming environment with the function "bugs" and the package R2WinBUGS (Sturtz, Ligges, & Gelman, 2005). The relevant code is available from the first author upon request.

[7]We use 20 items instead of five to address a criticism raised by one reviewer of this article who suggested that the application of either GIRM or GT to a measurement design with only 5 items was unrealistic.

[8]The desktop computer used was a PC Pentium 4, 3 GHz CPU, 1 GB RAM.

References

Adams, R. (2006). *Reliability as a measurement design effect*. Paper presented at the 2006 bi-annual meeting of the Institute for Objective Measurement Workshop, Berkeley, CA.

Adams, W. K., Perkins, K. K., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2005). A new instrument for measuring student beliefs about physics and learning physics: The Colorado learning attitudes about science survey submitted for publication to physical review special topics PER. Retrieved from http://class.colorado.edu/ on May 30, 2006.

Betebenner, D. W. (1998). *Improved confidence interval estimation for variance components and error variances in generalizability theory*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Bock, D., Brennan, R., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, *26*(4), 364–375.

Brennan, R. (2001a). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. (2001b). *Manual for urGENOVA*. Iowa City, IA: Iowa Testing Programs, University of Iowa.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, *27*, 907–949.

Crick, J. E. & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin No. 43). Iowa City, IA: ACT, Inc.

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

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, *14*, 191–203.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.) Boca Raton, FL: Chapman & Hall/CRC.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Green, B., Bock, R., Humphreys, L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*(4), 577–601.

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

Mao, X., Shin, D., & Brennan, R. (April, 2005). *Estimating the variability of the estimated variance components and related statistics using the MCMC procedure: An exploratory study*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.

Patz, R., & Junker, B. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178.

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.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: John Wiley.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2004). WinBUGS 1.4 (available at http://www.mrc-bsu.cam.ac.uk/bugs).

Sturtz, S., Ligges, U. & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, *12*(3). Available online at http://www.jstatsoft.org/v12/i03/v12i03.pdf.

Van den Noortgate, W., & Paek, I. (2004). Person regression models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 167–187). New York: Springer.

Verhelst, N., & Verstralen, H. (2001). An IRT model for multiple raters. In A. Boomsma, M. van Duijn, & T. Snijders (Eds.), *Essays on item response theory* (pp. 88–108). New York: Springer-Verlag.

Wiley, E. W. (2001). *Bootstrap strategies for variance component estimation: Theoretical and empirical results*. Doctoral dissertation, Stanford University.

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

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306.

Authors

DEREK C. BRIGGS is an Assistant Professor, School of Education, University of Colorado, 249 UCB, Boulder, CO 80309; derek.briggs@colorado.edu. His primary research interests include causal inference, educational statistics, psychometrics, and validity theory.

MARK WILSON is a Professor, Graduate School of Education, University of California, Berkeley, CA 94720; MarkW@berkeley.edu. His primary research interests include educational statistics, experimental design in education, psychometrics, quantitative methods, research methods, and testing.