Journal of Modern Applied Statistical Methods

Volume 18 | Issue 2 Article 26

10-2-2020

An Investigation of Chi-Square and Entropy Based Methods of Item-Fit Using Item level Contamination in Item Response Theory

William R. Dardick
The George Washington University, wdardick@gwu.edu

Brandi A. Weiss
The George Washington University, weissba@gwu.edu

Recommended Citation

Dardick, W. R., & Weiss, B. A. (2019). An Investigation of chi-square and entropy based methods of item-fit using item level contamination in item response theory. Journal of Modern Applied Statistical Methods, 18(2), eP3208. doi: 10.22237/jmasm/1604190480

Journal of Modern Applied Statistical Methods
November 2019, Vol. 18, No. 2, eP3208.
doi: 10.22237/jmasm/1604190480

בס"ד

ISSN 1538 − 9472

An Investigation of Chi-Square and Entropy Based Methods of Item-Fit Using Item Level Contamination in Item Response Theory

William R. Dardick
The George Washington University
Washington, DC

Brandi A. Weiss
The George Washington University
Washington, DC

New variants of entropy as measures of item-fit in item response theory are investigated. Monte Carlo simulation(s) examine aberrant conditions of item-level misfit to evaluate relative (compare EMR_j, X^2, G^2, S-X^2, and PV-Q_1) and absolute (Type I error and empirical power) performance. EMR_j has utility in discovering misfit.

Keywords: Item response theory, IRT item-fit, IRT model fit, Monte Carlo simulation, entropy

Introduction

In social sciences, item response theory (IRT) is a modeling technique used in the measurement of continuous latent psychological constructs which assesses the latent constructs and the items used to measure the latent constructs. The multitude of benefits from using IRT are recognized when the proposed assumptions and model-data fit hold. Inevitably when assessing the fit of our model to the data we find patterns of items and persons that behave in ways inconsistent with the model we have under consideration. In the psychometric tool bag for evaluating assessments we use statistical instruments to help compare models, or flag aberrant items and persons. The ability to determine if a model and data are aligned, or if the individual items fit adequately, is limited by the accuracy and utility of these tools.

Two reasonable measures of accuracy of a new item-fit statistic are empirical Type I error and power. A second consideration when evaluating a new fit statistic

doi: 10.22237/jmasm/1604190480 | Accepted: March 21, 2020; Published: October 2, 2020.
Correspondence: William R. Dardick, wdardick@gwu.edu

for item-fit would be the utility of considering statistics that complement but are not grounded in the χ^2 and likelihood ratio approaches (e.g. Ames & Penfield, 2015; Bock, 1960; McKinley & Mills, 1985; Orlando & Thissen, 2000; Wright & Panchapakesan, 1969; Yen, 1981) and can help parse out novel information regarding model-data fit. In this investigation as we introduce the entropy misfit ratio (*EMR*) as a new statistic for item-fit we keep in mind the principles of accuracy and utility.

The aim of this study is to: 1) introduce new measures of Entropy, Entropy Misfit, and Entropy Misfit Ratio (E_j, EM_j, and EMR_j, respectively) that can be used to estimate item-fit for IRT models with a focus on EMR_j; 2) establish empirically derived cut-off values for three test lengths and two models; 3) conduct a simulation study to examine empirical pseudo-Type I error rate and power for EMR_j compared to commonly utilized goodness-of-fit measures; and 4) discuss the utility of EMR_j along with S-X^2 (Orlando & Thissen, 2000) as a novel approach to item-fit. EMR_j is not intended to replace currently utilized fit statistics, but instead, to demonstrate that it will be useful along with other measures.

IRT and Fit

IRT may be useful to: 1) estimate scores on a continuous latent construct (e.g., ability); 2) construct test instruments; 3) create an item bank; or 4) equate scores across alternate forms of a test. In IRT an item characteristic curve (ICC) provides the probability of a correct answer will be obtained for an item given a particular ability level. For example, in the 2-parameter logistic model (PLM, e.g. 2PLM) the probability the item will be answered correctly can be expressed as

$$\mathrm{P}\left(X=1|\,\theta_i,b_j,a_j\right)=\frac{\mathrm{e}^{a_j\left(\theta_i-b_j\right)}}{1+\mathrm{e}^{a_j\left(\theta_i-b_j\right)}}, \tag{1}$$

where e is the exponential constant 2.718…, and θ is the latent trait of a given respondent's ability. For a given item, *j*, there are two parameters in the model, *a* and *b*, where: *a* is the slope at the inflection point of the model or discrimination and *b* is the item difficulty, the abscissa value at the point of inflection, *a*. Typical models can further be obtained by: 1) setting the constant *a* to 1 (the Rasch model); 2) estimating the constant *a* to be the same value across all items (1PLM); or 3) adding the lower asymptote and pseudochance parameter *c* to the model (3PLM). This unidimensional IRT model is just one amongst a set of extensions. For example, we could add additional parameters to unidimensional dichotomous

models, change to focus on nominal and ordinal polytomous models, nonparametric models, and unfolding models or consider more dimensions with mixed and multidimensional models. The current study utilizes the 2PLM for IRT to assess fit, focusing on the dichotomous, unidimensional, correct/incorrect models.

IRT is based on the assumption that a model is able to reproduce the observed data. Although it is not expected for any model to reproduce the data perfectly, the beneficial properties of IRT only hold provided that a model is able to reproduce the data well. Models that do not fit the data will result in biased item parameter estimates, inaccurate ability estimates (Ames & Penfield, 2015; Wainer & Thissen, 1987; Yen, 1981), and inaccurate standard error estimates, and often discarded during test development phases.

The evaluation of model-data fit in IRT separates out model level fit, item-fit, and person-fit (Rupp, 2013). Measures of model-data fit examine fit at the model level by considering whether the model fits for all items and persons within the observed dataset. However, measures of item-fit (i.e. item-fit analysis), are also useful during the test development phase as they can be used for item selection or revision. Many articles/book chapters intermingle these types of fit together. This may partially be because if all items fit then you will have model-level fit. However, having model-level fit does not necessarily mean all items fit (De Ayala, 2019). For more discussions on model fit in IRT we refer readers to Ames and Penfield (2015), De Ayala (2019), Hambleton and Swaminathan, (1985), and Swaminathan et al. (2007). In the current study we focus on fit at the item level.

Traditional methods of examining item-fit in IRT focus on quantifying how well observed item responses fit the predicted responses based on the ICCs. These measures utilize the residuals (i.e., the difference between a person's observed response and their predicted response based on the ICC). Frequently utilized residual-based measures of data-model fit in IRT include: Bock's (1960) χ^2 square, Yen's (1981) Q_1 (also referred to as χ^2 by many statistical software programs), and the G^2 statistic (aka, the likelihood ratio; McKinley & Mills, 1985). These three measures are often criticized because ability estimates (which are model dependent) are binned (grouped) together and used to calculate the statistic (Ames & Penfield, 2015). These bins are based on arbitrary cut points that are sample specific. The general issue with binning and averaging is an attempt to balance sampling error and bias (Fox, 2015). A small number of wide bins reduces sampling error but increase bias, while a large number of narrow bins increase sampling error and reduces bias. When we have extremely large sample sizes this issue can be mitigated. Additionally, ability estimates are model-dependent so true distributions of these statistics are unknown.

There have been advances in the use of chi-square-based item-fit statistics over the past few decades. Orlando and Thissen's (2000) S-χ^2 and S-G^2 statistics use observed test scores (summated scores) instead of theta ability estimates, and both measures have performed well with short test lengths (Chon et al., 2010). Consequently, these measures require a theoretical dissonance, because it is required to fall back on observed-score ability estimates from classical test theory even though ability is a latent variable. Orlando and Thissen discussed issues when estimating the 2PLM when generated data are 3PLM. S-χ^2 and S-G^2 do not perform well in detecting this type of contamination. Stone's (2000) χ^2* and G^2* statistics are pseudo-observed score methods that may improve on more traditional measures, however, past research has found these measures take excessively long time to compute for a single dataset (Chon et al., 2010). Chalmers and Ng's (2017) "plausible-value (PV) imputations (Mislevy, 1991) and parametric bootstrap techniques (Hope, 1968)" (p. 373) PV-Q_1 and PV-Q_1* tended to have conservative to normal Type I error rates respectively. Empirical power was somewhat lower than χ^2* but higher than S-χ^2. As PV-Q_1* is computationally, similar to χ^2* Chalmers and Ng recommend PV-Q_1.and S-χ^2 for test lengths over 20 and reserve the use of χ^2* PV-Q_1*when parametric bootstrapping is computationally feasible and when a local minimum is unlikely to occur. Evaluation of power in Chalmers and Ng's study was limited to three items generated to be aberrant conditions to the estimated 3PLM. We refer readers to Orlando and Thissen (2000), Stone and Zhang (2010), and Chalmers and Ng (2017) for a more in-depth discussion of some of these measures.

There is not a single superior measure of item-fit. Instead, methodologists recommend using multiple types of fit statistics to evaluate fit (Ames & Penfield, 2015; Sinharay, 2006; van der Linden & Hambleton, 1997). De Ayala (2019) recommends using both statistical and graphical analyses. One type of measure not yet explored in the item-fit IRT literature is entropy-based measures, which could be used to assess the amount of separation between groups. Previously, entropy was used as an approximate measure of data-model fit, when it falls within some recommended range, to quantify how well individuals are classified into latent classes (Celeux & Soromenho, 1996; Henson et al., 2007), to quantify quality of separation between groups in logistic regression models (Weiss & Dardick, 2016), and to detect person misfit in IRT models (Dardick & Weiss, 2017). Many measures of item-level fit focus on dichotomous hypothesis testing procedure (e.g., χ^2 and G^2). In contrast, approximate fit indices, are continuous measures used to supplement tests of statistical significance by indicating the degree of fit (Hu & Bentler, 1999). An entropy-based item-fit index may not discriminate between

items the same way as the residual-based measures (e.g., χ^2, G^2, S-X^2, S-G^2, χ^2*, and G^2*) and therefore provide alternative information to help make decisions on model fit. Finally, entropy-based measures do not rely on binning ability estimates based on model dependent arbitrary cut points as residual-based measures do. In summary, entropy-based measures of model fit may be useful in conjunction with existing residual-based analyses.

Entropy in Latent Class Analysis

Entropy is a classification-based approach assisting researchers in determining the number of latent classes within a dataset (Celeux & Soromenho, 1996; Clark & Muthén, 2009; Ramaswamy et al., 1993; Henson et al., 2007; Pastor & Gagné, 2013). Entropy captures the separation amongst classes when $K > 1$ and in Latent Class Analysis (LCA) the entropy index is calculated as

$$E_k = 1 - \frac{E_k^*}{N \ln K}, \tag{2}$$

where N is the total sample size, K is the number of latent classes, and E_k^* is defined as

$$E_k^* = \sum_{i=1}^{I} \sum_{k=1}^{K} \left(-p_{ik} * \ln p_{ik} \right), \tag{3}$$

where $0 \leq E_k \leq 1$ and p_{ik} represents the conditional posterior probabilities calculated for each observation, i, and represents the probability of membership in each of the k classes.

When posterior probabilities are equal across classes, $E_k = 0$. As posterior probabilities move further from a threshold used to differentiate between classes (In a two-class model this cut-point is often set to .50), the fuzziness of class membership decreases due to the increased separation of classes, and thus the value of entropy also increases to reflect this increased separation. As a clarification note, in other fields entropy can be thought of as disorder. Thus, the more entropy, the less separation between groups.

