4-2019

Considerations in S-χ2: Rest Score or Summed Score, Priors, and Violations of Normality

Christine E. DeMars
demarsce@jmu.edu

Derek Sauder
James Madison University

Recommended Citation

DeMars, C. E. & Sauder, D. (2019, April). Considerations in S-χ2: Rest score or summed score, priors, and violations of normality. Electronic poster presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

Considerations in S-χ^2: Rest Score or Summed Score, Priors, and Violations of Normality

Christine E. DeMars

Derek Sauder

James Madison University

DeMars, C. E. & Sauder, D. (2019, April*). Considerations in S-χ^2: Rest score or summed score, priors, and violations of normality.* Electronic poster presented at the annual meeting of the National Council on Measurement and Education, Toronto, CA.

Abstract

The S-χ^2 item fit index is one of the few item fit indices that appears to maintain accurate Type I error rates. This study explored grouping examinees by the rest score or summed score, prior distributions for the item parameters, and the shape of the ability distribution. Type I error was slightly closer to the nominal level for the total-score S-χ^2 for the longest tests, but power was higher for the rest-score S-χ^2 in every condition where power was < 1. Prior distributions reduced the proportion of estimates with extreme standard errors but slightly inflated the Type I error rates in some conditions. When the ability distribution was not normally distributed, integrating over an empirically-estimated distribution yielded Type I error rates closer to the nominal value than integrating over a normal distribution.

Considerations in S-χ^2: Rest Score or Summed Score, Priors, and Violations of Normality

The S-χ^2 item fit index (Orlando & Thissen, 2000) is one of the few item fit indices that appears to follow a standard distribution and thus maintains accurate Type I error rates at the nominal α (Chon & Sinharay, 2014; Glas & Suárez Falcón, 2003; Orlando & Thissen, 2000; Orlando & Thissen, 2003) without the need for bootstrapping the probability distribution. The S-χ^2 has been included in several item response theory (IRT) estimation packages, including Flexmirt, IRTPRO, and the mirt package in R. However, several issues have not been explored in the published literature, including whether the analyst should group examinees by the rest score or summed score, how prior distributions for the item parameters may impact the degrees of freedom, and whether the shape of the ability (θ) distribution must be taken into account or if it can be treated as normal.

To calculate S-χ^2, Orlando and Thissen (2000; 2003) grouped examinees by total summed score. Within Flexmirt (2017), examinees are instead grouped by rest score[1], the summed score not including the item for which the index is calculated. This is labeled the Orlando-Thissen-Bjorner index.

For the Orlando-Thissen index, for each summed score *k*, the number of examinees expected (based on the estimated item parameters) to answer the item correctly is estimated through:

$$E_{ik} = \frac{\int T_i S_{k-1}^{*i} \phi(\theta) \partial(\theta)}{\int S_k \phi(\theta) \partial(\theta)}, \quad (1)$$

[1] This is not documented in the user manual, but is obvious when the tables are printed to the output and was confirmed in a personal communication from the Flexmirt support desk, April 25, 2018. Cai (2015) used the rest score in an extension of the S-χ^2 to polytomous items in a hierarchical multidimensional model.

where T_i is the expected probability of correct response as a function of θ, S_k is the likelihood function for summed score *k*, S_{k-1}^{*i} is the likelihood function for the rest score omitting item *i*, and ϕ(θ) is the density of θ, which could be assumed to be normal or estimated through empirical histograms or other methods.

For the Orlando-Thissen-Björner index, the numerator in Equation 1 is the same but S_{k-1}^{*i} replaces S_k in the denominator. The examinees in each score group change from item to item.

After E_{ik} is estimated, $S-\chi^2 = \sum_{k=1}^{n-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1-E_{ik})}$, where N_k is the total number of students with score *k*, O_{ik} is the observed proportion correct for item *i* in score group *k*, and *n* is the number of items. If *k* is the summed score, scores of 0 and *n* must be omitted, but if *k* is the rest score the summation begins at 0. The statistical significance of S-χ^2 is assessed through a χ^2 test with df = # of score groups - # of parameters estimated for the studied item. If E_{ik} or 1 - E_{ik} < 1 scores are combined, so the degrees of freedom will vary across items.

