Chapter 1

Generalizability Theory and Item Response Theory

Cees A.W. Glas

Abstract Item response theory is usually applied to items with a selected-response format, such as multiple choice items, whereas generalizability theory is usually applied to constructed-response tasks assessed by raters. However, in many situations, raters may use rating scales consisting of items with a selected-response format. This chapter presents a short overview of how item response theory and generalizability theory were integrated to model such assessments. Further, the precision of the estimates of the variance components of a generalizability theory model in combination with two- and three-parameter models is assessed in a small simulation study.

Keywords: Bayesian estimation, item response theory, generalizability theory, Markov chain Monte Carlo

Introduction

I first encountered Piet Sanders when I started working at Cito in 1982. Piet and I came from different psychometric worlds: He followed generalizability theory (GT), whereas I followed item response theory (IRT). Whereas I spoke with reverence about Gerhard Fischer and Darrell Bock, he spoke with the same reverence about Robert Brennan and Jean Cardinet. Through the years, Piet invited all of them to Cito, and I had the chance to meet them in person. With a slightly wicked laugh, Piet told me the amusing story that Robert Brennan once took him aside to state, “Piet, I have never seen an IRT model work.” Later, IRT played an important role in the book *Test Equating, Scaling, and Linking* by Kolen and Brennan (2004). Piet’s and my views converged over time. His doctoral thesis “The Optimization of Decision Studies in Generalizability Theory” (Sanders, 1992) shows that he was clearly inspired by optimization approaches to test construction from IRT.

On January 14 and 15, 2008, I attended a conference in Neuchâtel, Switzerland, in honor of the 80th birthday of Jean Cardinet, the main European theorist of GT. My presentation was called “The Impact of Item Response Theory in Educational Assessment:

A Practical Point of View” and was later published in *Mesure et Evaluation en Education* (Glas, 2008). I remember Jean Cardinet as a very friendly and civilized gentleman. But he had a mission:

It soon became clear that he wanted to show the psychometric world that GT was the better way and far superior to modernisms such as IRT. I adapted my presentation to show that there was no principled conflict between GT and IRT, and that they could, in fact, be combined. Jean seemed convinced. Below, I describe how IRT and GT can be combined. But first I shall present some earlier attempts of analyzing rating data with IRT.

Some History

Although in hindsight the combination of IRT and GT seems straightforward, creating the combination took some time and effort. The first move in that direction, made by Linacre (1989, 1999), was not very convincing. Linacre considered dichotomous item scores given by raters. Let Y_{nri} be an item score given by a rater r ($r = 1, \dots, N_r$) on an item i ($i = 1, \dots, K$) when assessing student n ($n = 1, \dots, N$). Y_{nri} is equal to 0 or 1. Define the logistic function $\Psi(\cdot)$ as

$$\Psi(\tau) = \frac{\exp(\tau)}{1 + \exp(\tau)}. \quad (1)$$

Conditional on a person ability parameter θ_n , the probability of a positive item score is defined as $\Pr(Y_{nri} = 1 | \theta_n) = P_{nri} = \Psi(\tau_{nri})$, with

$$\tau_{nri} = \theta_n - \beta_i + \rho_r,$$

where β_i is an item parameter and ρ_r is a rater effect. The model was presented as a straightforward generalization of the Rasch model (Rasch, 1960); in fact, it was seen as a straightforward application of the linear logistic test model (LLTM) (Fischer, 1983). That is, the probability of the scores given to a respondent was given by

$$\prod_i \prod_r P_{nri}^{y_{nri}} (1 - P_{nri})^{1 - y_{nri}}.$$

In my PhD thesis, I argued that this is a misspecification, because the assumption of local independence made here is violated: The responses of the different raters are dependent because they depend on the response of the student (Glas, 1989).

Patz and Junker (1999) criticize Linacre's approach on another ground: LLTMs require that all items have a common slope or discrimination parameter; therefore, they suggest using the logistic model given in Equation (1) with the argument

$$\tau_{nri} = \alpha_i \theta_n - \beta_i + \rho_{ri},$$

where α_i is a discrimination parameter and ρ_{ri} stands for the interaction between an item and a rater. However, this does not solve the dependence between raters. Therefore, we consider the following alternative. The discrimination parameter is dropped for convenience; the generalization to a model with discrimination parameters is straightforward. Further, we assume that the students are given tasks indexed t ($t = 1, \dots, N_t$), and the items are nested within the tasks. A generalization to a situation where tasks and items are crossed is straightforward. Further, item i pertains to task $t(i)$. Consider the model given in Equation (1) with the argument

$$\tau_{nrti} = \theta_n + \delta_{nt(i)} - \beta_i + \rho_r.$$

The parameter $\delta_{nt(i)}$ models the interaction between a student and a task. Further, Patz and Junker (1999) define θ_n and $\delta_{nt(i)}$ as random effects, that is, they are assumed to be drawn from some distribution (i.e., the normal distribution). The parameters β_i and ρ_r may be either fixed or random effects.