Entropy in LCA, represented by E_k, however, is subtracted from 1 [equation (2)] so that values close to 0 indicate less separation (i.e., more disorder) and values close to 1 indicate more separation between classes (i.e., less disorder). In IRT, entropy also capitalizes on group separation, where larger values indicate order, or

more division of responses, clarifying group membership. However, values close to 0 indicate poor separation between the centroids of the conditional parametric distributions for each class (Ramaswamy et al., 1993). In LCA, higher values of entropy are more desirable because they are indicative of good classification quality.

Entropy Measures of Person-Fit

Entropy for Person-Fit

The standardized entropy index (bounded 0,1) from LCA was adapted for use in dichotomous IRT models to examine person-fit (Dardick & Weiss, 2017) where across J items and K response options (e.g., correct/incorrect, where previously K was the number of latent classes in LCA, now the equivalent concept is response option) E_i is defined by

$$E_i = 1 - \frac{E_i^*}{J \ln K}, \tag{4}$$

where E_i^* represents the unstandardized entropy of the set of items j for person i and is defined by

$$E_i^* = \sum_{j=1}^{J} \sum_{k=1}^{K} \left(-p_{ijk} * \ln p_{ijk} \right), \tag{5}$$

where $E_i^* \geq 0$ and p_{ijk} is the model predicted probabilities across all items for each person and each categories k. E_i values are bound between 0 and 1 and higher values are indicative of more distinct separation among category responses.

Entropy Weighted for Misfits

Knowing the observed scores for person i on item j, Dardick and Weiss (2017) modified the calculation of entropy to partition predicted probabilities of a response (0, 1) into those predicted probabilities that fit (correct classification), and those predicted probabilities that do not fit (incorrect classification). Following from equations (2) and (3),

$$EM_i = \frac{\sum_{j=1}^{J}\left\{\left(1-\frac{EM_i^*}{\ln K}\right)\left(1-x_j\right)\right\}}{J}, \tag{6}$$

where J is the number of items, K is the number of response categories (e.g., 2 for dichotomous correct/incorrect items), and EM_i^* is defined by

$$EM_i^* = \sum_{k=1}^{K}\left(-p_{ijk} * \ln p_{ijk}\right), \tag{7}$$

where p_{ijk} again represents the predicted probabilities from the PLM calculated across all items for each person and denotes the probability the person i endorses an item with a correct response. x_j is a weight and is defined as

$$x_j = \begin{cases} 1 & \text{correct classification} \\ 0 & \text{incorrect classification} \end{cases}. \tag{8}$$

EM_i weights each item-person combination so that only an item's predicted probabilities who are incorrectly classified contribute to EM_i. Large values of EM_i reflect poor classification, and thus small values of EM_i are desirable because they indicate there is a small amount of misfit.

Entropy Misfit Ratio

The entropic misfit ratio, EMR_i, is the ratio of EM_i, to, E_i, representing the strength of fit that is attributed to misfit and calculated as

$$EMR_i = \frac{EM_i}{E_i}. \tag{9}$$

This equation provides a relative understanding of entropy regardless of the initial magnitude of total entropy. EMR_i ranges from 0 to 1, where a value of 0 represents all predicted classifications were correct (i.e., no misfit), and a value of 1 indicates all predicted classifications were incorrect and thus EM_i and E_i are equal. Smaller values of EMR_i are indicative of less misfit (i.e., fewer incorrect predictions) and are thus more desirable.

This ratio may enhance interpretation of entropy as misfit is comparable with different item-θ combinations on different tests. Simulation studies comparing these entropy measures (E_i, EM_i,, and EMR_i) to likelihood-based statistics, l_z, (Drasgow et al., 1985) l_z* (Snijders, 2001) and residual based measures, *U,* and *W* (Wright & Masters, 1982; Wright & Stone, 1979), indicated EM_i, and EMR_i were successfully able to detect aberrant response patterns as approximate person-fit measures for IRT models (Dardick & Weiss, 2017).

Extending Entropy Measures to Item-Fit

Entropy for Item-Fit

The current study extends the entropic fit measures from person- to item-fit in IRT. Here we follow the same logic used to calculate the person-fit entropy measures but derive them across persons (instead of items) to acquire item level statistics. Item transformation for summing across persons instead of items impacts the denominator where across *N* persons and *K* response options (e.g., correct/incorrect) *Ej* is entropy for a given item and is defined by

$$E_j = 1 - \frac{E_j^*}{N \ln K}, \tag{10}$$

where E_j^* represents the unstandardized entropy of the set of persons for item *j*:

$$E_j^* = \sum_{i=1}^{N} \sum_{k=1}^{K} \left(-p_{ijk} * \ln p_{ijk} \right), \tag{11}$$

where $E_j^* \geq 0$ and as defined previously p_{ijk} is the model predicted probabilities across all persons and *k* categories. Entropy, E_j, is calculated for each item and represents a standardized format of E_j^* in which total entropy is bound between 0 and 1, and higher values are indicative of more distinct separation among categories of response.

Item-Level Entropy Weighted for Misfit

The entropy misfit calculations can also be modified at the item level. Knowing the observed scores for person *i* on item *j* allows the entropy to partition predicted probabilities of a response (where the threshold .5 splits our prediction response

above .5 is 1, at or below is 0) into those that fit (correct classification), and those that do not fit (incorrect classification). Following from equations (8) and (9),

$$EM_j = \frac{\sum_{i=1}^{N}\left\{\left(1-\frac{EM_{ij}^{*}}{\ln K}\right)\left(1-x_{ij}\right)\right\}}{N}, \tag{12}$$

where N is the number of persons, K is the number of response categories (e.g., 2 for dichotomous correct/incorrect items), and EM_j^{*} is defined by

$$EM_{ij}^{*} = \sum_{k=1}^{K}\left(-p_{ijk} * \ln p_{ijk}\right), \tag{13}$$

where p_{ijk} again represents the predicted probabilities from the PLM calculated across all persons for each item and denotes the probability that person i endorses an item with a correct response. x_{ij} is a weight (i.e. based on the observed item response data) and is defined as

$$x_{ij} = \begin{cases} 1 & \text{correct classification} \\ 0 & \text{incorrect classification} \end{cases}. \tag{14}$$

EM_j weights each person-item combination so that only person-items who are incorrectly classified contribute to EM_j. Large values of EM_j reflect poor classification, and thus small values of EM_j are desirable because they indicate there is a small amount of misfit.

Entropy Misfit Ratio at the Item Level

The ratio of entropic misfit ratio, *EMRj*, represents the amount of misfit relative to the to the total amount of entropy and is calculated the same way it is for person-level entropy

$$EMR_j = \frac{EM_j}{E_j}. \tag{15}$$

EMR_j provides a relative understanding of entropy regardless of the initial magnitude of total entropy, ranges from 0 (all predicted classifications were correct;

i.e., no misfit) to 1 (all predicted classifications were incorrect and thus EM_i and E_i are equal). Smaller values of EMR_j are indicative of less misfit and are thus more desirable. Chi-square based measures focus on correct classification without regard to the degree of separation amongst group membership. Entropy, however, is telling the story of degree of separation amongst groups, although not considering correct classification. Therefore, correct classification and separation are incorporated into the EMR_j measure.

Cell-Level Entropy

Dardick and Weiss (2017) showed how entropy can be calculated at the person level [equations (4) and (5)] whereas the current study shows how entropy can be calculated at the item level [equations (10) and (11)]. Although the focus of the current study is on item-fit, two important considerations are provided: 1) measures of person-fit, item-fit and model-fit entropy can be calculated simultaneously for the same data set, and 2) this reasoning carries to *EM* and *EMR* in that all entropy variants can be calculated at the person- [equations (6) to (8)] and item-level [equations (12) to (15)].

Many measures of fit (e.g., chi-square, OUTFIT, and INFIT) are based on person-by-item cell-level residuals. The cell-level residuals are combined either across persons (to create a person-level fit statistic) or down items (to create an item-level fit statistic). Entropy is calculated in a similar manner. Consider a data set that represents a matrix of correct/incorrect responses for n persons on j items. When calculating entropy, obtain a matrix of cell-level entropy values for every person-by-item combination. To calculate unstandardized entropy at the person level, values are summed across the item-level cells for each person; to calculate unstandardized entropy at the item level, values are summed down the person-level cells for each item. Entropy values are then standardized and averaged by the number of persons (or items) and response categories (e.g., correct/incorrect).

The Current Study

The entropy measures in item-fit in IRT are considered. There were two goals of this study. First, null conditions are explored for the distribution of EMR_j and establish empirically derived cut points under various test lengths and models. Second, the aim is to determine how well EMR_j could correctly identify misfitting items (empirical power) and control Type I error in comparison to other measures of item-fit (S-X^2, χ^2, and G^2). We investigated E_j, EM_j, and EMR_j in comparison to S-X^2, χ^2, and G^2. In determining which measures to investigate for item-fit in

comparison to entropy measures, some commonly used measures are considered, with examples for this study coming from PROC IRT (χ^2 and G^2; SAS Institute Inc., 2017) from SAS and the R 'mirt' package (S-X^2; Chalmers, 2019; Chalmers & Ng, 2017). Contamination is defined as participant responses that were not generated to conform to the IRT model.

Methods

Monte-Carlo simulation was used to first determine empirical cut points for entropy-based measures, and second using those cut points to evaluate empirical power and Type I error rates across various conditions of models, test lengths, and misfit.

Simulation Design Study 1

In this stage of investigation, baseline models were examined in which all participants were generated to fit the appropriate IRT model. Two models (1PLM and 2PLM) and three test lengths representing short, medium, and long tests (10-, 20-, and 40-items) were considered. There were 6 conditions for this study (2 models × 3 test lengths). Empirical one-tailed cut points were derived for entropy at the upper-bound 95th percentile of the distributions.

Simulation Design Study 2

Contamination is defined as participant responses that were generated to not conform to the IRT model. The assessment of item misfit could be evaluated in many ways, with or without contaminating persons specifically but as with any simulation design, the contamination conditions are limited to be able to add value to possible scenarios. More specifically, the contaminated persons within the contaminated items were modified to have probabilities of either .25 or .50 of endorsing a correct response, which might represent guessing 25% and modified guessing of 50% correct responses and were chosen to have two consistent but different levels of contamination. The subtest of contaminated items varied to contain different percentages of contaminated items (for the 10, 20 and 40 item tests): 10% (1-, 2-, or 4-items, respectively), 25% (2-, 5-, or 10-items), and 50% (5-, 10-, or 20-items) contamination. The percent of simulated persons contaminated within the subtests were varied to be 10%, 25%, and 50%. For example, when 25% of items are contaminated, in the 20-item condition, with 25% of simulated persons

contaminated to have probabilities of .25 guessing; then there are 5 items contaminated with 25% of simulated persons guessing at a probability .25.

The focus was on the 1PLM and 2PLM. Although there are numerous potential extensions of IRT models (e.g., 3PLM, 4PLM, polytomous, multidimensional, or unordered) this decision was made for four reasons. First, the primary interest is with discrimination and difficulty. Second, the 1PLM and 2PLM are popular IRT models. Third, the type of contamination may interact with the pseudo-guessing parameter in the 3PLM. Fourth, variables manipulated were more important than an extension to more complex IRT models.

