A Bayesian Nonparametric Model Fit statistic of Item Response Models

Purpose

As more and more states move to use the computer adaptive test for their assessments, item response theory (IRT) has been widely applied. Then investigating the fit of a parametric model becomes an important measurement process before building the item pool. If a misfitting item has been put in the item pool, it will cause the item selection and ability estimation error and then affect the validity of the test. Several researchers have developed IRT model fit assessing tests (Douglas & Cohen, 2001; Orlando & Thissen, 2000; Yen, 1981). However, these tests have some drawbacks might lead to over identify misfitting items and provide no help on investigating the reasons for item misfitting. Therefore, this study aims to propose a new Bayesian nonparametric item fit statistics and compare this new item fit statistics with more traditional item fit statistics to indicate the advantages of this method.

Theoretical Framework

While many of studies have developed and compared different approaches for model fit assessment, this topic remains a major hurdle to overcome for effective implementing IRT model (Hambleton & Han, 2005). Most of these approaches are X^2-based item fit statistics: Yen's (1981) Q_1 statistic, Orlando and Thissen's (2000) $S\text{-}X^2$ statistics, and G^2. However, these X^2-based item fit statistics have several drawbacks. First, they are sensitive to sample size (Hambleton & Swaminathan, 1985). When the sample size is large, this statistic test tends to over reject models because with large sample sizes statistical power is available to detect even very small discrepancies between the model and data. Since the state tests always have large sample size, almost all items have a significant X^2 statistic and do not fit the model. Second, several popular fit statistics do not have a X^2 distribution because ability estimation error, treating estimated parameters as true values (Stone & Zhang, 2003), and the degrees of freedom are in question (Orlando & Thissen, 2000). This drawbacks lead to falsely identify valid items as misfitting. Third, these fit statistics are not able to indicate the location and magnitude of misfit for a misfitting item. As a result, content experts cannot explain the item misfitting reasons and give suggestions on changing the model. Because of these limitations of X^2-based fit statistics, Douglas and Cohen (2001) developed a nonparameteric approach for assessing the model fit of dichotomous IRT model, hereafter refered to as kernel smoothing method. This method uses the kernel smoothing to draw a nonparametric item response function (IRF) and compares it with parametric IRF. Later, Liang and Wells generalized it for assessing the polytomouse IRT model fit. Liang, Wells and Hambleton (2014) compared this approach with traditional X^2-based item fit statistics (S-X^2 and G^2) under different conditions by manipulating test length, sample size, IRT models, and ability distribution. Their result indicated that this kernel smoothing method has exhibited controlled Type I error rates and adequate power. Moreover, this method provides a clear graphical representation of model misfit. However, in all these studies applying the kernel smoothing model fit assessment, the bootstrapping procedure was performed to construct an empirical distribution for determining the significance level of fitting statistics. This bootstrapping procedure did not count the uncertainty of the parameter estimation because it used the item parameter estimates previously obtained. The Bayesian statistic used in this study applying the posterior predictive model checking (PPMC) method (Gelman, Meng, & Stern, 1996; Guttman, 1967; Rubin, 1984) take the uncertainty of the parameter estimation into account by integrating over the parameters. Thus, this study will use the kernel smoothing and PPMC method to develop a Bayesian nonparametric model fit statistics and compare it with S-X^2 and bootstrapping statistics using simulation data and empirical data.

Method

The Monte Carlo simulation study was performed to examine the Type I error rate and power of the proposed statistic on detecting misfitting items in a mixed-format test under conditions differing on significance levels. In addition, this proposed Bayesian nonparametric statistic was being compared with: 1) S-X^2, provided by the computer software IRTPro (Cai, Thissen, & du Toit, 2011); 2) the bootstrapping kernel smoothing method. An empirical data set from a mixed-format test was analyzed to explore the use of the Bayesian nonparametric approach for assessing the model fit.

Kernel smoothing method. To apply the kernel smoothing method in educational assessment data, Y_i is a binary random variable denoting whether the score of one item is obtained or not and the latent trait variable is *Θ*. However, the latent trait variable *Θ* could not be observed directly. An estimator $\hat{\Theta}$ will replace *Θ*. The nonparametric IRF estimated by the kernel smoothing method is as follows:

$$\hat{P}_i^{non}\left(\hat{\Theta} = \theta\right) = \frac{\sum_{j=1}^{N} K(\frac{\theta - \hat{\theta}_J}{h}) y_{ij}}{\sum_{j=1}^{N} K(\frac{\theta - \hat{\theta}_J}{h})}. \quad (1)$$