To assess the dependency structure implied by this model, assume τ_{nrti} could be directly observed. For two raters, say r and s , scoring the same item i , it holds that $Cov(\tau_{nrti}, \tau_{nsti}) = Cov(\theta_n, \theta_n) + Cov(\delta_{nt(i)}, \delta_{nt(i)}) = \sigma_n^2 + \sigma_{nt}^2$. This also holds for two items related to the same task. If two items, say i and j , are related to the same task, that is, if $t(i) = t(j) = t$, then $Cov(\tau_{nrti}, \tau_{nstj}) = Cov(\theta_n, \theta_n) + Cov(\delta_{nt(i)}, \delta_{nt(j)}) = \sigma_n^2 + \sigma_{nt}^2$.

If items are related to different tasks, that is, if $t(i) \neq t(j)$, then $Cov(\tau_{nrti}, \tau_{nstj}) = \sigma^2$. So, σ_{nt}^2 models the dependence of item responses within a task.

Combining IRT and GT

The generalization of this model to a full-fledged generalizability model is achieved through the introduction of random main effects for tasks π_t , random effects for the interaction between students and raters γ_{nr} , and students and tasks η_{tr} . The model then becomes the logistic model in Equation (1) with the argument

$$\tau_{nri} = \theta_n - \beta_i + \rho_r + \pi_t + \delta_{nt(i)} + \gamma_{nr} + \eta_{tr}.$$

The model can be conceptualized by factoring it into a measurement model and structural model, that is, into an IRT measurement model and a structural random effects analysis of variance (ANOVA) model. Consider the likelihood function

$$\prod_i \prod_r P_{nri}^{y_{nri}}(\tilde{\tau}_{nri(i)}) (1 - P_{nri}(\tilde{\tau}_{nri(i)}))^{1-y_{nri}} N(\tilde{\tau}_{nri(i)})$$

where

$$\tilde{\tau}_{nri} = \theta_{ni} + \rho_r + \pi_t + \delta_{nt} + \gamma_{nr} + \eta_{tr} \quad (2)$$

is a sum of random effects, $P_{nri}(\tilde{\tau}_{nri(i)})$ is the probability of a correct response given $\tilde{\tau}_{nri}$ and the item parameter β_i , and $N(\tilde{\tau}_{nri(i)})$ is the density of $\tilde{\tau}_{nri}$, which is assumed to be a normal density. If the distribution of $\tilde{\tau}_{nri}$ is normal, the model given in Equation (2) is completely analogous to the GT model, which is a standard ANOVA model.

This combination of IRT measurement model and structural ANOVA model was introduced by Zwinderman (1991) and worked out further by Fox and Glas (2001). The explicit link with GT was made by Briggs and Wilson (2007).

They use the Rasch model as a measurement model and the GT model—that is, an ANOVA model—as the structural model. The structural model implies a variance decomposition

$$\sigma_\tau^2 = \sigma_n^2 + \sigma_t^2 + \sigma_r^2 + \sigma_{nt}^2 + \sigma_{nr}^2 + \sigma_{tr}^2 + \sigma_e^2,$$

and these variance components can be used to construct the well-known agreement and reliability indices as shown in Table 1.

Table 1 Indices for Agreement and Reliability for Random and Fixed Tasks

Type of Assessment	Index
Random tasks, agreement	$\frac{\sigma_n^2}{\sigma_n^2 + \sigma_t^2 / N_t + \sigma_r^2 / N_r + \sigma_{nt}^2 / N_t + \sigma_{nr}^2 / N_r + \sigma_{tr}^2 / N_r N_t + \sigma_e^2 / N_r N_t}$
Random tasks, reliability	$\frac{\sigma_n^2}{\sigma_n^2 + \sigma_{nt}^2 / N_t + \sigma_{nr}^2 / N_r + \sigma_{tr}^2 / N_r N_t + \sigma_e^2 / N_r N_t}$
Fixed tasks, agreement	$\frac{\sigma_n^2 + \sigma_{nt}^2}{\sigma_n^2 + \sigma_{nt}^2 + \sigma_r^2 / N_r + \sigma_{nr}^2 / N_r + \sigma_{tr}^2 / N_r + \sigma_e^2 / N_r}$
Fixed tasks, reliability	$\frac{\sigma_n^2 + \sigma_{nt}^2}{\sigma_n^2 + \sigma_{nt}^2 + \sigma_r^2 / N_r + \sigma_e^2 / N_r}$

Note: N_t = number of tasks; N_r = number of raters.