Overall test length was either 10-, 20-, or 40- items. These test lengths are similar to those used in other simulation studies (Chon et al., 2010; Dardick & Weiss, 2017; Haberman et al., 2013). Test lengths of 10 and 20 assist in considerations of sensitivity. Although large scale testing organizations often have more items, other testing scenarios (e.g. educational, certification) and future trends (e.g. formative and summative assessment, teacher empowered assessment, Dardick & Choi, 2016) will consider much shorter testing experiences that require psychometric rigor of data-model fit.

For the second simulation data were generated in a fully-crossed $2 \times 3 \times 3 \times 2 \times 3$ design: 2 types of contamination (25% and 50%), 3 percentages of items contaminated (10%, 25%, and 50%), 3 percentages of persons contaminated (10%, 25%, and 50%), 2 model types (1PLM and 2PLM), and 3 test lengths (10-, 20-, 40-). This resulted in 108 conditions evaluated across the cut points (95^{th} percentiles).

Evaluation Criteria

The empirically derived cut points for E_j, EM_j, and EMR_j from the first simulation study were used in the second simulation study to evaluate the empirical power and Type I error rates for entropy measures. Rupp (2013) described four methods of investigating empirical power and Type I error, and recommended researchers "compute the empirical sampling distributions and always use the appropriate empirically-derived cut-off values that ensure nominal Type I error rates for computing power rates (best method with highest precision)" (p. 19). For clarification purposes, reference to Type I error means the empirically derived Type I error rate sometimes referred to as empirical pseudo-Type I error.

Entropy measures E_j, EM_j, and EMR_j were compared with Orlando and Thissen's (2000) S-X^2, Yen's (1981) χ^2 statistic, McKinley and Mills (1985) likelihood ratio (i.e., G^2), and PV-Q_1 (Chalmer's chi-square) for all conditions. A customary alpha level of .05 was used to evaluate performance of the S-X^2, χ^2, G^2,

and PV-Q_1 statistics. Although other measures of item-fit exist, these measures were selected because they are taught in IRT textbooks, frequently used measures of item-fit in practice, easy to compute, readily available in many IRT statistical software packages, or sometimes the only measures of item-fit in statistical software packages. Other measures such as Stone's (2000) χ^2* and G^2* exist, however, previous research found inflated empirical pseudo-Type I error rates associated with these pseudo-observed score and noted that it took 40 minutes to calculate these measures for a single dataset for 40-items and 4,000 persons (Chon et al., 2010). Thus, they were not investigated in the current study. Chalmers and Ng's (2017) bootstrap adaptation was computed for a few replications, but it also took substantial time for a single replication in one condition, and given recommendations mentioned previously of computational intensity and local minima it was not investigated further for this study.

Data Generation

For each of the 6 (baseline) and 108 (contaminated) conditions datasets were generated to contain responses for 1,000 persons. One thousand replications were simulated for each condition. Previous studies used fewer than 1000 replications (Chon et al., 2010; Haberman et al., 2013; Stone & Zhang, 2003; Feinberg & Rubright, 2016). Data were simulated using either the 1PLM or 2 PLM with randomly generated error around the measures and incorporated misfit using two levels of probabilities .25 and .5. More specifically, item difficulty parameters *b* were generated from a random normal distribution N(0, 1) with constrains of $\pm$ 2 (e.g., Emons et al., 2002; Karabatsos, 2003; Zhang & Walker, 2008). The item discrimination parameter, slope *a*, was randomly generated from a realistic range of values, using a log normal distribution LN(0, 0.2) which also provides values similar to those recommended by Rupp (2013). In SAS (SAS Institute Inc., 2017), code used to simulate data to these distributions are RAND ("NORMAL", 0, 1) for the N(0, 1) and RAND("LOGNORMAL", 0, .2) for LN(0, .02). Data that were generated using a 1PLM were fit with a 1PLM, and data that were generated using a 2PLM and fit with a 2PLM.

Calibration and Distribution

The IRT parameters for the 1PLM and 2PLM were estimated using the Proc IRT procedure in SAS (SAS Institute Inc., 2017). The estimation method for item parameters in Proc IRT to obtain marginal maximum likelihood estimation is the gradient-based convergence criteria for the quasi-Newton algorithm. Statistics for χ^2 and G^2 were attained from the SAS IRT procedure, while EMR_i was calculated

using SAS/IML. SAS/IML was used to interface with R language as S-X^2 and PV-Q_1 are not currently available in SAS. We used the 'itemfit' function within the R package 'mirt' (Chalmers, 2019) to study fit statistics including S-X^2 and PV-Q_1. The 'mirt' package uses full information maximum likelihood estimation with a default of expected *a-posteriori* estimation.

Results

Results are presented for the two Monte Carlo simulation studies. In the first study, baseline models were simulated to obtain empirically-derived cut points for the various test lengths and models. A second simulation was conducted in which different amounts and types of contamination were incorporated in order to evaluate the relative (EMR_j , S-X^2, χ^2, G^2, PV-Q_1) and absolute (empirical pseudo-Type I error rate and empirical power) performance of entropy as an item-fit index. Results were evaluated in relation to the test length, percent of contaminated items, the percent of contaminated persons, the type of contamination, and the cut-point utilized.

Both simulation studies were conducted using E_j, EM_j, and EMR_j. However, results for EMR_j rather than E_j and EM_j were the focus of our analysis because insufficient statistical power and high Type I error rates were found for E_j in all conditions; and low statistical power was found for EM_j in all conditions. Furthermore, previous research emphasized the advantages of EMR_j for successful detection of aberrant response patterns, potential usefulness for subtests with a small number of items, and ability to identify separation between contaminated and uncontaminated subgroups (Dardick & Weiss, 2017).

Simulation Study 1

Results for baseline models (i.e., no contamination) are presented for 6 conditions: 3 test lengths × 2 models. Table 1 contains the descriptive statistics for the EMR_j statistic. Average EMR_j values increased as the number of test items increased. Also, average EMR_j values and their standard deviations were slightly higher (≈ .02 points) for the 2PLM in comparison to the 1PLM.

Table 1. Means, standard deviations, empirically derived 95 percentile cut-points for of baseline uncontaminated conditions of EMR_j

Model	Test length	EMR M(SD)	EMR 95%
1PLM	10	0.13(0.017)	0.156
	20	0.148(0.025)	0.185
	40	0.159(0.030)	0.203
2PLM	10	0.133(0.042)	0.208
	20	0.150(0.044)	0.226
	40	0.161(0.045)	0.237

Note: Means (M) and Standard Deviations for the Entropy Misfit Ration (*EMR*) for three test lengths, 10, 20, and 40 over two models the 1PLM and 2PLM

Cut points were empirically derived for the 95th percentiles in the distribution for each of the 6 baseline conditions with no contamination based on Rupp's (2013) recommendation. These cut points are shown in Table 1 and followed the same pattern as the descriptive statistics. That is, the 95th percentile for EMR_j increased as the test length increased and as the model increased from a 1PLM to 2PLM. Figure 1 shows the distributions of EMR_j collapsed across all items for each of the 6 baseline conditions. In the general, the distributions for the 1PLM model were a little more negatively skewed in comparison to the distributions for the 2PLM which were slightly more positively skewed.

True and estimated parameter estimates were compared for item difficulty and discrimination (slope) across test length. For the 1PLM, average biases for item difficulty were < 0.001 ($SD < 0.09$). For the 2PLM, average biases for item difficulty were < 0.005 ($SD < 0.125$) and item discrimination were less than 0.006 ($SD < 0.130$) for all test lengths. Given the lack of bias in the baseline conditions, 1000 replications are more than sufficient for obtaining empirically derived cut-points for EMR.

Simulation Study 2

The goal of the second simulation study was to investigate Type I error and empirical statistical power for the EMR_j, χ^2, and G^2 statistics, S-X^2, and PV-Q_1. Results for this study are presented for the 2 models (1PLM and the 2PLM), 3 test lengths (10-, 20-, 40-items), the percent of contaminated items (10%, 25%, and 50%), the percent of contaminated persons (10%, 25%, and 50%), and the 2 types of contaminated responses (25% and 50%).

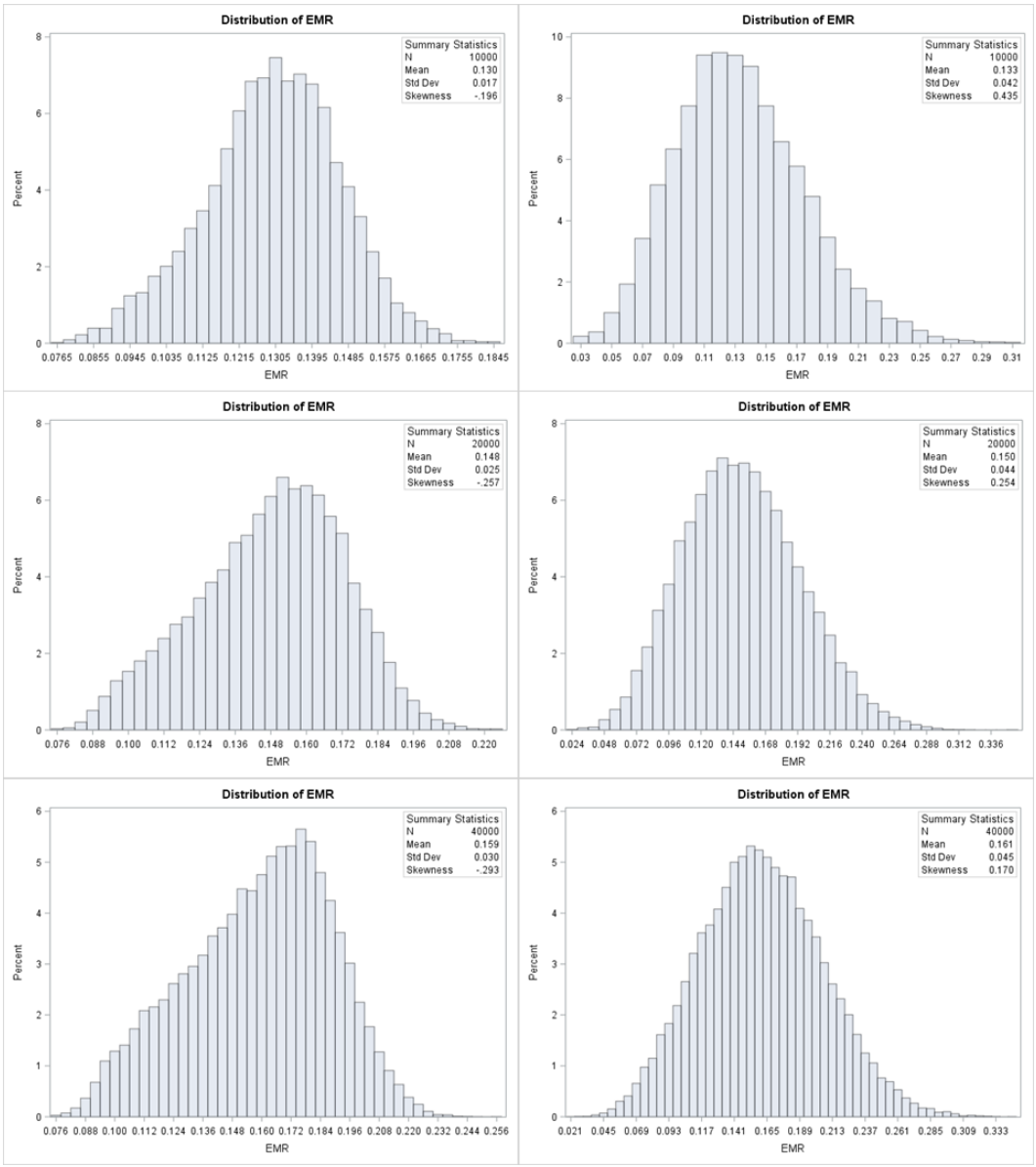

Figure 1. Distributions for the 6 baseline conditions in simulation one; left side is 1PLM, right side is 2PLM; from top to bottom 10, 20, 40 item test; summary statistics are included: *N* = number of items, Mean, Std Dev = standard deviation, Skewness

Means and SDs

The means and standard deviations for all statistics in this study (E_j, EM_j, EMR_j, χ^2, G^2, and S-X^2, PV-Q_1) are presented in Tables 2 and 3 for the 1PLM and 2PLM models, respectively. E_j, EM_j, and EMR_j increased as the number of items increased, and mean values were slightly higher for the 2PLM than in the 1PLM model. The average χ^2 and G^2 values decreased as the number of items increased. On the contrary, S-X^2 increased as the number of items increased, while PV-Q_1 values were similar across different numbers of items. In general, chi-square measures values were a little smaller in the 2PLM compared to the 1PLM, with the exception of χ^2 for 10-item test lengths which did not vary. In general, as the amount of contamination increased the chi-square-based statistics also increased.