Purpose

This study explored three research questions:

1) Do Type I error and power rates differ depending on whether examinees are grouped by total score or rest score? Using the rest score (Orlando-Thissen-Bjorner) instead of the total score (Orlando-Thissen) might provide slightly higher power when the studied item does not fit because the misfit does not contaminate the rest score.

2) Do prior distributions on the *a* and *c* parameters impact the Type I error rates? Although the parameters of the one and two parameter logistic (1PL and 2PL) models can often be estimated

well without imposing priors, priors are generally needed to obtain reasonable estimates and standard errors for the parameters of the 3PL model (Mislevy, 1986). But when priors are used, the so-called *free* parameters are not literally free. Does this make the nominal degrees of freedom too large and increase the Type I error rate?

3) When the ability distribution is not normal, does integrating over the estimated ability distribution yield different Type I error rates than integrating over a normal distribution? In calculating S-χ^2, one integrates over the ability distribution. Using the wrong ability distribution would obviously impact the estimated distribution of summed scores. But within score group *k*, there may be only a narrow range of ability where the relative likelihood is high for any pattern summing to *k*. Thus, the overall ability distribution may not be critical in the calculations.

Method

Following Orlando and Thissen (2000), three sample sizes (N = 500, 1000, 2000) were crossed with three test lengths (10, 40, or 80 items). The *b*-parameters were randomly selected from a N(0, 1) distribution, with any draws < -2 or > 2 replaced. The logs of the *a*-parameters were randomly selected from a N(0, $.35^2$) distribution, with any draws < -0.7 or > 0.7 replaced. The exponents were then multiplied by 1.7 to put the resulting *a*-parameters on a scale reasonable for the logit metric. All *c*-parameters were set to .2. We simulated 8000 items, divided across 800 10-item test forms, 200 40-item test forms, and 100 80-item test forms. Thus, the Type I error rates were averages across 8000 items because no single item was replicated.

Item parameters and the rest-score S-χ^2 were estimated in Flexmirt (2017). The item parameter estimates were read into the mirt() package (Chalmers, 2012) in R and fixed for the

calculation of the total-score S-χ^2, with the degrees of freedom adjusted to reflect that the item parameters were really estimated, not known.

The same simulated data sets were used for Research Questions 1 and 2. Examinee abilities were randomly simulated, ~N(0,1), for each replication. For Question 1, no priors were applied to the item parameters. The *c*-parameters were fixed to .2, because without priors many of the item parameter estimates were implausible. For Question 2, the prior for the *a*-parameters was logN(0,.5^2) and the prior for the *c*-parameters was β(21,81).

For Question 1, two additional runs were conducted for each replication, each including one misfitting item. Misfit 1 was the same as Bad Item 1 in Orlando and Thissen (2003). Misfit 2 added a sine curve to a 3PL item. The item response functions (IRF) are shown in Figure 1.

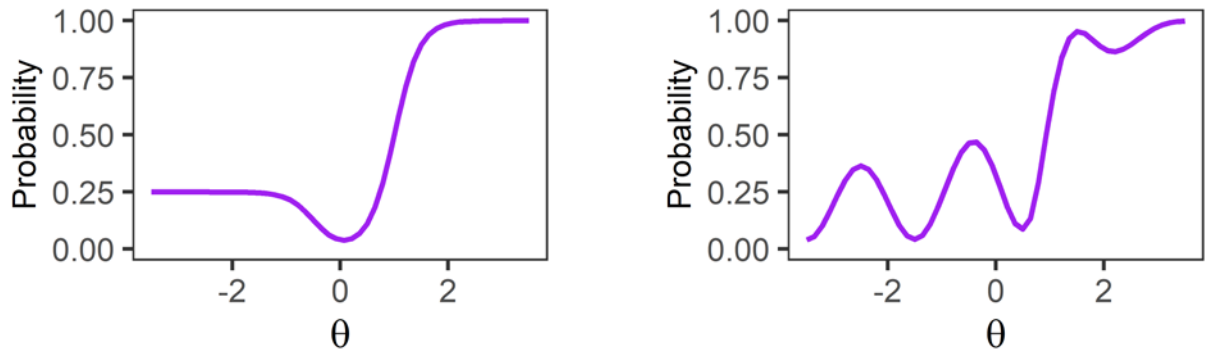

Figure 1. IRFs for misfiting items.

