10-2019

Examining the Effects of Specifying Bayesian Priors on the Wald's Test for DIF

Paulius Satkus
James Madison University, satkuspx@jmu.edu

Christine E. DeMars
demarsce@jmu.edu

Recommended Citation

Satkus, P., & DeMars, C. E. (2019, October). Examining the Effects of Specifying Bayesian Priors on the Wald's Test for DIF. Paper presented at the annual meeting of the Northeastern Educational Research Association. Rocky Hill, CT

Examining the Effects of Specifying Bayesian Priors on the Wald's Test for DIF

Paulius Satkus

Christine E. DeMars

James Madison University

Satkus, P., & DeMars, C. E. (2019, October). *Examining the Effects of Specifying Bayesian Priors on the Wald's Test for DIF.* Paper presented at the annual meeting of the Northeastern Educational Research Association. Rocky Hill, CT.

Examining the Effects of Specifying Bayesian Priors on the Wald's Test for DIF

In item response theory (IRT) models, differential item functioning (DIF) occurs when the item parameters are not invariant across groups, after scaling to a common metric. One popular method of assessing DIF is a Wald's test of the differences between item parameters. The Wald's test purportedly follows a χ^2 distribution with degrees of freedom equal to the number of parameters for the tested item. However, if prior distributions are applied to the item parameters, as is often needed when using small samples (Langer, 2008), the null distribution may not follow this assumed distribution. The purpose of this study is to examine the impact of priors on the distribution of the Wald's test when using the 3-parameter-logistic (3PL) model.

Theoretical Framework

The 3PL IRT model is parameterized as: $P(\theta) = c_i + (1 - c_i)\dfrac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$ or

$P(\theta) = c_i + (1 - c_i)\dfrac{e^{a_i\theta + d_i}}{1 + e^{a_i\theta + d_i}}$, where θ is the examinee's proficiency, a_i is the item discrimination for item i, b_i is the item difficulty or d_i is the item easiness, and c_i is a lower asymptote. One method of testing for differences between parameters is a Wald's χ^2 test, suggested by Lord (1980, Chapter 14). After scaling the parameters to the same metric, $\chi^2_\text{i} = \text{v}'_\text{i}\Sigma^{-1}_\text{i}\text{v}_\text{i}$, where v_i is the vector of item parameter differences and Σ^{-1}_i is the inverse of the error variance-covariance matrix of the item parameter differences. The degrees of freedom are equal to the number of item parameters, 3 for the 3PL model. The software flexMIRT (Cai, 2017) implements a version of the Wald's test. Initially, item parameters are held constant across groups so that the focal group mean and variance can be estimated relative to the reference group mean and variance. Holding the means and variances constant, the item parameters are then freed to be estimated separately for each group and tested for equality.

Several simulation studies have examined the empirical Type I error rate with the Wald's test (Langer, 2008, Wang & Woods, 2017, Woods, Cai, Wang, 2013). Langer (2008) found that Type I error rate was consistently too conservative (when compared to the nominal .05 rate) when data were simulated under the 3PL model. The author manipulated test length and sample size and found even with 1000 examinees and 40-item test, the empirical Type I rate was .02. Langer (2008) then conducted an additional simulation study and found that with 8000 examinees and 40 item-test, the empirical Type I error rate was not statistically different than the nominal .05 rate. Both Wang and Woods (2017) and Woods and colleagues (2013) simulated data under the 2PL model. In both of these simulation studies, the empirical Type I error rate was around .05 (Woods et al., 2013) and sometimes considerably larger depending on the anchor selection strategy (Wang & Woods, 2017).

Thus, it appears that the Type I error rate may be influenced depending on whether a 3PL model or a 2PL model is used. Langer (2008) noted that lower Type I error rates in the 3PL model may be attributed to near-zero Type I error rates for detecting the *c* parameter. When estimating the 3PL model using maximum marginal likelihood (MML), it is often necessary to put Bayesian priors on the item parameters, especially for small samples. Without priors, different combinations of item parameter may include unrealistic values (Langer, 2008, Mislevy, 1986). In fact, in Langer's (2008) study, up to 51% of estimated item slopes in small sample sizes were extreme. However, priors may change the distribution of the Wald χ^2 test. When using this test, we assume that we have freed 3 parameters, but the priors restrict those parameters somewhat. Thus, a 3-df test may not be appropriate.

Method

Three sample sizes (250, 500, or 2000 per group) were crossed with two levels of impact (0 or 1). The θ distribution was ~N(0,1) when impact = 0, or ~N(0.5,1) for the reference group and ~N(-0.5,1) for the focal group when impact = 1. The test length was 40 items, with c = .25, *a* =0.8 or 1.6, and *d* ranging from 2.2 to -2.2. The coefficient α for the resulting observed scores was .83. 1000 replications were simulated for each condition.