In order to examine which manipulated variables were related to E_j, EM_j, EMR_j, χ^2, G^2, S-X^2, and PV-Q_1: Factorial between-groups analysis of variance (ANOVA) was conducted for the 1PLM and 2PLM, separately. For each replication the outcomes were the values of the fit statistic (i.e., E_j, EM_j, EMR_j, χ^2, G^2, S-X^2, or PV-Q_1), and the between-groups variables were: item contamination (uncontaminated vs contaminated), test length (10, 20, 40 items), percentage of persons contamination (10%, 25%, 40%), and type of misfit (0.25 or 0.50). Due to the large number of replications, partial eta-squared values were examined (see Table 4) and tests of statistical significance were ignored. In order for an effect to be considered practically meaningful it had to have a partial eta-squared value of .03 or greater. The 4- and 3-way interactions did not explain a meaningful amount of variance in the outcomes. Test length explains the most variance in all the chi-square-based measures, with the exception of PV-Q_1, in which person contamination explained the most variance in values for the 1PLM. On the contrary, item contamination explains the most variance in EMR_j. Thus, EMR_j for item-fit is able to differentiate between contaminated and uncontaminated items more so than the chi-square measures. Item contamination is the only variable that explains a meaningful amount of variance in $E_{j.}$ Partial-eta squared values are an indication that EMR_j differentiates between items differently than the chi-square measures.

Table 2. Means and standard deviations of all conditions 1PLM

Test	Item	Rep	MF	*E* M(SD)	*EM* M(SD)	*EMR* M (SD)	χ^2 (SD)	G^2 M(SD)	S-χ^2 (SD)	PV-Q_1 (SD)
10	0	0	0.00	0.174(0.091)	0.021(0.008)	0.130(0.017)	52.196(16.209)	60.497(20.681)	7.745(3.932)	9.102(1.182)
10	10%	10%	0.25	0.171(0.091)	0.021(0.008)	0.131(0.018)	52.864(16.506)	61.040(20.720)	8.005(4.256)	9.214(1.396)
10	10%	10%	0.50	0.171(0.090)	0.021(0.008)	0.132(0.018)	52.959(16.611)	61.189(20.884)	8.044(4.213)	9.214(1.370)
10	10%	25%	0.25	0.165(0.088)	0.021(0.008)	0.134(0.017)	53.243(17.121)	61.425(21.597)	8.046(4.170)	9.218(1.346)
10	10%	25%	0.50	0.165(0.089)	0.021(0.009)	0.134(0.018)	53.197(16.608)	61.302(20.781)	8.125(4.267)	9.232(1.380)
10	10%	50%	0.25	0.160(0.084)	0.021(0.009)	0.136(0.016)	53.444(16.258)	61.658(20.513)	8.210(4.260)	9.228(1.309)
10	10%	50%	0.50	0.153(0.083)	0.020(0.009)	0.138(0.016)	54.045(16.383)	62.039(20.414)	8.119(4.184)	9.214(1.264)
10	20%	10%	0.25	0.164(0.089)	0.021(0.009)	0.135(0.021)	54.151(17.832)	62.388(22.259)	8.995(6.304)	9.622(2.556)
10	20%	10%	0.50	0.164(0.090)	0.021(0.009)	0.135(0.021)	54.349(17.808)	62.583(22.183)	8.981(6.122)	9.627(2.534)
10	20%	25%	0.25	0.157(0.090)	0.021(0.009)	0.139(0.020)	55.717(18.535)	63.883(22.851)	9.030(5.492)	9.595(2.128)
10	20%	25%	0.50	0.153(0.087)	0.020(0.009)	0.140(0.022)	56.182(18.862)	64.360(23.094)	9.500(6.037)	9.822(2.469)
10	20%	50%	0.25	0.143(0.084)	0.020(0.010)	0.144(0.017)	56.909(18.180)	65.307(22.453)	9.357(4.994)	9.565(1.573)
10	20%	50%	0.50	0.129(0.078)	0.019(0.009)	0.152(0.019)	58.383(18.592)	66.209(22.225)	9.153(4.921)	9.583(1.650)
10	50%	10%	0.25	0.160(0.090)	0.021(0.009)	0.138(0.026)	57.227(19.744)	65.751(24.243)	11.275(12.273)	10.589(5.496)
10	50%	10%	0.50	0.154(0.090)	0.020(0.009)	0.141(0.031)	58.020(20.424)	66.537(24.876)	12.131(14.282)	10.872(6.224)
10	50%	25%	0.25	0.148(0.088)	0.020(0.010)	0.146(0.027)	60.915(22.415)	69.374(27.060)	11.941(9.863)	10.724(4.170)
10	50%	25%	0.50	0.136(0.088)	0.019(0.009)	0.152(0.034)	62.365(24.707)	70.910(29.270)	13.598(11.998)	11.433(5.223)
10	50%	50%	0.25	0.129(0.083)	0.020(0.011)	0.160(0.019)	66.541(24.775)	76.072(29.883)	11.723(6.595)	10.011(1.970)
10	50%	50%	0.50	0.093(0.077)	0.015(0.010)	0.176(0.027)	68.379(26.745)	76.343(30.763)	11.732(6.628)	10.288(2.064)

Note: Item = percent of item contamination, except for the 10 item 25% contamination was rounded to 20%, Rep = percent of replicants contaminated, MF = misfit proportion value, *E* = entropy, *EM* = entropy misfit, *EMR* = entropy misfit ratio, χ^2 = Pearson chi-squared, G^2 = likelihood ratio chi-squared, S-χ^2 = Orlando and Thissen (2000) chi-squared, PV-Q_1 = Chalmers and Ng's (2017) adaptation of Q_1

Table 2 (continued).

Test	Item	Rep	MF	*E* M(SD)	*EM* M(SD)	*EMR* M (SD)	χ^2 (SD)	G^2 M(SD)	S-χ^2 (SD)	PV-Q_1 (SD)
20	0	0	0.00	0.187(0.087)	0.026(0.007)	0.148(0.025)	19.962(7.110)	21.859(8.186)	15.649(5.689)	9.122(1.510)
20	10%	10%	0.25	0.183(0.086)	0.026(0.007)	0.150(0.025)	20.363(7.505)	22.270(8.628)	15.815(5.928)	9.265(1.827)
20	10%	10%	0.50	0.181(0.084)	0.025(0.007)	0.150(0.025)	20.229(7.434)	22.142(8.571)	15.826(5.892)	9.256(1.824)
20	10%	25%	0.25	0.177(0.084)	0.025(0.007)	0.152(0.025)	20.500(7.401)	22.427(8.509)	15.854(5.805)	9.244(1.658)
20	10%	25%	0.50	0.175(0.083)	0.025(0.008)	0.153(0.025)	20.716(7.777)	22.647(8.940)	15.947(5.925)	9.360(1.855)
20	10%	50%	0.25	0.173(0.081)	0.025(0.008)	0.154(0.024)	20.734(7.546)	22.704(8.719)	16.114(5.875)	9.366(1.735)
20	10%	50%	0.50	0.166(0.079)	0.025(0.008)	0.157(0.023)	20.921(7.723)	22.894(8.869)	15.923(5.831)	9.315(1.734)
20	20%	10%	0.25	0.178(0.084)	0.025(0.007)	0.153(0.028)	21.175(8.066)	23.167(9.302)	16.598(7.363)	9.785(3.411)
20	20%	10%	0.50	0.176(0.085)	0.025(0.008)	0.154(0.029)	21.467(8.228)	23.489(9.506)	16.918(8.105)	10.012(3.979)
20	20%	25%	0.25	0.166(0.083)	0.025(0.008)	0.159(0.027)	21.994(8.487)	24.014(9.715)	16.885(6.858)	9.835(2.658)
20	20%	25%	0.50	0.161(0.081)	0.024(0.008)	0.162(0.030)	22.943(9.754)	25.039(11.193)	17.613(7.638)	10.440(3.625)
20	20%	50%	0.25	0.156(0.080)	0.024(0.009)	0.162(0.023)	23.044(9.034)	25.299(10.503)	17.891(6.962)	10.252(2.645)
20	20%	50%	0.50	0.140(0.075)	0.022(0.008)	0.170(0.026)	23.649(10.159)	25.812(11.508)	17.408(6.732)	10.169(2.602)
20	50%	10%	0.25	0.172(0.086)	0.025(0.008)	0.157(0.035)	23.783(8.613)	25.98(10.011)	19.170(14.439)	11.502(8.521)
20	50%	10%	0.50	0.167(0.086)	0.025(0.008)	0.161(0.042)	24.750(8.709)	27.016(10.119)	20.425(17.619)	12.254(10.44)
20	50%	25%	0.25	0.156(0.083)	0.024(0.009)	0.168(0.035)	26.304(10.896)	28.649(12.538)	20.434(11.368)	11.846(5.707)
20	50%	25%	0.50	0.140(0.083)	0.022(0.008)	0.177(0.047)	28.771(13.401)	31.400(15.417)	22.591(13.558)	13.576(7.668)
20	50%	50%	0.25	0.141(0.079)	0.024(0.011)	0.179(0.024)	28.987(12.121)	31.87(14.023)	21.862(9.061)	11.842(4.094)
20	50%	50%	0.50	0.102(0.073)	0.018(0.009)	0.197(0.037)	30.300(17.157)	32.898(19.199)	21.288(9.000)	12.136(3.739)

Note: Item = percent of item contamination, except for the 10 item 25% contamination was rounded to 20%, Rep = percent of replicants contaminated, MF = misfit proportion value, *E* = entropy, *EM* = entropy misfit, *EMR* = entropy misfit ratio, χ^2 = Pearson chi-squared, G^2 = likelihood ratio chi-squared, S-χ^2 = Orlando and Thissen (2000) chi-squared, PV-Q_1 = Chalmers and Ng's (2017) adaptation of Q_1

Table 2 (continued).