For misfitting item 1, $P(\theta) = \frac{.25}{1+e^{1.7(2.5)(\theta+0.5)}} + \frac{e^{1.7(2.5)(\theta-1)}}{1+e^{1.7(2.5)(\theta-1)}}$

For misfitting item 2, $P(\theta) = P^*(\theta) + .8(.5 - |.5 - P^*(\theta)|)\sin(\pi(\theta-1))$, where $P^*(\theta) = .2 + .8\frac{e^{1.7(\theta-1)}}{1+e^{1.7(\theta-1)}}$

For Question 3, two examinee distributions were used to draw examinee abilities. For one distribution, abilities were ~$\chi^2(3)$, standardized by subtracting 3 and dividing by √6. This distribution was not intended to be realistic, but instead was intended to represent an extreme

case; if integrating over a normal distribution made little difference with this extreme case, it would be unlikely to matter with real data. The other distribution was skew-normal, with parameters (0, 1, -2). This yielded approximate skew = -0.80 and excess kurtosis = 0.49. These distributions are shown in Figure 2. Item parameters and the Orlando-Thissen-Bjorner index were estimated twice, once integrating over a normal distribution and again estimating the ability distribution with empirical histograms. Twenty-one quadrature points were evenly spaced from -4 to 4. However, sometimes the estimated densities, item fit, and standard errors could not be estimated (although the parameter estimates seemed reasonably accurate). For the χ^2 data, this problem occurred for 11% of the 500 examinee/80 item condition replications and 2% of the 1000 examinee/80 item condition replications. For the skew-normal data, this problem occurred for 0.5% of the 500 examinee/40 item condition replications, 0.5% of the 1000 examinee/40 item replications, 24% of the 500 examinee/80 item condition replications, and 9% of the 1000 examinee/80 item condition replications. For these replications, the quadrature distribution was narrowed to -2.8 to 2.8, with 15 nodes; with this narrower range, the estimation terminated normally and produced estimates for the density, item fit, and standard errors.

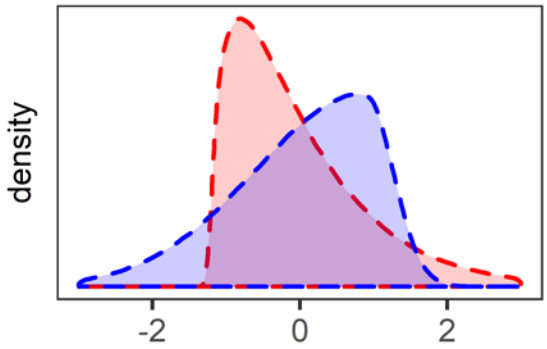

Figure 2: Non-normal distributions.

Results

Rest Score or Total Summed Score

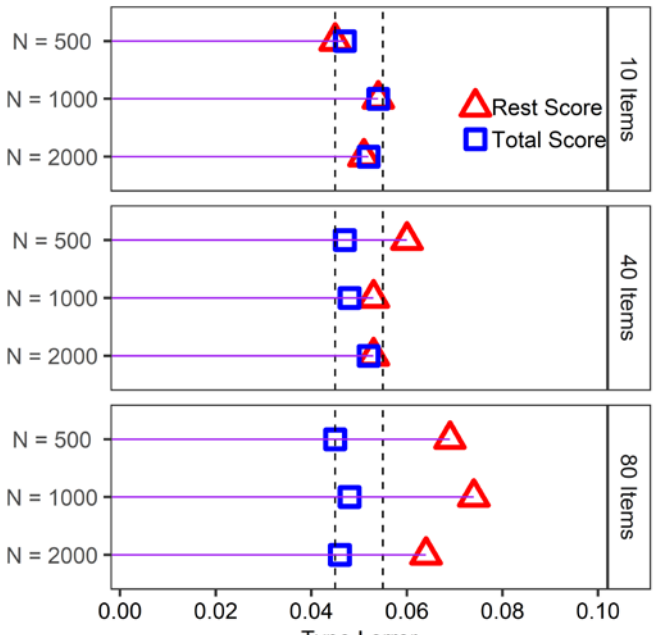

Figure 3: Type I error grouping on total score or rest score.