Parameter estimation and the Wald tests were conducted in Flexmirt (Cai, 2017). All items were used in the anchor and tested for DIF. Parameters were estimated with no priors and with a β(21,81) prior on the lower asymptotes and a N(1.2,0.5^2) prior on the *a*-parameters. The number and range of the quadrature points was decreased to 22 points spanning -3.5 to 3.5, due to convergence problems with the smaller sample. Additionally, the θs were assumed to be normal because, when priors for the item parameters were omitted and the sample size was small, the process often stopped without producing results when attempting to estimate the shape of the θ distribution[1].

Results

First, our findings highlight the need to apply Bayesian priors when modeling 3PL data. With priors, all six conditions were estimated with no estimation problems. When priors were not applied, a substantial number of replications produced problems (Table 1). For example, when sample size was 250 per group, and the impact was 1, 22.4% of the estimated Wald's statistic values were negative[2], 21.7 % were extremely high (i.e., > 100), and 7.1% of

[1] In general, it is not recommended to assume normality, as misspecification of the shape of the θ distribution can manifest as differences in item parameters (Woods, 2008). However, in this study the simulated θ distribution was normal, so this should not have impacted the results.

[2] Negative χ^2 values should not be possible. They occurred when the logit of the lower asymptote was estimated to be extremely negative for one group and the estimate of the error variance was negative (clearly, a variance can not be negative). Similarly, the extremely high χ^2 values occurred when the error variances for the lower asymptote of one group was extremely high.

replications crashed without converging to a solution. In fact, when the impact was 1, at least 20 % of all replications showed these problems, though fewer problems were observed with higher sample sizes. When the impact was 0, the same pattern of faulty replications was observed, however, the overall rate ranged from 6.8% when the sample size was 2000 to 15.1% when the sample size was 250. Difficult, more discriminating items had the fewest problems. Because the frequency of problems varied by condition, the number of replications used to calculate the Type I error rates, discussed next, varied depending on the condition.

The empirical Type I error rates for all 12 conditions are presented in Table 2. Overall, in all but one condition (no priors, N = 2000, Impact = 0) the Type I error rates did not approach the nominal .05 rate. In other words, all but one condition showed conservative Type I error rates, which is consistent with Langer's (2008) study. When priors were applied, the error rate was somewhat lower in all conditions. The Type I error rate became less conservative as sample size increased, again consistent with Langer (2008). The level of impact did not seem to affect the empirical Type I rate.

The Type I error rates in Table 2 were averaged over all items within a condition. Generally, the rates did not differ very much by item. Type I error rates were slightly higher for the less-discriminating items, especially when there was no impact.

The Type I error rate only measures the tail of the distribution. Figure 1 shows the distribution of the Wald's test and compares it to the theoretical χ^2 distribution. Results were pooled across items because the distribution did not vary greatly across items. It appears that especially in small sample sizes (with or without priors), the empirical distribution for the Wald’s test does not follow the theoretical χ^2 distribution. At sample sizes of 2000, the empirical distributions with no priors seems to approximate the theoretical distribution better than when priors were applied. In other words, when priors were applied, even at sample size of 2000, the empirical distribution was still noticeably different from the theoretical distribution. To sum up, the empirical distribution of the Wald's test did not seem to follow the theoretical χ^2 distribution regardless of whether the priors were applied or not, even at large sample sizes.

Conclusions and Educational Implications

Fairness in measuring educational and psychological constructs remains one of the more prominent issues today. Thus, it is imperative that the popular methods for detecting items that function differently across various groups are well studied and supported empirically. In this study, we examined the Type I error rate for the commonly used Wald's test statistic. Specifically, we hypothesized that applying Bayesian priors to the estimation process would change the Wald's test distribution. Consistent with findings in a previous study (Langer, 2008), we found that the Wald's test was too conservative under a 3PL model. Somewhat surprisingly, even when no priors were applied, the empirical χ^2 distribution was noticeably different from the theoretical distribution. Lastly, applying priors to the estimation method seems to be necessary for model convergence given the various problems that emerged when no priors were applied.

The findings from the current study directly inform practitioners who use the Wald's test to identify potentially problematic items. Conservative Type I error rate for the Wald's test implies that the test may lack power, and thus, the practitioners who use Wald's test may not "catch" all items that have smaller amounts of DIF. However, as usual, the current study has limitations, given only a specific subset of conditions (levels of impact, sample sizes) were examined. Therefore, results and the implications of the results should be interpreted with caution.

References

Cai L. (2017). flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring (version 3.5) [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Langer, M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation). Retrieved from https://cdr.lib.unc.edu/record/uuid:3ef47e17-9b76-45be-bf49-febddc17f4e0. University of North Carolina, Chapel Hill.

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

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

Wang, M., & Woods, C. M. (2017). Anchor selection using the Wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, 41, 17-29.

Woods, C. M. (2008). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement, 68*, 571-586.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*, 532–547.