Test	Item	Rep	MF	*E* M(SD)	*EM* M(SD)	*EMR* M (SD)	χ^2 (SD)	G^2 M(SD)	S-χ^2 (SD)	PV-Q_1 (SD)
40	0	0	0.00	0.195(0.084)	0.029(0.006)	0.159(0.030)	11.852(5.101)	12.627(5.682)	29.924(7.885)	9.102(1.845)
40	10%	10%	0.25	0.191(0.083)	0.029(0.006)	0.161(0.030)	12.109(5.212)	12.897(5.823)	30.101(8.047)	9.284(2.208)
40	10%	10%	0.50	0.190(0.082)	0.029(0.006)	0.161(0.030)	12.154(5.225)	12.935(5.829)	30.098(8.096)	9.328(2.261)
40	10%	25%	0.25	0.186(0.081)	0.028(0.007)	0.163(0.029)	12.154(5.195)	12.936(5.794)	30.063(8.025)	9.282(2.057)
40	10%	25%	0.50	0.184(0.080)	0.028(0.007)	0.164(0.030)	12.328(5.361)	13.114(5.979)	30.192(8.074)	9.418(2.263)
40	10%	50%	0.25	0.181(0.078)	0.028(0.007)	0.165(0.028)	12.593(5.401)	13.408(6.030)	30.505(8.038)	9.538(2.234)
40	10%	50%	0.50	0.175(0.077)	0.028(0.007)	0.168(0.028)	12.581(5.429)	13.403(6.050)	30.148(7.901)	9.430(2.147)
40	20%	10%	0.25	0.186(0.082)	0.028(0.007)	0.164(0.034)	12.995(5.523)	13.816(6.125)	30.929(9.709)	10.031(4.358)
40	20%	10%	0.50	0.184(0.081)	0.028(0.007)	0.165(0.034)	13.341(5.666)	14.174(6.241)	31.269(10.419)	10.339(5.126)
40	20%	25%	0.25	0.175(0.080)	0.028(0.007)	0.170(0.031)	13.584(5.792)	14.434(6.432)	31.198(9.104)	10.240(3.523)
40	20%	25%	0.50	0.170(0.079)	0.027(0.007)	0.173(0.035)	14.479(6.360)	15.380(7.120)	31.831(9.937)	11.012(4.643)
40	20%	50%	0.25	0.165(0.077)	0.027(0.008)	0.174(0.028)	15.258(6.825)	16.289(7.719)	32.949(9.292)	11.245(3.882)
40	20%	50%	0.50	0.148(0.072)	0.025(0.008)	0.182(0.030)	15.203(6.970)	16.158(7.716)	32.118(9.106)	11.009(3.751)
40	50%	10%	0.25	0.181(0.082)	0.028(0.008)	0.168(0.040)	15.591(7.316)	16.407(7.315)	33.483(16.933)	12.281(10.897)
40	50%	10%	0.50	0.174(0.082)	0.027(0.007)	0.173(0.049)	16.882(8.586)	17.728(8.274)	34.981(20.783)	13.417(13.916)
40	50%	25%	0.25	0.165(0.081)	0.027(0.009)	0.179(0.040)	17.702(7.155)	18.645(7.741)	34.971(14.259)	13.309(8.000)
40	50%	25%	0.50	0.148(0.080)	0.025(0.008)	0.190(0.055)	20.493(7.965)	21.692(8.793)	37.244(17.146)	15.722(10.715)
40	50%	50%	0.25	0.150(0.078)	0.027(0.010)	0.190(0.028)	21.308(9.034)	22.845(10.406)	38.071(11.795)	14.604(6.514)
40	50%	50%	0.50	0.111(0.071)	0.0210(0.008)	0.208(0.043)	21.487(11.529)	22.796(12.807)	36.773(11.516)	14.631(5.749)

Note: Item = percent of item contamination, except for the 10 item 25% contamination was rounded to 20%, Rep = percent of replicants contaminated, MF = misfit proportion value, *E* = entropy, *EM* = entropy misfit, *EMR* = entropy misfit ratio, χ^2 = Pearson chi-squared, G^2 = likelihood ratio chi-squared, S-χ^2 = Orlando and Thissen (2000) chi-squared, PV-Q_1 = Chalmers and Ng's (2017) adaptation of Q_1

Table 3. Means and standard deviations of all conditions 2PLM

Test	Item	Rep	MF	*E* M(SD)	*EM* M(SD)	*EMR* M (SD)	χ^2 (SD)	G^2 M(SD)	S-χ^2 (SD)	PV-Q_1 (SD)
10	0	0	0.00	0.180(0.106)	0.021(0.008)	0.133(0.042)	46.877(16.297)	55.876(22.586)	6.736(3.620)	8.634(0.874)
10	10%	10%	0.25	0.176(0.106)	0.021(0.008)	0.136(0.044)	47.229(16.177)	56.133(22.235)	6.911(3.831)	8.635(0.881)
10	10%	10%	0.50	0.176(0.105)	0.021(0.008)	0.135(0.043)	47.183(16.231)	56.149(22.459)	6.745(3.654)	8.634(0.872)
10	10%	25%	0.25	0.173(0.106)	0.021(0.008)	0.137(0.043)	47.858(16.550)	56.769(22.632)	6.830(3.706)	8.646(0.878)
10	10%	25%	0.50	0.172(0.102)	0.021(0.008)	0.138(0.044)	47.705(16.600)	56.448(22.669)	6.835(3.718)	8.664(0.850)
10	10%	50%	0.25	0.166(0.098)	0.020(0.009)	0.139(0.042)	48.051(16.607)	57.206(22.826)	6.971(3.734)	8.750(0.892)
10	10%	50%	0.50	0.162(0.097)	0.020(0.009)	0.141(0.043)	48.601(17.197)	57.741(23.530)	6.883(3.736)	8.673(0.861)
10	20%	10%	0.25	0.172(0.106)	0.020(0.008)	0.139(0.049)	47.342(16.492)	56.211(22.717)	6.813(3.749)	8.634(0.865)
10	20%	10%	0.50	0.170(0.106)	0.020(0.008)	0.140(0.049)	47.194(16.143)	55.971(22.291)	6.805(3.703)	8.639(0.867)
10	20%	25%	0.25	0.166(0.107)	0.020(0.009)	0.143(0.050)	48.068(17.216)	56.993(23.812)	6.963(3.790)	8.663(0.865)
10	20%	25%	0.50	0.163(0.106)	0.020(0.009)	0.145(0.053)	47.701(16.853)	56.500(23.343)	6.813(3.729)	8.651(0.861)
10	20%	50%	0.25	0.149(0.091)	0.020(0.010)	0.148(0.045)	51.070(18.681)	60.520(25.488)	7.955(4.335)	8.975(1.053)
10	20%	50%	0.50	0.137(0.094)	0.018(0.009)	0.157(0.052)	50.883(19.318)	60.043(26.509)	7.099(3.849)	8.716(0.836)
10	50%	10%	0.25	0.170(0.108)	0.020(0.009)	0.144(0.060)	47.959(17.289)	57.120(23.897)	6.757(3.652)	8.666(0.870)
10	50%	10%	0.50	0.167(0.111)	0.019(0.009)	0.147(0.067)	48.110(17.210)	57.073(23.643)	6.752(3.636)	8.655(0.866)
10	50%	25%	0.25	0.161(0.108)	0.020(0.010)	0.153(0.068)	49.145(19.268)	58.349(26.746)	7.092(3.873)	8.662(0.865)
10	50%	25%	0.50	0.150(0.112)	0.018(0.009)	0.161(0.079)	48.699(19.907)	58.059(27.607)	6.806(3.764)	8.651(0.856)
10	50%	50%	0.25	0.138(0.089)	0.020(0.011)	0.166(0.057)	56.828(24.514)	67.436(34.213)	10.260(6.290)	9.267(1.255)
10	50%	50%	0.50	0.105(0.101)	0.014(0.009)	0.192(0.085)	54.665(29.851)	65.213(41.076)	7.445(3.937)	8.728(0.830)

Note: Item = percent of item contamination, except for the 10 item 25% contamination was rounded to 20%, Rep = percent of replicants contaminated, MF = misfit proportion value, *E* = entropy, *EM* = entropy misfit, *EMR* = entropy misfit ratio, χ^2 = Pearson chi-squared, G^2 = likelihood ratio chi-squared, S-χ^2 = Orlando and Thissen (2000) chi-squared, PV-Q_1 = Chalmers and Ng's (2017) adaptation of Q_1

Table 3 (continued).

Test	Item	Rep	MF	*E* M(SD)	*EM* M(SD)	*EMR* M (SD)	χ^2 (SD)	G^2 M(SD)	S-χ^2 (SD)	PV-Q_1 (SD)
20	0	0	0.00	0.193(0.104)	0.025(0.007)	0.150(0.044)	18.352(5.326)	20.359(6.540)	14.511(5.394)	8.410(1.039)
20	10%	10%	0.25	0.189(0.103)	0.025(0.007)	0.152(0.044)	18.485(5.408)	20.480(6.648)	14.530(5.433)	8.439(1.035)
20	10%	10%	0.50	0.187(0.102)	0.025(0.007)	0.153(0.044)	18.564(5.483)	20.561(6.730)	14.546(5.420)	8.436(1.032)
20	10%	25%	0.25	0.183(0.100)	0.025(0.007)	0.155(0.044)	18.757(5.483)	20.758(6.705)	14.568(5.452)	8.461(1.045)
20	10%	25%	0.50	0.181(0.100)	0.025(0.007)	0.156(0.044)	18.820(5.553)	20.826(6.803)	14.514(5.430)	8.441(1.022)
20	10%	50%	0.25	0.177(0.094)	0.024(0.007)	0.156(0.042)	19.098(5.606)	21.174(6.876)	14.851(5.575)	8.617(1.170)
20	10%	50%	0.50	0.173(0.095)	0.024(0.008)	0.159(0.043)	19.161(5.671)	21.270(6.957)	14.635(5.468)	8.478(1.008)
20	20%	10%	0.25	0.184(0.105)	0.025(0.007)	0.156(0.048)	18.699(5.669)	20.712(6.977)	14.444(5.450)	8.425(1.038)
20	20%	10%	0.50	0.182(0.104)	0.024(0.007)	0.158(0.050)	18.735(5.672)	20.749(6.979)	14.584(5.541)	8.435(1.050)
20	20%	25%	0.25	0.173(0.100)	0.024(0.008)	0.162(0.047)	19.47(5.899)	21.518(7.238)	14.936(5.605)	8.550(1.098)
20	20%	25%	0.50	0.169(0.101)	0.024(0.008)	0.165(0.052)	19.318(6.122)	21.360(7.545)	14.493(5.496)	8.453(1.011)
20	20%	50%	0.25	0.161(0.088)	0.024(0.008)	0.165(0.042)	21.068(6.532)	23.365(8.045)	16.268(6.135)	9.246(1.723)
20	20%	50%	0.50	0.146(0.090)	0.022(0.008)	0.175(0.048)	20.651(6.579)	22.879(8.047)	14.886(5.535)	8.601(1.041)
20	50%	10%	0.25	0.182(0.105)	0.024(0.008)	0.160(0.058)	18.934(6.072)	21.002(7.481)	14.407(5.426)	8.447(1.041)
20	50%	10%	0.50	0.176(0.107)	0.023(0.008)	0.165(0.065)	18.958(6.120)	21.054(7.576)	14.370(5.417)	8.442(1.025)
20	50%	25%	0.25	0.166(0.103)	0.024(0.009)	0.173(0.063)	20.160(7.158)	22.290(8.856)	15.600(6.135)	8.559(1.087)
20	50%	25%	0.50	0.153(0.110)	0.021(0.008)	0.184(0.079)	20.080(7.679)	22.293(9.509)	14.413(5.514)	8.469(0.988)
20	50%	50%	0.25	0.147(0.085)	0.024(0.010)	0.183(0.047)	25.332(8.964)	28.057(11.073)	19.770(8.382)	10.417(2.700)
20	50%	50%	0.50	0.114(0.101)	0.018(0.008)	0.209(0.078)	23.023(10.319)	25.46(12.674)	15.499(5.948)	8.648(1.004)

Note: Item = percent of item contamination, except for the 10 item 25% contamination was rounded to 20%, Rep = percent of replicants contaminated, MF = misfit proportion value, *E* = entropy, *EM* = entropy misfit, *EMR* = entropy misfit ratio, χ^2 = Pearson chi-squared, G^2 = likelihood ratio chi-squared, S-χ^2 = Orlando and Thissen (2000) chi-squared, PV-Q_1 = Chalmers and Ng's (2017) adaptation of Q_1

Table 3 (continued).