With 8000 items, one would expect the empirical Type I rate to fall between .045 and .055 95% of the time for a nominal α = .05. The total-score S-χ^2rates were within this range, but the rest-score Type I error rates were slightly above .055 in four of the nine conditions (Figure 3). Power was somewhat higher for the rest-score S-χ^2 (Figure 4).

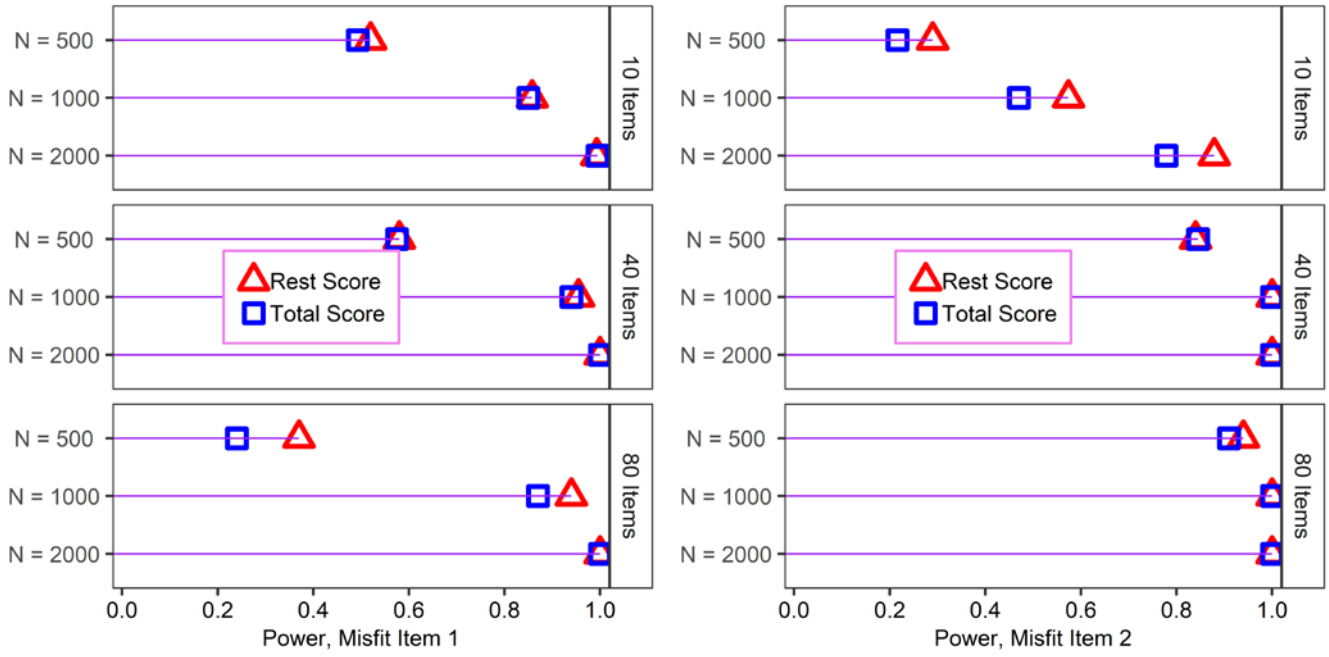

Figure 4: Power for Misfitting Items

Prior Distributions

In the first analysis, with fixed *c*-parameters but no prior distribution on the *a*-parameters, some of the estimated standard errors were extremely large. This problem was greater when the *c*-parameters were also freely estimated. The proportion of *a*-parameters with estimated standard errors greater than one (arbitrarily chose because one seemed large given the metric) is shown in Figure 5. Patterns were similar for large standard errors for the difficulty parameters, and for extreme estimates of the parameters. Thus, priors on the *a*'s and *c*'s are helpful for preventing extreme estimates of parameters and standard errors.

