SCRIPTION
GLOBALIZE YOUR SCIENTIFIC RESEARCH

Journal of Psychology and Personality Disorders

Original Article

Volume: 1; Issue: 1

Investigating the Practical Utility of Proportion Agreement, Simulation-based Proportion Agreement, and Concordance Index for Item Fit Assessment in Item Response Theory

Insu Paek, Hirotaka Fukuhara, Lanrong Li

Florida State University, College of Nursing, 98 Varsity Way, Tallahassee, FL 32306-4310, (Tel) 850-644-6007

Pearson Assessments,

Amplify Education, Inc.

Corresponding Author: Insu Paek, College of Nursing, Florida State University, (Tel) 850-644-6007

Email: ipaek@fsu.edu

Received Date: 15th August 2022 **Acceptance Date:** 26th August 2022 **Published Date:** 30th August 2022

Abstract

Item response theory (IRT) is a modern measurement technique and widely used in psychology and education. Popular IRT models for dichotomously scored item response data are the 2-parameter logistic model (2PLM) and the 3-parameter logistic model (3PLM). In applications of IRT, item fit investigation is typically conducted. To complement the existing well-known statistical significance test of item fit such as Q_1 and $S-X^2$, this study proposes and explores the utility of proportion agreement (PA), its refined version, PA2, and concordance index (CD), which are descriptive item fit measures equipped with efficient computation and easy understanding of the degree of item fit. The performance of PA, PA2, and CD was also compared with the existing item fit test procedures, Q_1 and $S-X^2$.

Acknowledgment

The simulation results presented in this paper were used for a presentation at the 2021 annual meeting of the National Council on Measurement in Education Conference.

Item response theory (IRT) is a modern measurement technique which uses a latent variable modeling. IRT now is widely used for educational and psychological measurement. Ensuring reasonable model-data fit in item response theory (IRT) modeling is a cornerstone for its diverse applications, for example, for test construction, item-banking, equating, and computer adaptive testing **[7]** de Ayala, 2009; **[8]** Embretson & Reise, 2000; **[9]** Hambleton, Swaminathan, & Rogers, 1991; **[10]** Hambleton & Swaminathan, 1985; **[23]** Yen & Fitzpatrick, 2006). In IRT, both overall model/relative model fit and item fit investigations are of interest in IRT applications.

For the item fit evaluation, a graphical checking approach, statistical test approach, and fit index approaches are available. In the most popular graphical check approach, a plot in which the semi-empirical item characteristic curve (ICC) is compared with the model-predicted ICC. The empirical ICC is at best, semi-empirical because it is based on estimated person theta (ability) scores, not the true person thetas. For the statistical test approach, several procedures are available. Three of conventional well-known procedures in the literature are Yen's Q_1 **[22]** 1981), Bock's X^2 **[3]** 1960), and McKinley and Mills' G^2 **[14]** 1985). Yen's Q_1 and Bock's X^2 are a chi-square test approach which is based on the differences between observed responses and estimated probabilities. They are essentially the same except for choosing the number of intervals where those differences are computed. McKinley and Mills' G^2 is based on the likelihood ratio. The use of Q_1 is observed sometimes in some large-scale assessment (e.g., **[15]** New York State Education Department, 2018). The item fit test procedure $S-X^2$ **[16]** Orlando & Thissen, 2000, **[17]** 2003) was proposed as an enhancement of those Q_1 and G^2 and it is currently widely available in recent IRT programs such as IRTPRO ([5] Cai, Thissen, & du Toit, 2011), [4] flexMIRT (Cai, 2017), and [6] "mirt" (Chalmers, 2012) in the statistical computing environment R **[18]** R core team, 2021).