Parameter Estimation

The model considered here seems quite complicated; however, conceptually, estimation in a Bayesian framework using Markov chain Monte Carlo (MCMC) computational methods is quite straightforward. The objective of the MCMC algorithm is to produce samples of the parameters from their posterior distribution. Fox and Glas (2001) developed a Gibbs sampling approach, which is a generalization of a procedure for estimation of the two-parameter normal ogive (2PNO) model by Albert (1992). For a generalization of the three-parameter normal ogive (3PNO) model, refer to Béguin and Glas (2001). Below, it will become clear that to apply this approach, we first need to reformulate the model from a logistic representation to a normal-ogive representation. That is, we assume that the conditional probability of a positive item score is defined as $\Pr(Y_{nrti} = 1 | \tau_{nrti}) = P_{nrti} = \Phi(\tau_{nrti})$, where $\Phi(\cdot)$ is the cumulative normal distribution, i.e.,

$$\Phi(\tau) = (2\pi)^{-1/2} \int_{-\infty}^{\tau} \exp(-t^2 / 2) dt.$$

In the 3PNO model, the probability of a positive response is given by

$$P_{nrti} = \gamma_i + (1 - \gamma_i) \Phi(\tau_{nrti})$$

where γ_i is a guessing parameter.

Essential to Albert's approach is a data augmentation step (Tanner & Wong, 1987), which maps the discrete responses to continuous responses. Given these continuous responses, the posterior distributions of all other parameters become the distributions of standard regression models, which are easy to sample from. We outline the procedure for the 2PNO model. We augment the observed data Y_{nrti} with latent data Z_{nrti} , where Z_{nrti} is a truncated normally distributed variable, i.e.,

$$Z_{nrti} | Y_{nrti} \sim \begin{cases} N(\tau_{nrti}, 1) \text{ truncated at the left by } 0 & \text{if } Y_{nrti} = 1 \\ N(\tau_{nrti}, 1) \text{ truncated at the right by } 0 & \text{if } Y_{nrti} = 0. \end{cases} \quad (3)$$

Note that this data augmentation approach is based on the normal-ogive representation of the IRT model, which entails the probability of a positive response is equal to the probability mass left from the cut-off point τ_{nrti} .

Gibbs sampling is an iterative process, where the parameters are divided into a number of subsets, and a random draw of the parameters in each subset is made from its posterior distribution given the random draws of all other subsets. This process is iterated until convergence. In the present case, the augmented data Z_{nrti} are drawn given starting values of all other parameters using Equation (3). Then the item parameters are drawn using the regression model $Z_{nrti} = \tilde{\tau}_{nrt} - \beta_i + \varepsilon_{nrti}$, with $\tilde{\tau}_{nrt} = \theta_n + \rho_r + \pi_t + \delta_{nt} + \gamma_{nr} + \eta_{tr}$ where all parameters except β_i have normal priors. If discrimination parameters are included, the regression model becomes $Z_{nrti} = \alpha_i \tilde{\tau}_{nrt} - \beta_i + \varepsilon_{nrti}$.

The priors for β_i can be either normal or uninformative, and the priors for α_i can be normal, lognormal, or confined to the positive real numbers. Next, the other parameters are estimated using the standard ANOVA model $Z_{nrti} - \beta_i = \theta_n + \rho_r + \pi_t + \delta_{nt} + \gamma_{nr} + \eta_{tr} + \varepsilon_{nrti}$. These steps are iterated until the posterior distributions stabilize.

A Small Simulation Study

The last section pertains to a small simulation to compare the use of the 2PNO model with the use of the 3PNO model. The simulation is related to the so-called bias-variance trade-off. When estimating the parameters of a statistical model, the mean-squared error (i.e., the mean of the squared difference between the true value and the estimates over replications of the estimation procedure) is the sum of two components: the squared bias and the sampling variance (i.e., the squared standard error). The bias-variance trade-off pertains to the fact that, on one hand, more elaborated models with more parameters tend to reduce the bias, whereas on the other hand, adding parameters leads to increased standard errors. At some point, using a better fitting, more precise model may be counterproductive because of the increased uncertainty reflected in large standard errors. That is, at some point, there are not enough data to support a too elaborate model.

In this simulation, the 3PNO model is the elaborate model, which may be true but hard to estimate, and the 2PNO model is an approximation, which is beside the truth but easier to estimate. The data were simulated as follows. Sample sizes of 1,000 and 2,000 students were used. Each simulation was replicated 100 times. The test consisted of five tasks rated by two raters both scoring five items per task. Therefore, the total number of item responses was 50, or 25 for each of the two raters. The responses were generated using the 3PNO model. For each replication, the item location parameters β_i were drawn from a standard normal distribution, the item discrimination parameters α_i were drawn from a normal distribution with a mean equal to 1.0 and a standard deviation equal to 0.25, and the guessing parameters γ_i were drawn from a beta distribution with parameters 5 and 20. The latter values imply an average guessing parameter equal to 0.25. These distributions were also used as priors in the estimation procedure.