In order to estimate IRF, y_js are averaged in the small range around every evaluation point θ by their weights. The weights of y_js are $K(u)$, a nonnegative symmetric kernel function with mode at 0,which is the monotonic decreasing of the absolute value of *u* (Copas, 1983; Douglas, 1997). In this research, Gaussian function is used as the kernel function in equation 1 because it is a typical choice of the kernel function (Douglas, 1997):

$$K(u) = \exp\left(-\frac{u^2}{2}\right). \quad (2)$$

In the IRF expression, h is a parameter called the bandwidth, which controls the smoothing amount. The choice of h is a trade-off between the fluctuation of the regression function and the bias of the regression function estimation (Copas, 1983; Douglas, 1997). Ramsay (1991) suggested an optimal value of h for psychometric binary data is depending on the sample size (N): h=1.1*$N^{0.2}$. In the present research, this value is used for h because Ramsay (1991) also showed that this value functioned well under the Gaussian kernel function. In order to plot the IRF estimated from equation 1, an estimate of the proficiency of each examinee is needed. The $\hat{\theta}_j$ (the estimate of the j-th examinee's proficiency) is estimated by the ordinal ability estimation (Douglas, 1997). In the ordinal ability estimation method, the empirical percentile of the examinee in the latent trait distribution is determined using the sum of item scores and latent trait distribution G's inverse function G^{-1} is used to calculate the proficiency based on the empirical percentile. In this research, we choose the standard normal distribution as the latent trait distribution.

PPMC. Suppose $\boldsymbol{\omega}$ is the unknown parameters of the assumed model H, $\boldsymbol{y}$ is the observed data, $p(\boldsymbol{\omega})$ is the prior distribution of the unknown parameters, and $p(\boldsymbol{y}|\boldsymbol{\omega})$ is the likelihood distribution of observed data assuming that the model H is true (Sinharay, 2005), then the posterior distribution of unknown parameters is $p(\boldsymbol{\omega}|\boldsymbol{y})$ and $p(\boldsymbol{\omega}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\omega})p(\boldsymbol{\omega})$. The replicated data, $\boldsymbol{y}^{rep}$, could be interpreted as the data that will be observed in the future or predicted and are then replicated using the same model H and parameters drawn from the $p(\boldsymbol{\omega}|\mathbf{y})$. The posterior predictive distribution of $\boldsymbol{y}^{rep}$ is as follows:

$$p(\boldsymbol{y}^{rep}|\boldsymbol{y}) = \int p(\boldsymbol{y}^{rep}|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{y})d\boldsymbol{\omega}. \tag{33}$$

This distribution was calculated from a Bayesian perspective to eliminate the nuisance parameters by integrating them out (Bayarri & Berger, 2000). A statistic for measuring model fit that was calculated from the observed data is compared to the distribution of the same statistic calculated from the replicated data drawn from this distribution and the posterior predictive p-value (PPP-value) is calculated. This was done to check the model fit (Gelman, Meng, & Stern, 1996). PPP-value provides a quantitative measure of the degree to which the model is able to capture the features of the observed data, in other words, the fit of the model to the observed data. The PPP-value is defined as (Bayarri & Berger, 2000):

$$p = p(T(\boldsymbol{y}^{rep}) \geq T(\boldsymbol{y})|\boldsymbol{y}) = \int_{T(\boldsymbol{y}^{rep}) \geq T(\boldsymbol{y})} p(\boldsymbol{y}^{rep}|\boldsymbol{y})\, d\boldsymbol{y}^{rep}. \tag{34}$$

In this equation, $T(\boldsymbol{y})$ is a discrepancy measure which is the statistic for measuring model fit. Extreme PPP-values, those close to 0, 1, or both (depending on the nature of the discrepancy measure), indicate the model does not fit the data (Sinharay et al., 2006).

Procedure of Bayesian nonparametric fit statistic. The model *H* in this study was IRT model. The parameters $\boldsymbol{\omega}$ were item parameters, and proficiency parameters, θs. The observed data $\boldsymbol{y}$ were the responses of examinees for all the items. The discrepancy measure $T(\boldsymbol{y})$ for the *i*-th item was as follows:

$$T(\boldsymbol{y}) = \sqrt{\frac{\sum_{k=1}^{K-1} \left[\frac{\sum_{q=1}^{Q} (\hat{P}_{qk} - \hat{P}_{qk}^{non})^2}{Q} \right]}{K-1}}$$

where $\hat{P}_{qk}$ and $\hat{P}_{qk}^{non}$ are the estimated probability of IRT model and nonparametric model for each point and score category; Q is the number of evaluation points (e.g., Q=100); and K is the total number of categories.

The following were procedures of calculating the Bayesian nonparametric fit statistic:

1. The MCMC algorithm was used to simulate the posterior distributions of item parameters and proficiency parameters using the observed data and the IRT model.