Table 1

The percentage of problematic replications when Priors were not applied

	Negative χ^2	$\chi^2 > 100$	Crashed	Total
Impact=0				
N= 250	6.0%	5.2%	3.9%	15.1%
N= 500	3.9%	3.7%	0.2%	7.8%
N= 2000	3.3%	3.3%	0.2%	6.8%
Impact=1				
N= 250	22.4%	21.7%	7.1%	51.2%
N= 500	18.8%	18.3%	0.4%	37.5%
N= 2000	11.0%	10.9%	0.0%	21.9%

Note. No problems were encountered when Priors were applied.

Table 2

Type I Error Rate

	No Priors	Priors
Impact=0		
N= 250	0.8%	0.1%
N= 500	0.5%	0.4%
N= 2000	4.6%	0.9%
Impact=1		
N= 250	0.6%	0.1%
N= 500	0.8%	0.4%
N= 2000	3.4%	1.1%

Note: the nominal α = .05

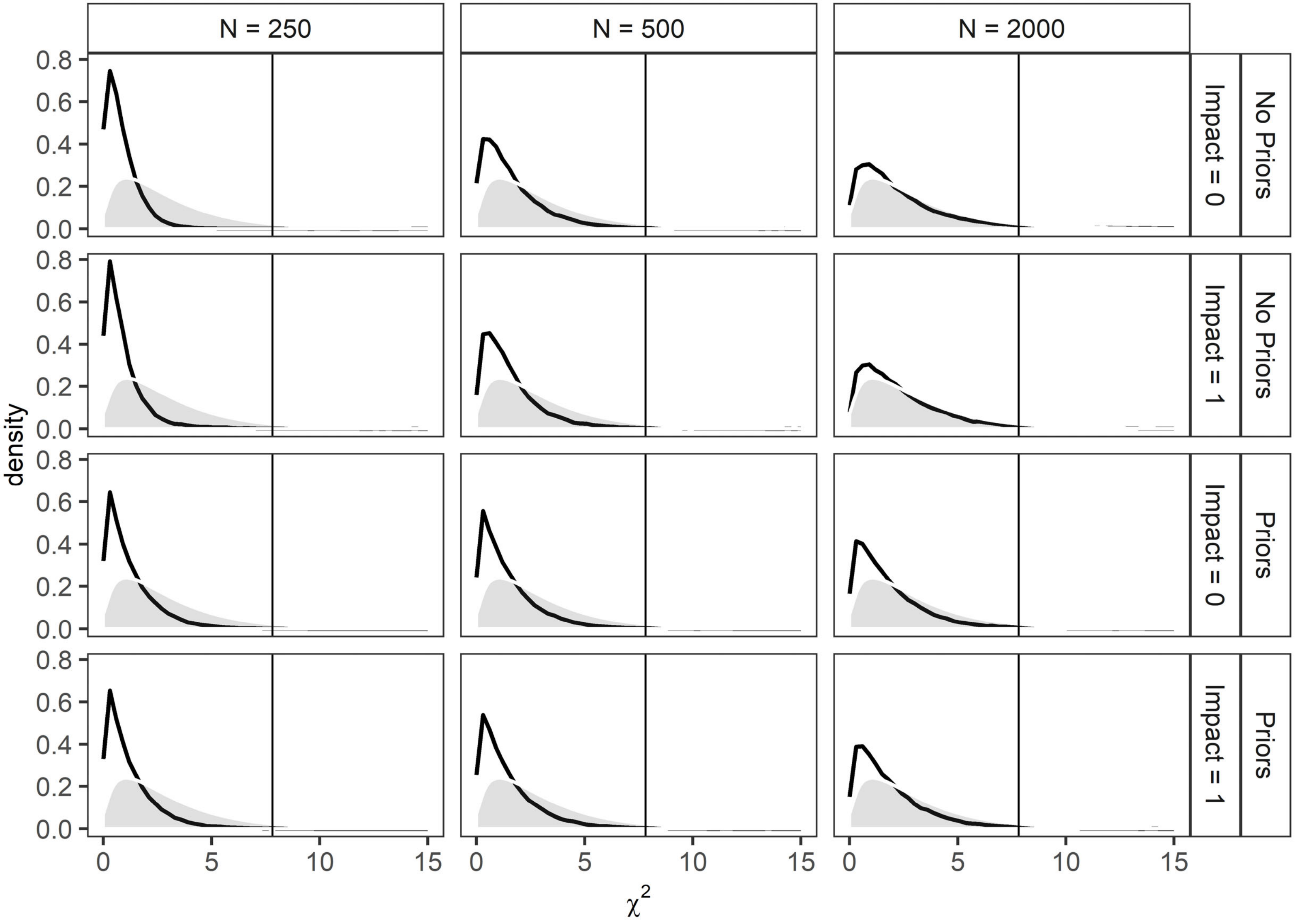

Figure 1: Empirical (black line) and theoretical (grey area) χ^2 with 3 df. The vertical line indicates the critical value for $\alpha = .05$ (7.815). With or without priors, far more values than expected are at the low end.