Test	Item	Rep	MF	*E* M(SD)	*EM* M(SD)	*EMR* M (SD)	χ^2 (SD)	G^2 M(SD)	S-χ^2 (SD)	PV-Q_1 (SD)
40	0	0	0.00	0.199(0.101)	0.028(0.006)	0.161(0.045)	10.828(3.944)	11.575(4.353)	28.727(7.651)	8.281(1.340)
40	10%	10%	0.25	0.196(0.100)	0.028(0.006)	0.163(0.045)	10.879(3.967)	11.619(4.365)	28.665(7.690)	8.300(1.359)
40	10%	10%	0.50	0.196(0.100)	0.028(0.006)	0.163(0.046)	10.868(3.973)	11.607(4.378)	28.646(7.630)	8.291(1.347)
40	10%	25%	0.25	0.192(0.098)	0.028(0.007)	0.165(0.044)	11.036(4.047)	11.789(4.453)	28.721(7.741)	8.356(1.378)
40	10%	25%	0.50	0.189(0.097)	0.028(0.007)	0.166(0.045)	10.995(3.966)	11.733(4.367)	28.667(7.627)	8.305(1.323)
40	10%	50%	0.25	0.185(0.091)	0.027(0.007)	0.167(0.043)	11.461(4.234)	12.254(4.679)	29.211(7.752)	8.632(1.603)
40	10%	50%	0.50	0.180(0.092)	0.027(0.007)	0.170(0.043)	11.261(4.045)	12.060(4.480)	28.678(7.583)	8.398(1.348)
40	20%	10%	0.25	0.193(0.101)	0.028(0.007)	0.166(0.049)	10.975(3.986)	11.718(4.391)	28.556(7.644)	8.306(1.354)
40	20%	10%	0.50	0.190(0.101)	0.027(0.007)	0.168(0.049)	10.970(3.973)	11.716(4.389)	28.453(7.639)	8.311(1.357)
40	20%	25%	0.25	0.181(0.098)	0.027(0.007)	0.172(0.048)	11.515(4.218)	12.287(4.652)	29.045(7.806)	8.551(1.559)
40	20%	25%	0.50	0.177(0.100)	0.026(0.007)	0.176(0.052)	11.249(4.084)	11.986(4.512)	28.365(7.711)	8.332(1.318)
40	20%	50%	0.25	0.170(0.085)	0.027(0.008)	0.176(0.042)	13.659(5.297)	14.612(5.916)	31.068(8.445)	9.882(2.709)
40	20%	50%	0.50	0.155(0.089)	0.025(0.007)	0.185(0.047)	12.380(4.443)	13.270(4.963)	29.287(7.877)	8.780(1.571)
40	50%	10%	0.25	0.189(0.104)	0.027(0.007)	0.171(0.057)	11.061(4.051)	11.815(4.484)	28.231(7.709)	8.317(1.340)
40	50%	10%	0.50	0.185(0.107)	0.026(0.007)	0.175(0.064)	11.056(4.065)	11.818(4.513)	28.115(7.621)	8.311(1.338)
40	50%	25%	0.25	0.174(0.102)	0.026(0.008)	0.183(0.061)	11.739(4.281)	12.491(4.728)	29.771(8.444)	8.575(1.561)
40	50%	25%	0.50	0.161(0.111)	0.024(0.009)	0.195(0.077)	11.437(4.254)	12.196(4.761)	27.823(7.650)	8.315(1.277)
40	50%	50%	0.25	0.155(0.082)	0.027(0.009)	0.193(0.043)	18.298(7.453)	19.673(8.540)	35.346(10.456)	12.482(4.595)
40	50%	50%	0.50	0.122(0.101)	0.020(0.009)	0.219(0.073)	13.224(5.006)	14.041(5.595)	29.743(8.228)	8.830(1.481)

Note: Item = percent of item contamination, except for the 10 item 25% contamination was rounded to 20%, Rep = percent of replicants contaminated, MF = misfit proportion value, *E* = entropy, *EM* = entropy misfit, *EMR* = entropy misfit ratio, χ^2 = Pearson chi-squared, G^2 = likelihood ratio chi-squared, S-χ^2 = Orlando and Thissen (2000) chi-squared, PV-Q_1 = Chalmers and Ng's (2017) adaptation of Q_1

Table 4. Proportions of partial variance accounted for by main and interaction effects (partial eta-squared)

Source	*EMR* 1PLM	2PLM	*EM* 1PLM	2PLM	*E* 1PLM	2PLM	χ^2 1PLM	2PLM	G^2 1PLM	2PLM	S-χ^2 1PLM	2PLM	PV-Q_1 1PLM	2PLM
Item Contamination (IC)	**0.16**	**0.16**	0.00	0.02	0.01	**0.05**	**0.10**	**0.05**	**0.11**	**0.05**	0.00	0.01	0.00	**0.03**
Test Length (TL)	**0.09**	0.02	**0.07**	**0.05**	0.00	0.00	**0.47**	**0.53**	**0.48**	**0.49**	**0.42**	**0.61**	0.01	0.00
Person Contamination (PC)	0.00	0.01	0.01	0.00	0.00	0.00	**0.04**	**0.04**	**0.04**	**0.04**	**0.03**	0.00	**0.04**	0.01
Type of Misfit (TM)	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
IC*TL	0.00	0.00	0.00	0.00	0.00	0.00	**0.06**	**0.03**	**0.06**	**0.03**	0.00	0.00	0.00	0.01
IC*PC	**0.03**	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
TL*PC	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.02	0.01	0.00	0.00	0.00	0.00
IC*TM	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01
TL*TM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PC*TM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IC*TL*PC	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
IC*TL*TM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
IC*PC*TM	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01
TL*PC*TM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IC*TL*PC*TM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01

Note: *E* = entropy, *EM* = entropy misfit, *EMR* = entropy misfit ratio, χ^2 = Pearson chi-squared, G^2 = likelihood ratio chi-squared, S-χ^2 = Orlando and Thissen's (2000) chi-squared, PV-Q_1 = Chalmers and Ng's (2017) adaptation of Q_1; partial eta-squared values of .03 or greater are bolded

Empirical Power and Type I Error for χ^2 and G^2 Statistics

Empirical power was calculated as the proportion of items that were correctly identified as misfit divided by the proportion of items that were generated to misfit in each condition. Proportion of items that were correctly identified as misfit are presented, which can be thought of as an unstandardized measure of empirical power (i.e., referred to as unstandardized in the tables). Most methodologists desire power values of .70 or higher. Type I error rates were calculated as the proportion of items that were flagged as misfit when the population-generating item-fit. Type I error values of .05 are desirable for the 95% cut points.

Tables 5 and 6 contain empirical power and Type I error rates for the χ^2 and G^2 statistics, respectively. For both the χ^2 and G^2 statistics, test length was the major variable resulting in variation of empirical power and Type I error rates. As the number of items on the test increased, both empirical power and Type I error decreased. Empirical power was very high for all of the 10-item test conditions and very low for the 20- and 40-item test conditions. Type I error rates were very high in all conditions. Similar results were found for the 1PLM and 2PLM models, although Type I error rates were a little lower (albeit still largely inflated) for the 2PLM.

Table 5. 1 and 2PLM χ^2 power to detect and Type I error

				% contaminated persons					
				Type I error			Power		
Model	Test length	% items	Contam. type	10%	25%	50%	10%	25%	50%
1PLM	10 items	10%	Misfit .25	99.99	99.99	99.99	99.30	93.90	77.80
1PLM	10 items	10%	Misfit .50	99.98	99.97	99.99	99.30	95.60	70.30
1PLM	10 items	20%	Misfit .25	99.96	100.00	100.00	99.75	98.20	91.95
1PLM	10 items	20%	Misfit .50	100.00	100.00	100.00	99.70	97.90	83.00
1PLM	10 items	50%	Misfit .25	99.96	99.96	99.98	99.98	99.92	99.74
1PLM	10 items	50%	Misfit .50	99.98	100.00	100.00	99.94	99.92	99.78
1PLM	20 items	10%	Misfit .25	76.03	81.17	88.56	39.70	19.15	35.55
1PLM	20 items	10%	Misfit .50	75.84	82.40	90.33	37.35	18.15	46.80
1PLM	20 items	25%	Misfit .25	78.16	86.67	95.93	58.70	45.70	40.98
1PLM	20 items	25%	Misfit .50	81.14	91.83	98.77	48.08	26.26	20.38
1PLM	20 items	50%	Misfit .25	71.95	78.39	93.93	75.81	79.77	78.89
1PLM	20 items	50%	Misfit .50	84.88	96.04	99.87	63.52	56.93	48.09
1PLM	40 items	10%	Misfit .25	24.52	30.02	39.21	11.48	20.75	78.40
1PLM	40 items	10%	Misfit .50	25.00	31.71	43.79	9.85	27.63	92.08
1PLM	40 items	25%	Misfit .25	26.06	36.98	60.06	15.68	20.88	51.33
1PLM	40 items	25%	Misfit .50	29.35	46.79	76.82	11.40	17.05	55.43
1PLM	40 items	50%	Misfit .25	23.14	36.79	75.87	29.49	49.17	66.82
1PLM	40 items	50%	Misfit .50	34.45	63.65	95.05	18.34	22.00	29.97
2PLM	10 items	10%	Misfit .25	99.83	99.83	99.96	99.60	98.30	88.30
2PLM	10 items	10%	Misfit .50	99.86	99.88	99.92	99.40	97.20	86.10
2PLM	10 items	20%	Misfit .25	99.83	99.88	99.93	99.55	98.45	88.45
2PLM	10 items	20%	Misfit .50	99.88	99.90	99.95	99.80	98.00	82.20
2PLM	10 items	50%	Misfit .25	99.80	99.66	99.30	99.88	99.62	98.60
2PLM	10 items	50%	Misfit .50	99.88	99.98	100.00	99.72	99.10	92.68
2PLM	20 items	10%	Misfit .25	70.02	72.82	76.78	56.30	37.50	17.70
2PLM	20 items	10%	Misfit .50	70.84	73.23	76.41	53.60	36.40	15.15
2PLM	20 items	25%	Misfit .25	71.61	78.65	86.83	65.66	58.04	32.48
2PLM	20 items	25%	Misfit .50	73.39	80.13	88.10	61.28	45.92	19.48
2PLM	20 items	50%	Misfit .25	70.51	77.92	90.25	74.17	81.42	84.07
2PLM	20 items	50%	Misfit .50	76.49	87.09	97.44	67.66	67.96	49.06
2PLM	40 items	10%	Misfit .25	12.27	13.06	13.94	12.05	11.25	8.53
2PLM	40 items	10%	Misfit .50	12.42	12.93	14.19	12.20	11.75	8.00
2PLM	40 items	25%	Misfit .25	12.84	15.32	18.01	14.35	19.01	15.33
2PLM	40 items	25%	Misfit .50	13.11	15.44	18.89	12.62	11.51	7.86
2PLM	40 items	50%	Misfit .25	15.94	30.24	60.28	16.13	33.19	59.78
2PLM	40 items	50%	Misfit .50	14.26	21.09	39.93	14.51	22.22	19.77