The popular existing fit procedures, Q_1 and $S-X^2$ were proposed as statistical tests. Using the test statistics of Q_1 and $S-X^2$ is not the most appealing way to deliver the extent of item fit. This is particularly true when the information about item fit should be communicated with those who are strangers to statistics or IRT (e.g., reporting the degree of model-data fit including item fit to teachers who are not familiar to IRT and/or statistics in assessment programs), due to the lack of easy interpretation for the degree of fit. Also, when item responses have examinee guessing, following the 3-parameter logistic model (3PLM) ([2] Birnbaum, 1968), Q_1 and $S-X^2$ do not show very satisfactory power to distinguish the 3PLM and the 2-parameter logistic model (2PLM). In addition, in the IRT literature, model-fit-assessment in terms of a model's predictive power (i.e., sensitivity and specificity) has not been given due attention. Given these, the use of descriptive measures with easy computation and interpretation which are based on the predictor power of a model could help with facilitating the understanding of the degree of item fit in practice and communication between IRT practitioners and clients.

In this study, therefore, we explore the utility of three descriptive item fit measures below, which are built upon the model's predictive power, computationally very efficient, and very straightforward to understand for dichotomously scored item response data with the 2PLM and the 3PLM. The three descriptive measures are proportion agreement (PA), simulation-based proportion agreement (PA2) and concordance index (CD). (See, e.g., **[1]** Agresti (2019) for CD in the context of logistic regression analysis.) To assess their practical utility, they are compared with the existing

popular item fit procedures of Q_1 and $S-X^2$ via simulations in this study.

Method

For the details of the two existing well-known methods, Q_1 and $S-X^2$, readers are referred to the aforementioned references. The three descriptive fit measures, IRT models used in the study, and the simulation design are described below.

Three descriptive item fit measures

Proportion agreement (PA). PA shows the extent of agreement between observed and predicted item responses. PA ranges from 0 (none) to 1 (perfect agreement). The higher the better the item fit is. The steps to calculate PA are:

1) Fit an IRT model and calculate $P(\theta)$ which is the probability of a correct answer given a theta score for a particular item of interest. $P(\theta)$ is the model-predicted item response.

2) Generate a predicted item score ($\hat{y}$). $\hat{y}=1$ if $P(\theta)>.5$ and $\hat{y}=0$ otherwise.

3) Construct a 2x2 (frequency) contingency table ($\boldsymbol{A}$) where row (y) = empirical category of 0 and 1 and column ($\hat{y}$) = predicted category of 0 and 1.

4) PA $=trace(\boldsymbol{A})/N$, i.e., sum of diagonal elements/total number of test takers (N).

5) PA is a combined measure of specificity and sensitivity. In an item fit assessment situation, specificity can be defined as $\Pr(\hat{y}=0|y=0)$ and sensitivity as $\Pr(\hat{y}=1|y=1)$.

6) Let the item difficulty index in classical test theory, which is the marginal proportion of $y=1$, be denoted by P. It can be shown that PA is equal to $P\times\text{sensitivity}+(1-P)\times$ specificity, i.e., a weighted average of CTT item difficulty and 1- CTT item difficulty.

Concordance index (CD).

If a model describes data better, we expect that the model shows higher sensitivity given a particular value of specificity. Thus, we would like to see the fitted model show the highest sensitivity given specificity. CD quantifies the degree of prediction power of a model by considering sensitivity across all different levels of specificity. CD is technically equivalent to the area under what is called "receiver operating characteristic curve". Another way of understanding CD is that it is a probability that shows the extent of concordance between observed and predicted responses. CD ranges from 0 to 1. CD = 0.5 is equivalent to the level of random guessing, thus we want to see CD greater than 0.5.

Simulation-based proportion agreement (PA2).