2. A total of N θs was drawn from corresponding posterior distributions. The sample size was N. For example, θ_j of the j-th examinee was drawn from the posterior distribution $p(\theta_j|\mathbf{y})$.

3. The item parameters of n items were drawn from their posterior distributions.

4. A data set was generated from the IRT model using item parameters and proficiency parameters.

5. The discrepancy measure was calculated for this data set and compared with the same discrepancy measure calculated from the observed data.

6. Steps 2 to 5 were repeated for M (e.g. M=100) times to compute the PPP-value for every item.

Data

Simulation data. The simulation data were generated from the simulated item parameter estimates follow the distributions, a~*log-N* (0, 0.4) and b~*N* (0, 1). Both two-parameter logistic model (2PLM) and the graded response model (GRM) were used. Among all items, 80% of them are generated from 2PLM and 20% of them are generated from GRM. In each simulated test, the 20% of misfitting items were generated, and 50% of misfitting items were dichotomous items and 50% of misfitting items were polytomous items. The samples size was 3000 and test length was 60 items. Two significance levels, α, were 0.05 and 0.01. For each significance levels, 100 replications were conducted.

Empirical data. The data came from a large-scale assessment with 53 items at a sample size of 3804. Ten partial credit polychromous items were fitted by the GRM and the rest of dichotomous items were fitted by the 2PLM. The fit were tested using S-X^2, bootstrapping kernel smoothing method, and Bayesian nonparametric method. Three statistics were compared for the items flagged as misfitting. These misfitting items were further explored by via a graphical representation.

Result

For simulation data, Table 1 and Table 2 summarize the Type I error rate and power for three fit statistics being compared. Table 1 reports the empirical type I error rates of PPMC, bootstrapping and S-X^2 for the mix format test, 2PLM items and GRM items.

Table1

Empirical type I error rates of PPMC, bootstrapping and S-X^2

Significant level	Model	PPMC	Bootstrapping	S-X^2
0.05	Mix format	0.088	0.095	0.063
	2PLM	0.090	0.109	0.062
	GRM	0.072	0.000	0.07
0.01	Mix format	0.020	0.029	0.015
	2PLM	0.020	0.033	0.015
	GRM	0.020	0.000	0.020

The comparison results of 0.05 and 0.01 significant level indicated that S-X^2 has the lowest type I error rate for the mix format test and 2PLM items and bootstrapping method has lowest type I error for the GRM items. In general, beside the 0 type I error rate of bootstrapping method for GRM items, the type I error rates of these three methods under two significant levels are very close.

Table 2 reports the empirical detection rate of PPMC, bootstrapping and S-X^2 for the mix format test, 2PLM items and GRM items.

Table2.

Empirical detection rates of PPMC, bootstrapping and S-X^2

Significant level	Model	PPMC	Bootstrapping	S-X^2
	Mix format	0.898	0.498	0.983
0.05	2PLM	0.888	0.995	0.987
	GRM	0.908	0.000	0.978
0.01	Mix format	0.765	0.439	0.932
	2PLM	0.823	0.878	0.923
	GRM	0.707	0.000	0.940

The comparison results of both significant levels indicated that S-X^2 method has the highest empirical detection rate for the mix format test, 2PLM items and GRM items. It should be noticed that the empirical detection rate for bootstrapping method is very low for GRM items. When the significant level is at 0.05, the empirical detection rate of bootstrapping and S-X^2 method is very close. But when the significant level is at 0.01, the S-X^2 method has much higher empirical detection rate.

Based on the simulation result, three methods all provide the accurate model checking rate for the 2PL model which is indicating by high empirical detection rate and low type I rate under both significant levels. But when the model is the GRM, the bootstrapping method cannot identify misfit items. The other two methods can identify the misfit GRM items accurately.

For the empirical data, among 53 items, S-X^2 method has identified 22 items as misfitting items, bootstrapping method has identified 14 items as misfitting items and PPMC method has identified 13 items as misfitting items. Only four items were identified misfitting by three methods. The following graphs includes the IRFs of nonparametric and parametric model for these four misfitting items of the assessment.