Table 6. 1 and 2PLM G^2 power to detect and Type I error

				% contaminated persons					
				Type I error			Power		
Model	Test length	% items	Contam. type	10%	25%	50%	10%	25%	50%
1PLM	10 items	10%	Misfit .25	99.99	100.00	100.00	99.60	94.80	79.90
1PLM	10 items	10%	Misfit .50	99.99	99.98	100.00	99.60	96.80	71.40
1PLM	10 items	20%	Misfit .25	99.98	100.00	100.00	99.85	98.45	93.00
1PLM	10 items	20%	Misfit .50	100.00	100.00	100.00	99.75	98.70	85.30
1PLM	10 items	50%	Misfit .25	99.98	99.98	99.98	100.00	99.92	99.76
1PLM	10 items	50%	Misfit .50	99.98	100.00	100.00	99.98	99.96	99.84
1PLM	20 items	10%	Misfit .25	80.82	85.28	91.23	45.05	21.15	32.55
1PLM	20 items	10%	Misfit .50	80.82	86.26	92.72	43.05	18.55	43.15
1PLM	20 items	25%	Misfit .25	82.45	89.87	97.01	64.52	50.78	43.04
1PLM	20 items	25%	Misfit .50	85.09	93.77	99.09	54.22	29.16	20.40
1PLM	20 items	50%	Misfit .25	77.22	82.17	95.47	80.20	82.55	80.49
1PLM	20 items	50%	Misfit .50	88.27	97.02	99.88	69.71	63.08	53.11
1PLM	40 items	10%	Misfit .25	30.39	36.24	46.11	13.03	18.50	74.48
1PLM	40 items	10%	Misfit .50	30.89	38.32	51.02	10.75	23.93	90.05
1PLM	40 items	25%	Misfit .25	32.07	43.47	65.99	19.28	23.03	49.12
1PLM	40 items	25%	Misfit .50	35.76	53.36	81.10	13.38	16.09	52.70
1PLM	40 items	50%	Misfit .25	28.34	42.37	79.63	34.86	53.03	68.03
1PLM	40 items	50%	Misfit .50	41.03	68.81	96.17	22.74	25.36	31.63
2PLM	10 items	10%	Misfit .25	99.87	99.90	99.98	99.60	98.40	88.90
2PLM	10 items	10%	Misfit .50	99.89	99.90	99.93	99.50	97.40	86.90
2PLM	10 items	20%	Misfit .25	99.85	99.93	99.95	99.70	98.65	89.75
2PLM	10 items	20%	Misfit .50	99.90	99.91	99.98	99.85	98.20	83.70
2PLM	10 items	50%	Misfit .25	99.82	99.72	99.40	99.94	99.64	98.80
2PLM	10 items	50%	Misfit .50	99.90	100.00	100.00	99.80	99.16	93.68
2PLM	20 items	10%	Misfit .25	77.39	79.75	82.93	63.10	43.85	19.60
2PLM	20 items	10%	Misfit .50	78.20	80.22	83.05	61.40	42.30	17.00
2PLM	20 items	25%	Misfit .25	79.02	84.20	90.30	73.28	64.92	35.50
2PLM	20 items	25%	Misfit .50	80.30	85.26	91.71	69.04	52.44	22.14
2PLM	20 items	50%	Misfit .25	77.36	82.94	92.44	80.46	85.23	85.72
2PLM	20 items	50%	Misfit .50	82.75	90.33	97.99	75.67	74.65	53.91
2PLM	40 items	10%	Misfit .25	17.54	18.51	19.81	15.53	13.10	9.28
2PLM	40 items	10%	Misfit .50	17.56	18.61	20.08	15.03	13.43	8.50
2PLM	40 items	25%	Misfit .25	18.10	21.30	25.51	18.81	23.56	17.07
2PLM	40 items	25%	Misfit .50	18.54	21.85	26.45	16.59	13.83	8.58
2PLM	40 items	50%	Misfit .25	21.49	36.95	66.13	21.72	39.36	63.40
2PLM	40 items	50%	Misfit .50	19.85	27.85	49.54	20.51	28.72	22.41

Empirical Power for EMR_j and S-X^2 and PV-Q_1

Tables 7, 8, and 9 contain the empirical power and Type I error rates for S-X^2, PV-Q_1, and EMR_j, respectively. For ease of interpretation we present empirical power as percentages rather than proportions. Figure 2 graphically displays the empirical statistical power for S-X^2, PV-Q_1, and EMR_j, for the 1PLM (top panel) and 2PLM (bottom panel), and corresponds with values in Tables 7, 8, and 9. In general, for the 1PLM the pattern of empirical power estimates was similar between S-X^2, PV-Q_1, and EMR_j within a condition, however, values for S-X^2 were better than PV-Q_1, and values for EMR_j were better than both S-X^2 and PV-Q_1. In general, S-X^2 and PV-Q_1 did not yield adequate power for any conditions in the 2PLM, whereas EMR_j values were noticeably higher than both S-X^2 and PV-Q_1 values in most conditions.

For all three measures, empirical power for the 1PLM was highest for all conditions in which 50% of persons were misfit, and lowest when only 10% of persons misfit. In general, empirical power increased as the percentage of contaminated persons increased, and as the type of contamination increased from 25% to 50%. Interestingly, as the percentage of contaminated items increased, empirical power decreased, but this finding should be considered in conjunction with changes in Type I error (discussed in the next section). Similar trends were observed for the 2PLM as for the 1PLM for *EMR*, with empirical power being lower for similar conditions in the 2PLM.

Type I Error for EMR_j

Figure 3 graphically displays Type I error rates for S-X^2, PV-Q_1, and EMR_j for the 1PLM (top panel) and the 2PLM (bottom panel). Type I error for PV-Q_1 was extremely low in nearly all conditions. In general, Type I error values were near the desirable values of .05 in many conditions and similar between S-X^2 and EMR_j across conditions. For the EMR_j and S-X^2 measures, Type I error was higher than anticipated in the many of the conditions in which 50% of the items and 25-50% of persons were contaminated. For EMR_j in most conditions, Type I error for the 2PLM was underestimated as the percent of contaminated items and persons increased. This finding is unsurprising given that the empirical power was greater for the 1PLM than for the 2PLM. The exception was for the conditions in which 50% items and 25-50% persons were contaminated, particularly when type of misfit was 25%; in these conditions Type I error was overestimated for some conditions. This was particularly true for the conditions in which 50% of persons and 50% of items were contaminated and subtests were only 10-20 items.

Table 7. 1 and 2PLM S-X^2 Power to detect and Type I error

				% contaminated persons					
				Type I error			Power		
Model	Test length	% items	Contam. type	10%	25%	50%	10%	25%	50%
1PLM	10 items	10%	Misfit .25	7.83	7.92	7.56	23.90	65.30	98.60
1PLM	10 items	10%	Misfit .50	7.90	8.02	8.14	23.90	71.50	99.50
1PLM	10 items	20%	Misfit .25	7.68	8.55	9.63	15.05	40.35	84.80
1PLM	10 items	20%	Misfit .50	7.74	8.23	11.78	17.25	52.00	95.15
1PLM	10 items	50%	Misfit .25	9.60	13.02	22.76	9.96	19.16	36.16
1PLM	10 items	50%	Misfit .50	8.52	10.58	23.02	10.68	18.80	34.94
1PLM	20 items	10%	Misfit .25	6.84	6.34	6.28	20.25	55.90	97.15
1PLM	20 items	10%	Misfit .50	6.66	6.10	6.21	20.40	68.15	99.65
1PLM	20 items	25%	Misfit .25	6.97	7.07	9.13	11.36	28.80	74.28
1PLM	20 items	25%	Misfit .50	6.55	6.55	10.31	14.86	41.60	89.46
1PLM	20 items	50%	Misfit .25	7.96	11.40	24.36	9.71	19.53	42.95
1PLM	20 items	50%	Misfit .50	6.85	8.41	23.45	9.76	19.72	39.88
1PLM	40 items	10%	Misfit .25	6.09	5.73	5.24	17.45	52.28	96.23
1PLM	40 items	10%	Misfit .50	5.99	5.53	4.82	20.00	64.70	99.53
1PLM	40 items	25%	Misfit .25	6.08	5.97	6.69	11.18	26.70	73.64
1PLM	40 items	25%	Misfit .50	5.82	5.38	6.92	14.19	38.90	88.33
1PLM	40 items	50%	Misfit .25	7.18	10.51	23.41	8.73	19.61	44.12
1PLM	40 items	50%	Misfit .50	5.55	7.12	19.65	9.07	19.76	42.78
2PLM	10 items	10%	Misfit .25	5.57	5.43	5.02	7.40	6.30	5.30
2PLM	10 items	10%	Misfit .50	4.67	5.03	4.64	5.20	5.20	4.80
2PLM	10 items	20%	Misfit .25	5.31	5.54	6.45	4.75	6.95	7.65
2PLM	10 items	20%	Misfit .50	5.11	5.13	5.68	4.85	6.05	5.40
2PLM	10 items	50%	Misfit .25	5.02	8.10	21.18	5.74	11.46	23.66
2PLM	10 items	50%	Misfit .50	6.00	6.42	8.46	4.62	5.40	6.52
2PLM	20 items	10%	Misfit .25	5.03	5.11	5.39	6.40	5.15	5.75
2PLM	20 items	10%	Misfit .50	4.91	5.38	5.22	5.70	6.00	5.40
2PLM	20 items	25%	Misfit .25	5.22	5.81	8.82	4.80	7.32	12.20
2PLM	20 items	25%	Misfit .50	5.01	5.47	6.21	5.02	5.12	4.86
2PLM	20 items	50%	Misfit .25	5.25	7.96	23.69	6.61	12.13	27.52
2PLM	20 items	50%	Misfit .50	5.68	6.29	11.94	4.88	5.54	7.59
2PLM	40 items	10%	Misfit .25	5.03	4.90	5.14	5.10	5.18	5.68
2PLM	40 items	10%	Misfit .50	4.83	4.85	4.91	4.60	5.43	5.03
2PLM	40 items	25%	Misfit .25	5.33	5.63	8.89	5.18	7.04	11.23
2PLM	40 items	25%	Misfit .50	5.03	5.61	6.28	4.55	4.69	4.39
2PLM	40 items	50%	Misfit .25	5.44	8.49	26.34	6.28	12.41	27.84
2PLM	40 items	50%	Misfit .50	5.30	8.03	15.46	4.41	5.76	7.72