PA2 is similar to PA. However, in generating the predictive response $\hat{y}$ in step 2 for PA, a simulation is accommodated to avoid the use of the pre- specified cut off (0.5) to generate predicted response values. In step 2, instead of using 0.5, draw a value, u, from a standard uniform distribution. Then $\hat{y}=1$ if $u<P(\theta)$ and $\hat{y}=0$ otherwise. The rest of the steps are the same as in PA. Also, to reduce chaptalizing on chance, repeat the whole steps up to R number of times. In each of the R number times, we obtain PA, producing R number of Pas. With the set of $\{\mathrm{PA}_1,\mathrm{PA}_2,\cdots,\mathrm{PA}_\mathrm{R}\}$, use $\sum_{r=1}^{R}PA_r/R$ as the final estimate of proportion agreement in PA2. PA2 obviates the choice of cut-off such as 0.5 and reduces capitalizing on chance in PA.

IRT Models and response data generation

The IRT models used for data generation and fitting the data are the 2PLM and the 3PLM (1968). The 3PLM is:

$$P(X_j=1|\theta_i)=g_j+(1-g_j)\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}},$$

where a_j, b_j, and c_j are the *j*th item discrimination, difficulty, and pseudo-guessing parameter, respectively, and θ_i is the *i*th person latent trait score. Constraining $g_j = 0$ results in the 2PLM.

The data-generating parameters for θ, a, b, and g were drawn from the standard normal distribution and the truncated normal distributions (TNs) with the lower bound (Lb) and the upper bound (Ub) as follows: $\theta \sim N(0,1)$; a (discrimination) $\sim TN(\mu_a = 1.5, \sigma = 0.5, Lb = 0.8, Ub = 2.5)$
b (difficulty) $\sim TN(\mu_a = 0, \sigma = 1, Lb = -2, Ub = 2)$; and g $(pseuo\ guessing) \sim TN(\mu_a = 0.2, \sigma = .05, Lb = .05, Ub = .3)$. The steps to generate item response data (which follow either 2PLM or 3PLM) are: for every response, (a) with the (true) data-generating parameters, calculate the probability of a correct answer, P_{ij}, using the above equation (either 2PLM or 3PLM), (b) draw a value u randomly from the standard uniform distribution, and (c) assign 1 if $P_{ij} \geq u$ and assign 0 otherwise.

Simulation Design

The simulation design has two big scenarios where dichotomous responses are collected. The first is when data are from a cognitive test where low or very low latent trait test-takers make guesses, thus data are contaminated by test-taker guessing. The second scenario is where data are from a psychological test. The first scenario is often found when multiple-choice (MC) item type cognitive tests are administered in large-scale assessment programs. In those programs, a long test (30 items to 50 items for example) with a large sample size (at least a few thousand test-takers) is calibrated typically by the 3PLM. The second scenario is usually observed when a psychological test with yes and no response options (e.g., in an attitude test) is administered to relatively smaller groups of people, for which the 2PLM is often used.

The first scenario: data from cognitive test with multiple-choice (MC) items.

Given the data sizes for applications of the 3PLM in practice described above, a 41-item test with a sample size of N = 2000 was simulated. One item in a test was designated as a studied item (SI) whose item fit was assessed. Responses of the studied item were generated by either the correct model (3PLM) or a misfit model (2PLM), while the rest items (RIs) in a test were generated using the 3PLM. To each of the simulated data sets, the five item-fit assessment procedures (Q_1, $S - X^2$, PA, CD, and PA2) were applied.

The second scenario: data from a psychological test.

The 2PLM is a popular choice for a psychological test. The number of items in a test and the sample size tend to be smaller than those to which the 3PLM is applied. The simulated test length and the sample size were 21 and 500, respectively. Again there was a studied item for the assessment of the item fit. Responses of the studied item were generated by either the correct model (2PLM) or a misfit model (3PLM), while the rest items in a test were generated using the 2PLM. As above, the five item-fit assessment procedures were applied to each simulation data set.

Variation in the studied item.