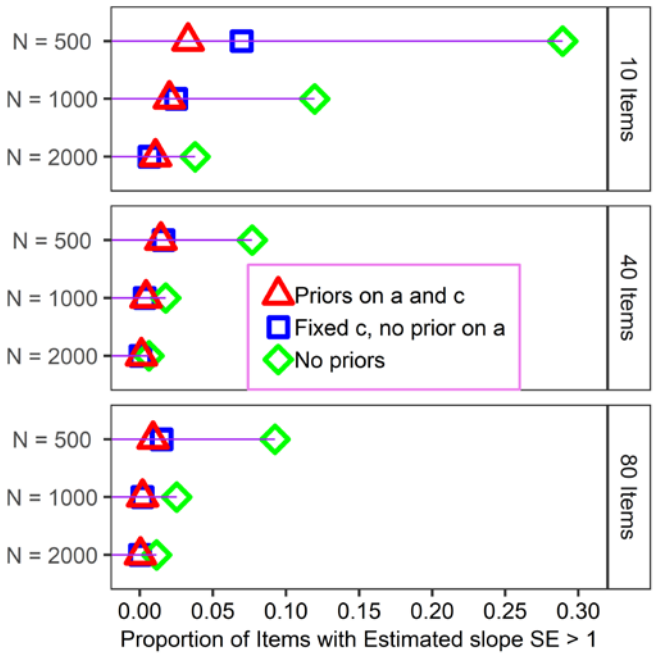

Figure 5. Effect of priors on standard error.

However, using priors impacts the distribution of S-χ^2 for the shortest tests. In Figure 6, Type I error rates were more inflated for 10 items with priors than they were without priors (Figure 3). The inflation for the 40 and 80-item test with the rest score was comparable to the same conditions without priors (Figure 3).

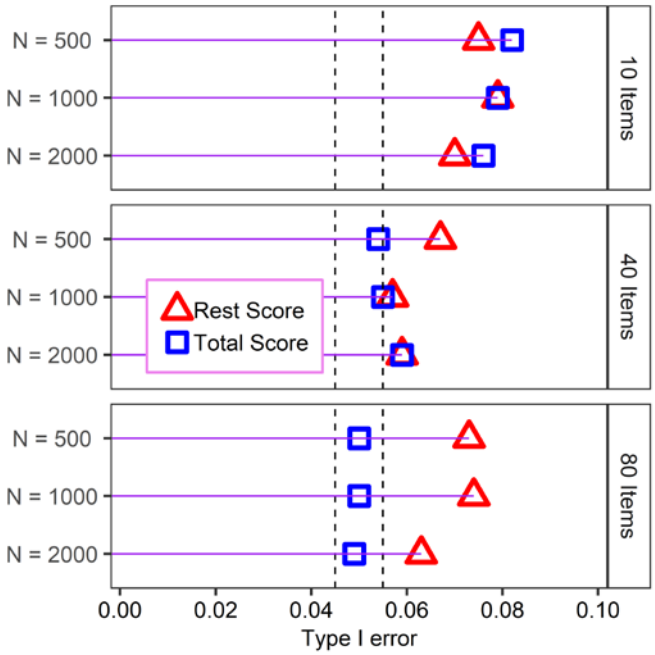

Figure 6. Type I error with priors

Nonnormal Ability Distribution

The item prior distributions from Question 2 were applied to avoid extreme values. Only the rest score was used for matching. Integrating over a normal distribution led to inflated Type I errors, especially for small samples and short tests, but estimating the ability distribution brought the error rate closer to the nominal value (Figure 7). Although Woods (2008) showed that 10 dichotomous items were not adequate for estimating the ability distribution, for the purposes of item fit the estimated ability distribution improved on using the incorrect normal distribution.