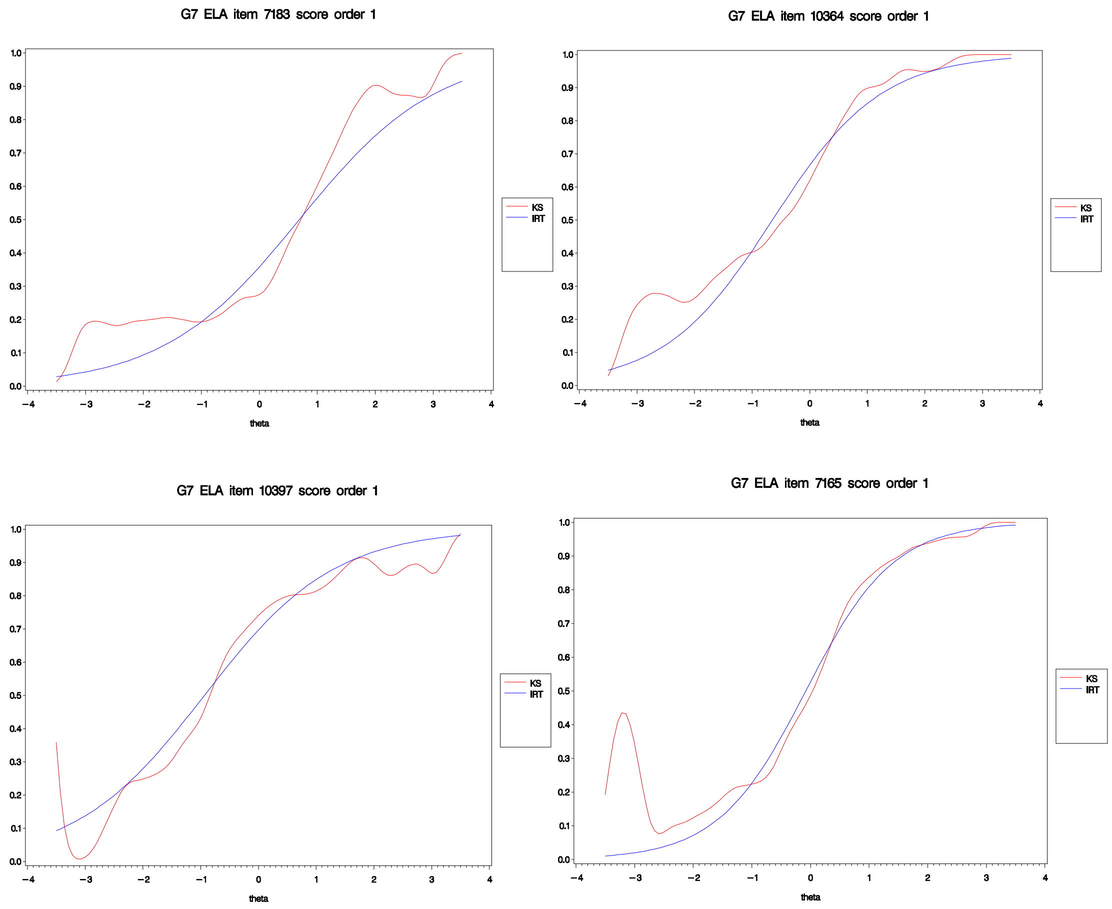

Figure 1. IRFs of nonparametric and parametric model for four misfitting items.

The nonparametric IRFs indicate the misfitting location is at the middle and high theta range for the first item, misfitting location is at the lower theta range for the second item, misfitting location is at the low and high theta range for the third item, and mifitting location is at the low theta range.

Significance

The new proposed Bayesian nonparametric model fit assessing method can be easily generalized to any IRT model and count the uncertainty of parameter estimation. This method also provides the graph representation of misfitting items for investigation of the location and magnitude of misfit. The result from the simulation study indicate this Bayesian nonparametric model fit statistics have low type I error and high detection rate. In the empirical study, the

Bayesian nonparametric method detected reasonable number of misfitting items. These results indicate Bayesian nonparametric model fit statistics can be used as a model fit assessing method. When Bayesian nonparametric method was compared to other model fit methods in simulation study, the detection rates are very similar between the Bayesian nonparametric method and the S-X^2 method. This result indicates these two model fit methods can check model fit equally well and the selection of model fit method depends on the estimation method used for item calibration because each method used different item calibration results for model fit checking. For example, if the MCMC method is used for estimation, Bayesian nonparametric method should be used for model fit checking. If the maximum likelihood method is used for estimation, S-X^2 method should be used for model fit checking.

Reference

Bayarri, M., & Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association, 95*(452), 1127-1142.

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

Copas, J. (1983). Plotting p against x. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 32*(1), 25-31.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, 62(1), 7-28.

Douglas, J., & Cohen, A. S. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, 25, 234–243.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733-759.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological), 29*(1), 83-100.

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Washington, DC: Degnon Associates.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.

Liang, T., & Wells, C. S. (2009). A model fit statistic for the generalized partial credit model. *Educational and Psychological Measurement*, 69, 913–928.

Liang, T., Wells, C. S. & Hambleton, R. K. (2014). An assessment of the nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement*, 51 (1), 1-17.

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

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151-1172.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, 42(4), 375-394.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit in item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40, 331–352.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245–262.