To reflect that the studied item can be different in terms of item discrimination (a), item difficulty (b), and pseudo-guessing (g), their variations were accommodated in the simulation. In the 2PLM fit to the 2PLM responses, $a = 1, 1.5, \text{and } 2$ and $b = -1, 0, \text{and } 1$, thus $3 \times 3 = 9$ studied item conditions. In the 3PLM fit to the 3PLM responses, $a = 1, 1.5, \text{and } 2$ and $b = -1, 0, \text{and } 1$, and $g = 0.2$, thus $3 \times 3 \times 1 = 9$ studied item conditions. When a misfit model was used for the studied item, that is, 2PLM fit to 3PLM responses, the studied item was generated with $a = 1, 1.5, \text{and } 2$; $b = -1, 0, \text{and } 1$, and $g = 0.2$, thus $3 \times 3 = 9$ studied item conditions.

In summary, in the first scenario in which the studied item is based on fitting the correct model (3PLM) or a misfit model (2PLM), we have 9 conditions for the correct model fit for studied item (3PLM fit to 3PLM responses) and also another set of 9 conditions for the misfit model for the studied item (2PLM fit to the 3PLM responses). Note that the 3PLM is used for the rest items in a test. In second scenario, all item responses follow the 2PLM and the 2PLM is used for all items, thus, there are 9 conditions for the correct model fit for the studied item. In each of these conditions, 100 replications were made. To each of the replications, the five item fit assessment procedures (Q_1, $S-X^2$, PA, CD, and PA2) were applied. The two existing well-known item fit procedures (Q_1, and $S-X^2$) are statistical test procedures and their rejection rates (for the null hypothesis of item fit against item misfit) were recorded.

The IRT model estimation and the implementation of Q_1 and $S-X^2$ were conducted using the "mirt" program (Chalmers, 2012) in R. For the calculation of CD, the "pROC" program **[19]** Robin, et al., 2011) in R was used. All other computations and simulation of the data were made using R.

Results

The results were organized for the correct model fit case for the studied item and for a misfit model case for the studied item. Because Q_1, and S-X^2 are statistical inferential procedure, their rejection rate (or percentage of rejections indicating a misfit item) under α=.05 were computed. PA, CD, and PA2 are descriptive measures for item fit. (The higher they are the better item fit is.) Their averages across all replications were recorded for the correct model condition for the studied item. In the misfit model fit condition for the studied item, for every replication, the results of PA, CD, and PA2 from fitting the misfit model were compared with those from fitting the correct model. Then the rejection of the misfit model was recorded if PA (or CD or PA2) for the correct model > PA (or CD or PA2) for the misfit model.

Correct model fit case: Type I error rates and Averages of PA, CD and PA2

In the correct model fit case (i.e., 2PLM fit to 2PLM response and 3PLM to 3PLM responses), the rejection rates of Q_1 and $S-X^2$ represent Type I error rates and we expect around 5% of rejection rates for the existing fit test procedures of $S-X^2$ and Q_1, The results here (Table 1) also tell us the overall expected predictive power when using PA, PA2, and CD when correct models are used.

			Type I error rates				Average					
			SX2		Q1		PA		CD		PA2	
a	*b*	*g*	2PLM	3PLM	2PLM	3PLM	2PLM	3PLM	2PLM	3PLM	2PLM	3PLM
1	-1	0 (2PLM) and .2 (3PLM)	0.03	0.04	0.08	0.17	0.74	0.77	0.75	0.73	0.65	0.68
	0	0 (2PLM) and .2 (3PLM)	0.06	0.03	0.05	0.05	0.69	0.66	0.75	0.71	0.59	0.58
	1	0 (2PLM) and .2 (3PLM)	0.07	0.06	0.06	0.03	0.75	0.64	0.76	0.67	0.65	0.55
2	-1	0 (2PLM) and .2 (3PLM)	0.04	0.07	0.07	0.13	0.81	0.82	0.84	0.81	0.73	0.74
	0	0 (2PLM) and .2 (3PLM)	0.07	0.03	0.04	0.13	0.75	0.71	0.83	0.77	0.65	0.62
	1	0 (2PLM) and .2 (3PLM)	0.06	0.07	0.09	0.13	0.81	0.69	0.84	0.71	0.73	0.59
2	-1	0 (2PLM) and .2 (3PLM)	0.05	0.05	0.08	0.19	0.86	0.86	0.9	0.87	0.79	0.79
	0	0 (2PLM) and .2 (3PLM)	0.08	0.1	0.02	0.08	0.8	0.74	0.88	0.82	0.71	0.65
	1	0 (2PLM) and .2 (3PLM)	0.05	0.04	0.11	0.13	0.86	0.72	0.89	0.73	0.79	0.61