The used variance components are shown in the first column of Table 2. The following columns give estimates of the standard error and bias obtained over the 100 replications, using the two sample sizes and the 2PNO and 3PNO models, respectively.

In every replication, the estimates of the item parameters and the variance components were obtained using the Bayesian estimation procedure by Fox and Glas (2001) and Béguin and Glas (2001), outlined above. The posterior expectation (EAP) was used as a point estimate. Besides a number of variance components, the reliability ρ^2 for an assessment with

random tasks was estimated. The bias and standard errors for the reliability are given in the last row of Table 2.

Note that, overall, the standard errors of the EAPs obtained using the 2PNO model are smaller than the standard errors obtained using the 3PNO model. On the other hand, the bias for the 2PNO model is generally larger. These results are in accordance with the author's expectations.

Table 2 Comparing Variance Component Estimates for 2PNO and 3PNO Models

Variance		N = 1,000				N = 2,000			
Components/ Reliability Coefficient	True Values	2PNO	3PNO	2PNO	3PNO	2PNO	3PNO	2PNO	3PNO
		SE	Bias	SE	Bias	SE	Bias	SE	Bias
$\hat{\sigma}_n^2$	1.0	.0032	.0032	.0036	.0028	.0021	.0024	.0028	.0009
$\hat{\sigma}_{nt}^2$	0.2	.0027	.0024	.0033	.0022	.0023	.0021	.0021	.0010
$\hat{\sigma}_{nr}^2$	0.2	.0043	.0039	.0054	.0036	.0022	.0036	.0043	.0027
$\hat{\sigma}_{tr}^2$	0.2	.0056	.0041	.0066	.0033	.0036	.0047	.0046	.0039
$\hat{\sigma}_\varepsilon^2$	0.2	.0047	.0015	.0046	.0014	.0028	.0012	.0037	.0012
ρ^2	0.85	.0396	.0105	.0401	.0106	.0254	.0101	.0286	.0104

Note: 2PNO = two-parameter normal ogive; 3PNO = three-parameter normal ogive; SE = standard error

Conclusion

This chapter showed that psychometricians required some time and effort to come up with a proper method for analyzing rating data using IRT. Essential to the solution was the distinction between a measurement model (i.e., IRT) and a structural model (i.e., latent linear regression model). The parameters of the combined measurement and structural models can be estimated in a Bayesian framework using MCMC computational methods.

In this approach, the discrete responses are mapped to continuous latent variables, which serve as the dependent variables in a linear regression model with normally distributed components. This chapter outlined the procedure for dichotomous responses in combination with the 2PNO model, but generalizations to the 3PNO model and to models for polytomous responses—e.g., the partial credit model (Masters, 1982), the generalized partial credit model

(Muraki, 1992), the graded response model (Samejima, 1969), and the sequential model (Tutz, 1990)—are readily available (see, for instance, Johnson & Albert, 1999).

However, nowadays, developing specialized software for combinations of IRT measurement models and structural models is no longer strictly necessary. Many applications can be created in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2004). Briggs and Wilson (2007) give a complete WinBUGS script to estimate the GT model in combination with the Rasch model. Although WinBUGS is a valuable tool for the advanced practitioner, it also has a drawback that is often easily overlooked: It is general-purpose software, and the possibilities for evaluation of model fit are limited.

Regardless, the present chapter may illustrate that important advances in modeling data from rating have been made over the past decade, and the combined IRT and GT model is now just another member of the ever-growing family of latent variable models (for a nice family picture, see, for instance, Skrondal & Rabe-Hesketh, 2004).

References

- Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics*, 17, 251-269.
- Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66, 541-562.
- Briggs, D.C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, 44, 131-155.
- Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, 48, 3-26.
- Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66, 271-288.
- Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models*. Unpublished PhD thesis, Enschede, University of Twente.
- Glas, C. A. W. (2008). Item response theory in educational assessment and evaluation. *Mesure et Evaluation en Education*, 31, 19-34.
- Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.
- Linacre, J. M. (1999). FACETS (Version 3.17) [Computer software]. Chicago: MESA Press.

- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Patz, R. J., & Junker, B. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342-366.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17*.
- Sanders, P. F. (1992). *The optimization of decision studies in generalizability theory*. Doctoral thesis, University of Amsterdam.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2004). WinBUGS 1.4. Retrieved from <http://www.mrc-bsu.cam.ac.uk/bugs>
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation [with discussion]. *Journal of the American Statistical Association*, 82, 528-540.
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39-55.
- Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, 56, 589-600.