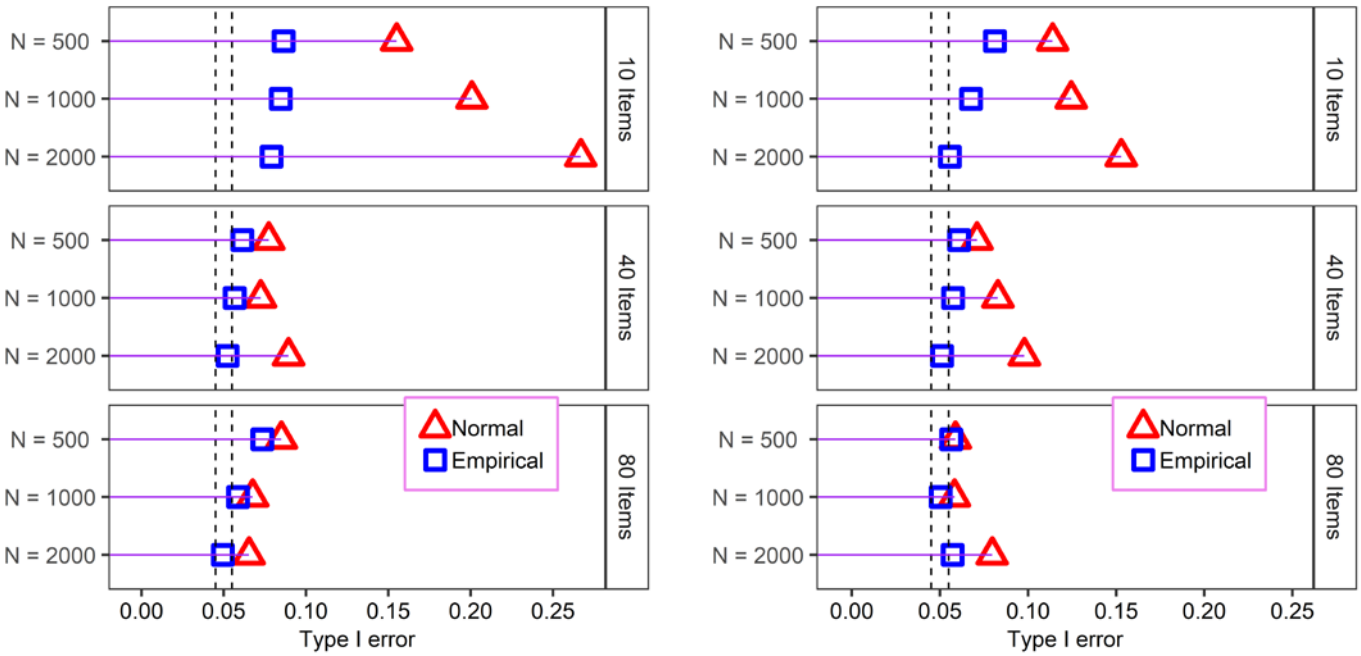

Figure 7: Type I error using either a normal distribution or the empirical estimation of the distribution. The left panel shows the χ^2 distributed θ and the right panel shows the skew-normal θ. Note the scale on the X-axis has changed from Figures 3 and 6.

Implications

The rest-score S-χ^2 provided slightly higher power with approximately the same Type I error rate as the total-score S-χ^2 for the conditions studied. The choice of index would not have

large practical implications. Utilizing prior distributions for the *a* and *c*-parameters reduced extreme estimates, but it appeared to decrease the degrees of freedom. Appropriate adjustment of the degrees of freedom merits further study. Finally, when the ability distribution was non-normal, using a normal distribution in the calculation produces inflated Type I error rates. Christensen, Bjorner, Kreiner, and Petersen (2004) noted: ". . . properties of the items cannot be separated from the properties of the latent distribution in a marginal maximum likelihood framework (Zwinderman and van den Wollenberg, 1990). A consequence of this is that it is impossible to distinguish lack of fit of the measurement model from a misspecified latent distribution." (p. 1310). Although Christensen et al were writing in the context of DIF, the concept generalizes to item fit. Further research could explore the impact of varying degrees of non-normality on item fit.

References

Cai, L. (2015). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika, 80*, 535-559.

Cai, L. (2017). flexMIRT_ version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO for Windows [Computer software]. Chicago, IL: Scientific Software International.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chon, K. H., & Sinharay, S. (2014). A note on the Type I error rate of the PARSCALE G^2 statistic for long tests. *Applied Psychological Measurement, 3*8, 245-252.

Christensen, K. B., Bjorner, J. B., Kreiner, S., and Petersen, J. H. (2004). Latent regression in loglinear Rasch models. *Communications in Statistics–Theory and Methods*, *33*, 1295-1313.

Glas, C. A. W., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87-106.

Mislevy, R. J.(1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Orlando, M. & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298.

Woods, C. M. (2008). Ramsay-curve item response theory for the three-parameter-logistic item response model. *Applied Psychological Measurement, 32*, 447-465.