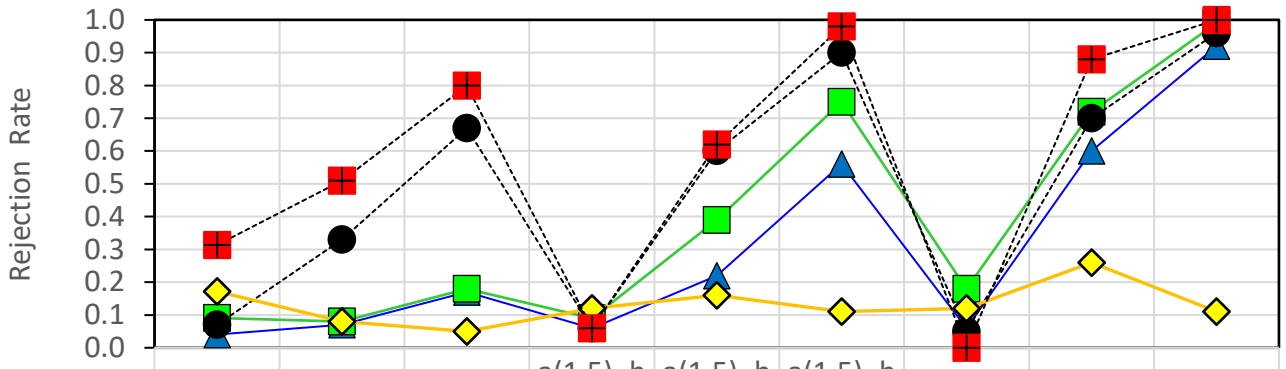
Table 1. Type I error rates of $S-X^2$ and Q_1 and averages of PA, CD, and PA2 when a correct model is fit

The averages of PA and PA2 represent the overall correct classification rates. PA2 avoids capitalizing on chance, thus its values are generally lower than PA. The Type I error rates of $S - X^2$ ranged from 0.03 to 0.08 for the 2PLM data and from 0.03 to 0.10 for the 3PLM data. The range of Q_1 Type I error rates were from 0.02 to 0.11 for the 2PLM data and from 0.03 to 0.17 for the 3PLM data. In general, Q_1 showed higher Type I error rates for the 3PLM data. The predictive power or the overall correct classification rates using PA was 69% to 86%. PA2 showed lower values ranging between 55% to 79%. Again, this is because PA2 avoids capitalizing on chance by incorporating more variations through simulation in its process.

All CD values which represent an overall correspondence between data and predicted values were greater than 0.5, ranging between .71 to .90.

Misfit model case: rejection rates for detecting a misfit item when the 2PLM is fit to the 3PLM responses for the studied item.

The misfit model cases have two occasions. The first one is when the 2PLM was fit to the 3PLM responses for the studied item (SI) while the rest items (RIs) in a test are fitted by the 3PLM. The rejection rates of a misfit model (or the rates for detecting a misfit item) and their graphic presentations are given in Figure 1.

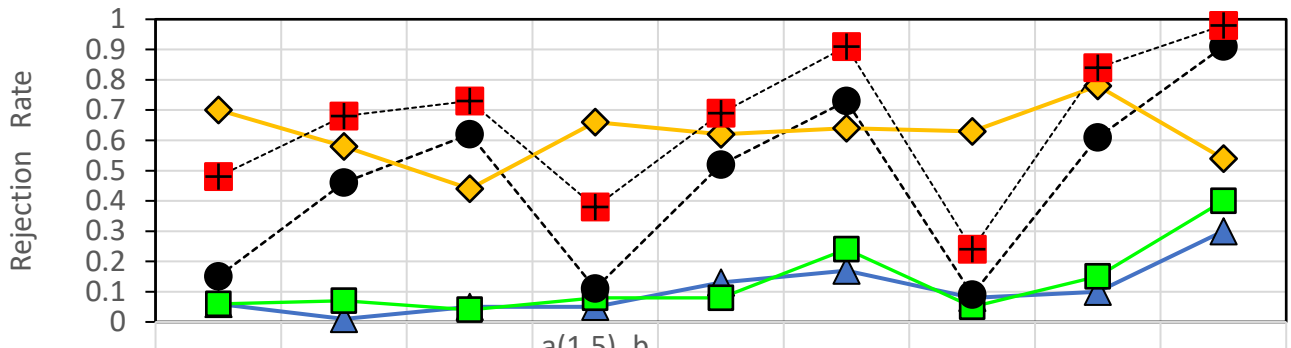


	a(1)_b(-1)_g(.2)	a(1)_b(0)_g(.2)	a(1)_b(1)_g(.2)	a(1.5)_b(-1)_g(.2)	a(1.5)_b(0)_g(.2)	a(1.5)_b(1)_g(.2)	a(2)_b(-1)_g(.2)	a(2)_b(0)_g(.2)	a(2)_b(1)_g(.2)
SX2	0.04	0.07	0.17	0.06	0.22	0.56	0.06	0.60	0.92
Q1	0.09	0.08	0.18	0.09	0.39	0.75	0.18	0.72	0.99
PA	0.07	0.33	0.67	0.06	0.60	0.90	0.05	0.70	0.96
CD	0.17	0.08	0.05	0.12	0.16	0.11	0.12	0.26	0.11
PA2	0.31	0.51	0.80	0.06	0.62	0.98	0.00	0.88	1.00

Figure 1. Item misfit detection rates (2PLM fit to 3PLM responses) for the studied item (SI) when the rest items (RIs) follow the 3PLM.

Note. The x-axis represents the item parameter values for the condition: item discrimination (a), item difficulty (b), and pseudo-guessing parameter (g). For example, a(1)_b(-1)_g(.2) means $a = 1, b = -1$, and $g = 0.2$.

The results of the second occasion, where the 2PLM was fit to the 3PLM responses for SI while RIs in a test were fitted by the 3PLM, are shown in Figure 2.



	a(1)_b(-1)_g(.2)	a(1)_b(0)_g(.2)	a(1)_b(1)_g(.2)	a(1.5)_b(-1)_g(.2)	a(1.5)_b(0)_g(.2)	a(1.5)_b(1)_g(.2)	a(2)_b(-1)_g(.2)	a(2)_b(0)_g(.2)	a(2)_b(1)_g(.2)
SX2	0.06	0.01	0.05	0.05	0.13	0.17	0.08	0.1	0.3
Q1	0.06	0.07	0.04	0.08	0.08	0.24	0.05	0.15	0.4
PA	0.15	0.46	0.62	0.11	0.52	0.73	0.09	0.61	0.91
CD	0.7	0.58	0.44	0.66	0.62	0.64	0.63	0.78	0.54
PA2	0.48	0.68	0.73	0.38	0.69	0.91	0.24	0.84	0.98

Figure 2. Item misfit detection rates (2PLM fit to 3PLM responses) for the studied item (SI) while the rest items (RIs) follow the 2PLM.

Note. The x-axis represents the item parameter values for the condition: item discrimination (a), item difficulty (b), and pseudo-guessing parameter (g). For example, a(1)_b(-1)_g(.2) means $a = 1, b = -1$, and $g = 0.2$.

The difference between the first and the second occasion is whether the data for the RIs were responses following either the 3PLM or the 2PLM. A higher value of rejection rates indicates a better item misfit detection for a given item measure. The results show that, overall, PA2 resulted in the best performance for detecting a misfit item. Also, the performance of PA, in general, though less than PA2, was better than or very comparable to the existing statistical item fit test procedures, Q_1 and $S - X_2$.

Summary of Major Findings

Summaries of the major findings from the results are described below.

When an item is fitted by a correct model

First, the Type I error rates of S-X^2 were closer to the nominal α level of .05 (06 for the 2PLM data and .5 for the 3PLM data on average) than those of Q_1, whose Type I error rates were .07 and .12 on average for the 2PLM data and for the 3PLM data, respectively.

Second, when item response data follow the 3PLM, Type I error rates of Q1 were inflated in general, reaching the maximum of .19, which is almost 4 times as large as the nominal significance rate.

Third, the predictive power indicated by PA, PA2, and CD was higher in the 2PLM data than the 3PLM data. PA showed 79% and 74%, on average, of correct classification for the 2PLM data and the 3PLM data, respectively. PA2 showed 70% and 64 % on average for the 2PLM and the 3PLM data, respectively. The average values of CD which indicates overall correspondence between predicted and observed response data were .83 and .76 on average for the 2PLM and the 3PLM data.

Fourth, the predictive power indicated by PA, PA2, and CD increased as the item discrimination (a) increased. For a=1, 1.5, and 2, PA exhibited 71%, 77%, and 81% of correct classification on average, respectively. As to PA2, the correct classification rate was, 61%, 68%, and 72%, respectively for a=1, 1.5, and 2. The values of CD also increased as a=1,1.5, and 2, showing .73, .80, and .85.

When an item is fitted by a misfit model

First, the rejection rates of a misfit model or the detection rates of a misfit item, in general, increased as the item slope parameter (a) and item difficulty parameter (b) increased.

Second, the worst performance of all item fit measures except for CD, with data for the RIs following the 3PLM was found when an item was easy (b=-1)

Third, when data for RIs follow the 2PLM, PA, PA2 and CD were always better performers for detecting a misfit item across all conditions than the existing item fit test procedures of Q_1 ad S-X_2.

Fourth, among the three item fit measures of PA, PA2, and CD, the overall performance of CD was the poorest when data for RIs follow the 3PLM, while its performance was better than Q_1 and S-X^2 when data for the RIs follow the 2PLM.

Fifth, when data for RIs follow the 3PLM, the order of the item fit procedures in their detection of a misfit item, according to the average detection rates, was PA2 (57%) > PA (48%) > Q_1 (39%) > S-X^2 (30%) > CD (13%). When data for RIs follow the 2PLM, the order was PA2 (66%) > CD (62%) > PA (47%) > Q_1 (13%) > S-X^2 (11%). The better performance of Q_1 than S-X^2 is partly due to due to the liberal nature of Q_1 whose Type I error rate is higher than S-X^2. Sixth, PA2 and PA in general were better in detecting a misfit item than Q_1 and S-X^2.

Seventh, the differences of the values in PA, PA2, and CD between a correct model and a misfit model were small. The differences on average were found at the second through the fourth decimal places. For example, in a condition where PA2 shows the detection rate of 98% and CD shows 54%, PA2 for the correct model (3PLM) minus PA2 for a misfit model (2PLM) = .6109 - .5987 = .0123; and CD for 3PLM minus CD for 2PLM = 0.7261 – 0.7257 = 0.0004, on average.

Discussion and Conclusion

Item fit assessment is part of the model-data fit assessment when IRT models are used. The popular conventional item fit measures in research and, especially in practice are Q_1 and $S-X^2$. To make up for the fact that they were developed as statistical significance test procedures and to facilitate the understanding and communication of the item fit results with clients in assessment programs or consumers in practice, this study proposed the use of

descriptive item measures that are based on predictive power of a fitted model, which are PA, PA2 and CD. Also, the performance for detecting a misfit item by comparing these descriptive item fit measures between different models (correct and incorrect model in this study) was investigated and compared with the statistical test procedures of Q_1 and $S-X^2$.

The overall results in the current study support the potential utility of PA and PA2 for detecting a misfit item. They may be considered for item fit assessment for 2PLM or 3PLM applications in addition to using the existing Q_1 or $S-X^2$. Unlike $S-X^2$ or Q_1, PA and PA2 have the advantage of easy interpretation. (e.g., 0.75 means 75% of the item responses are correctly predicted by the model.) In addition to the performance shown by PA and PA2, this straightforward interpretation of these item fit measures, which shows the extent of an item fit makes the use of PA or PA2 more suitable for communicating with the end users and consumers in practice Of the two item fit measures, PA and PA2, PA2 is is preferable because of its avoidance of "capitalizing on chance". PA2 also obviates the choice of cut off (such as .5) to generate the predicted item responses. In addition, the detection power of a misfit item by PA2 was higher than PA. Though there is some more computational demand to generate PA2 (for this, see the description of PA and PA2 in the beginning of the method section.), the performance of PA2 was clearly better than PA. Therefore, the additional computational cost deems to be justified.
It was mentioned previously that PA2 lessens the issue of capitalizing on chance. One may use other approaches which also avoid that issue. For instance, using PA with a cross-validation technique may be considered in a future study. Also, the model violation or a misfit item in this study was created using the 3PLM because a major focus, as described in the introduction, was on detecting item response data where data are contaminated by test-taker guessing in the way the 3PLM posits. Investigations of PA2 and PA with more diverse types of violation could be considered (e.g., item response data which show monotonicity in theta (latent trait) but not exactly following the 2PLM or 3PLM item response function form.). In addition, varying the test specifications and data size such as different test lengths and sample sizes with a more complex situation where multiple misfit items exist could be further considered in the future. Lastly, comparisons of PA2 and PA with more recently proposed item statistics **[11]** Kondratek, 2022; **[12]** Kolher, Robitzsch, & Hartig, 2020; **[13]** Maydeu-Olivares & Liu, 2015; [20] Sinharay, 2006, **[21]** Suarez-Falcon & Glas, 2003) could be another line of future research.

References

1. Agresti, A. (2019). *Categorical data analysis* (second edition). Hoboken, NJ: John Wiley & Sons, Inc.

2. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee'sability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

3. Bock, R. D. (1960). *Methods and applications of optimal scaling*. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory.

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

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

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

7. de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

8. Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah,NJ: Lawrence-Erlbaum.

9. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

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

11. Kondratek, B. (2022). Item-fit statistic based on posterior probabilities of membership in ability groups. *Applied Psychological Measurement*. First published online June 20,2022. https://doi.org/10.1177/01466216221108061

12. Kolher, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected RMSD item fit statistic: an evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*. *45*, 251-273.

13. Maydeu-Olivares, A., & Liu, Y. (2015). Item diagnostics in multivariate discrete data. Psychological Methods, 20(2), 276–292.

14. McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics, *Applied Psychological Measurement*, *9*, 49-57

15. New York State Education Department. (2018). *New York stat testing program 2018: English Language Arts and Mathematics Grades 3-8*. Apple Valley, MN: Questar Assessment Inc.

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

17. Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: Anitem fit index for use with dichotomous item response theory models. *AppliedPsychological Measuremen*t, *27*, 289-298.

18. R core team (2021). R: A language and environment for statistical computing. R Foundation For statistical computing, Vienna, Austria.

19. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves, *BMC Bioinformatics*, *12*, p. 77. DOI: 10.1186/1471-2105-12-77

20. Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. British Journal of Mathematical and Statistical Psychology, 59, 429-449.

21. Suarez-Falcon, J. C., & Glass, C. A. W. Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*, 127-143.

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

23. Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.). *Educational Measurement* (fourth edition) (pp. 111-154). West Port, CT: Praeger Publishers.