Table 8. 1 and 2PLM PV-Q_1 Power to detect and Type I error

				% contaminated persons					
				Type I error			Power		
Model	Test length	% items	Contam. type	10%	25%	50%	10%	25%	50%
1PLM	10 items	10%	Misfit .25	0.02	0.00	0.00	2.90	28.00	90.20
1PLM	10 items	10%	Misfit .50	0.00	0.01	0.00	1.30	29.50	95.00
1PLM	10 items	20%	Misfit .25	0.00	0.00	0.00	1.00	8.20	47.50
1PLM	10 items	20%	Misfit .50	0.00	0.00	0.01	1.05	13.65	69.50
1PLM	10 items	50%	Misfit .25	0.10	0.10	0.10	0.18	0.82	2.26
1PLM	10 items	50%	Misfit .50	0.00	0.00	0.02	0.16	1.04	2.68
1PLM	20 items	10%	Misfit .25	0.11	0.07	0.06	6.55	40.55	96.70
1PLM	20 items	10%	Misfit .50	0.13	0.04	0.04	5.95	58.25	99.65
1PLM	20 items	25%	Misfit .25	0.10	0.05	0.15	1.22	10.96	56.50
1PLM	20 items	25%	Misfit .50	0.06	0.05	0.53	2.84	24.56	83.54
1PLM	20 items	50%	Misfit .25	0.16	0.36	0.32	0.79	4.91	18.31
1PLM	20 items	50%	Misfit .50	0.07	0.10	1.93	0.99	5.52	16.65
1PLM	40 items	10%	Misfit .25	0.32	0.19	0.20	8.58	54.13	98.65
1PLM	40 items	10%	Misfit .50	0.28	0.17	0.27	10.88	70.30	99.98
1PLM	40 items	25%	Misfit .25	0.25	0.24	0.77	2.96	20.39	77.09
1PLM	40 items	25%	Misfit .50	0.16	0.42	2.58	4.73	37.10	94.20
1PLM	40 items	50%	Misfit .25	0.54	1.25	3.02	1.79	14.14	50.30
1PLM	40 items	50%	Misfit .50	0.20	0.82	10.69	1.83	12.65	37.45
2PLM	10 items	10%	Misfit .25	0.00	0.00	0.00	0.00	0.00	0.00
2PLM	10 items	10%	Misfit .50	0.00	0.00	0.00	0.00	0.00	0.00
2PLM	10 items	20%	Misfit .25	0.00	0.00	0.00	0.00	0.00	0.00
2PLM	10 items	20%	Misfit .50	0.00	0.00	0.00	0.00	0.00	0.00
2PLM	10 items	50%	Misfit .25	0.00	0.00	0.00	0.00	0.02	0.16
2PLM	10 items	50%	Misfit .50	0.00	0.00	0.00	0.00	0.00	0.00
2PLM	20 items	10%	Misfit .25	0.00	0.00	0.01	0.00	0.15	0.10
2PLM	20 items	10%	Misfit .50	0.00	0.01	0.01	0.05	0.15	0.00
2PLM	20 items	25%	Misfit .25	0.01	0.01	.	0.04	0.10	0.24
2PLM	20 items	25%	Misfit .50	0.01	0.01	0.01	0.02	0.02	0.00
2PLM	20 items	50%	Misfit .25	0.00	0.02	0.08	0.11	1.63	11.43
2PLM	20 items	50%	Misfit .50	0.00	0.00	0.00	0.00	0.00	0.02
2PLM	40 items	10%	Misfit .25	0.09	0.07	0.07	0.38	0.50	0.43
2PLM	40 items	10%	Misfit .50	0.07	0.08	0.06	0.23	0.58	0.35
2PLM	40 items	25%	Misfit .25	0.11	0.11	0.10	0.22	1.23	1.41
2PLM	40 items	25%	Misfit .50	0.07	0.06	0.05	0.10	0.18	0.15
2PLM	40 items	50%	Misfit .25	0.13	0.49	3.99	0.67	8.24	36.56
2PLM	40 items	50%	Misfit .50	0.14	0.14	0.04	0.07	0.56	0.46

Table 9. 1 and 2PLM *EMR*: empirical power and Type I error

				% contaminated persons					
				Type I error			Power		
Model	Test length	% items	Contam. type	10%	25%	50%	10%	25%	50%
1PLM	10 items	10%	Misfit .25	5.22	5.93	6.39	32.50	71.50	88.40
1PLM	10 items	10%	Misfit .50	5.57	6.06	6.80	29.00	88.00	100.00
1PLM	10 items	20%	Misfit .25	6.34	7.56	9.04	22.20	63.10	90.00
1PLM	10 items	20%	Misfit .50	6.01	7.15	9.73	25.55	81.35	100.00
1PLM	10 items	50%	Misfit .25	11.32	22.84	46.74	8.62	23.92	75.50
1PLM	10 items	50%	Misfit .50	7.80	16.00	50.66	18.92	67.50	99.46
1PLM	20 items	10%	Misfit .25	5.14	4.72	5.14	26.45	63.25	80.65
1PLM	20 items	10%	Misfit .50	4.82	5.02	4.91	25.75	78.05	100.00
1PLM	20 items	25%	Misfit .25	5.79	6.57	7.91	13.76	43.60	77.86
1PLM	20 items	25%	Misfit .50	4.94	5.50	6.13	21.02	67.22	100.00
1PLM	20 items	50%	Misfit .25	9.12	15.86	26.38	7.02	15.11	56.91
1PLM	20 items	50%	Misfit .50	5.97	7.82	16.17	14.19	50.28	99.17
1PLM	40 items	10%	Misfit .25	5.05	4.81	4.67	24.98	63.68	76.73
1PLM	40 items	10%	Misfit .50	5.01	4.80	4.57	25.58	74.38	100.00
1PLM	40 items	25%	Misfit .25	5.29	6.27	6.78	13.03	38.44	72.92
1PLM	40 items	25%	Misfit .50	5.07	4.61	5.07	19.97	64.58	99.96
1PLM	40 items	50%	Misfit .25	8.19	14.38	21.42	6.83	14.38	49.79
1PLM	40 items	50%	Misfit .50	5.56	6.53	8.84	12.85	43.34	98.78
2PLM	10 items	10%	Misfit .25	4.87	4.81	4.39	13.10	47.10	84.20
2PLM	10 items	10%	Misfit .50	4.60	4.92	4.52	14.70	50.70	99.80
2PLM	10 items	20%	Misfit .25	4.89	4.44	3.90	11.50	36.15	80.55
2PLM	10 items	20%	Misfit .50	4.41	4.46	4.03	13.55	48.15	98.95
2PLM	10 items	50%	Misfit .25	7.10	11.50	16.80	5.40	6.92	27.28
2PLM	10 items	50%	Misfit .50	4.98	4.54	3.28	8.86	27.88	87.38
2PLM	20 items	10%	Misfit .25	4.91	5.11	4.39	12.10	41.35	74.60
2PLM	20 items	10%	Misfit .50	4.86	4.97	4.50	14.30	50.80	99.75
2PLM	20 items	25%	Misfit .25	5.05	4.91	4.59	8.42	23.78	67.40
2PLM	20 items	25%	Misfit .50	4.86	4.67	3.76	11.18	38.58	98.98
2PLM	20 items	50%	Misfit .25	6.38	9.89	15.09	5.06	6.34	18.79
2PLM	20 items	50%	Misfit .50	4.96	4.67	3.68	8.81	24.91	90.48
2PLM	40 items	10%	Misfit .25	5.14	4.98	4.73	11.05	40.20	71.58
2PLM	40 items	10%	Misfit .50	5.13	4.81	4.61	13.68	46.08	99.75
2PLM	40 items	25%	Misfit .25	4.95	5.37	4.78	7.64	21.45	62.03
2PLM	40 items	25%	Misfit .50	4.93	4.63	4.54	10.84	36.58	99.24
2PLM	40 items	50%	Misfit .25	6.38	8.89	11.70	5.33	6.29	16.93
2PLM	40 items	50%	Misfit .50	5.07	4.95	4.15	8.32	21.29	91.36

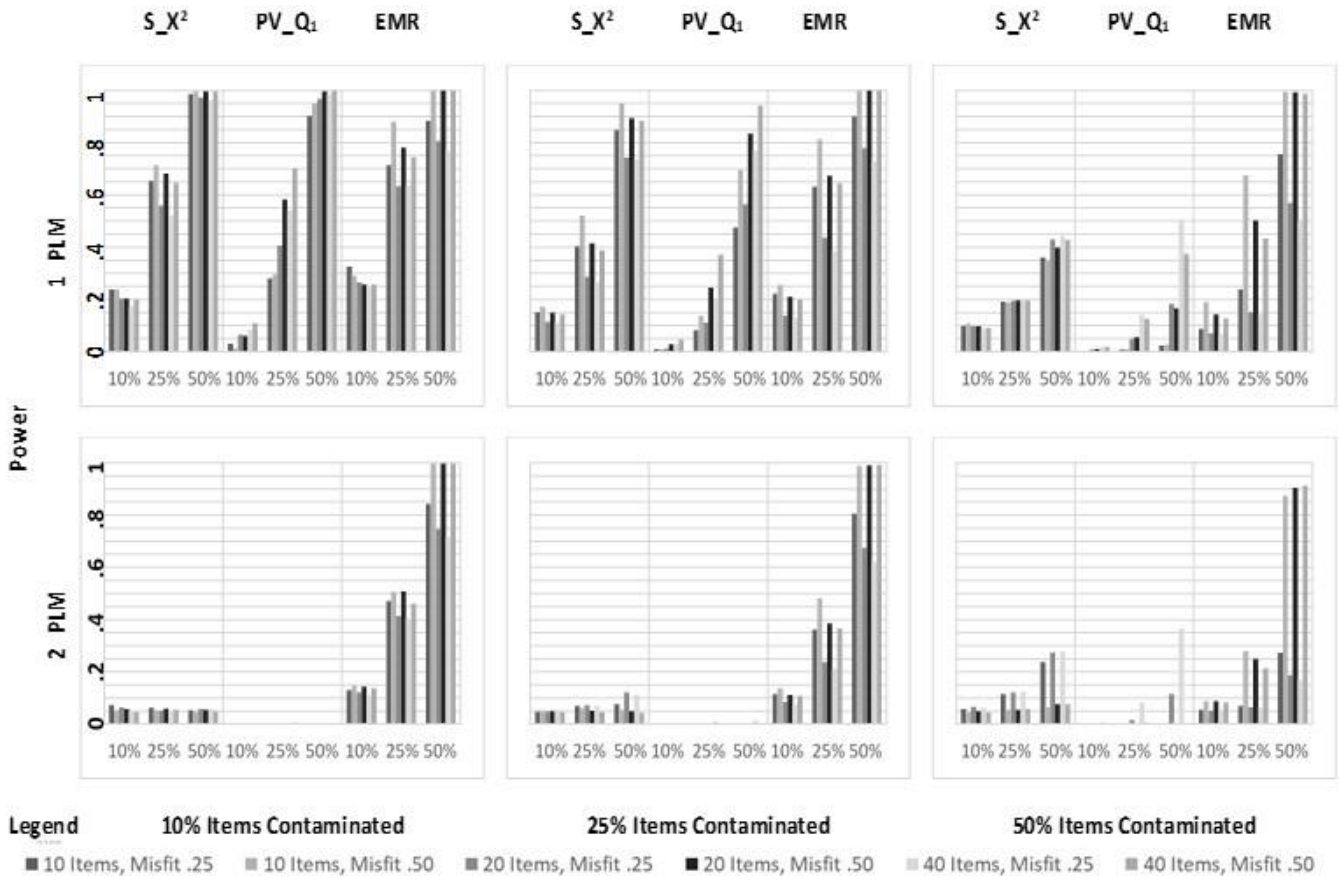

Figure 2. Power for S-X^2, PV-Q_1, and *EMR*; note: within each panel there are nine sets of six bars; the legend represents what each bar represents; the *x*-axis in each panel represents person contamination

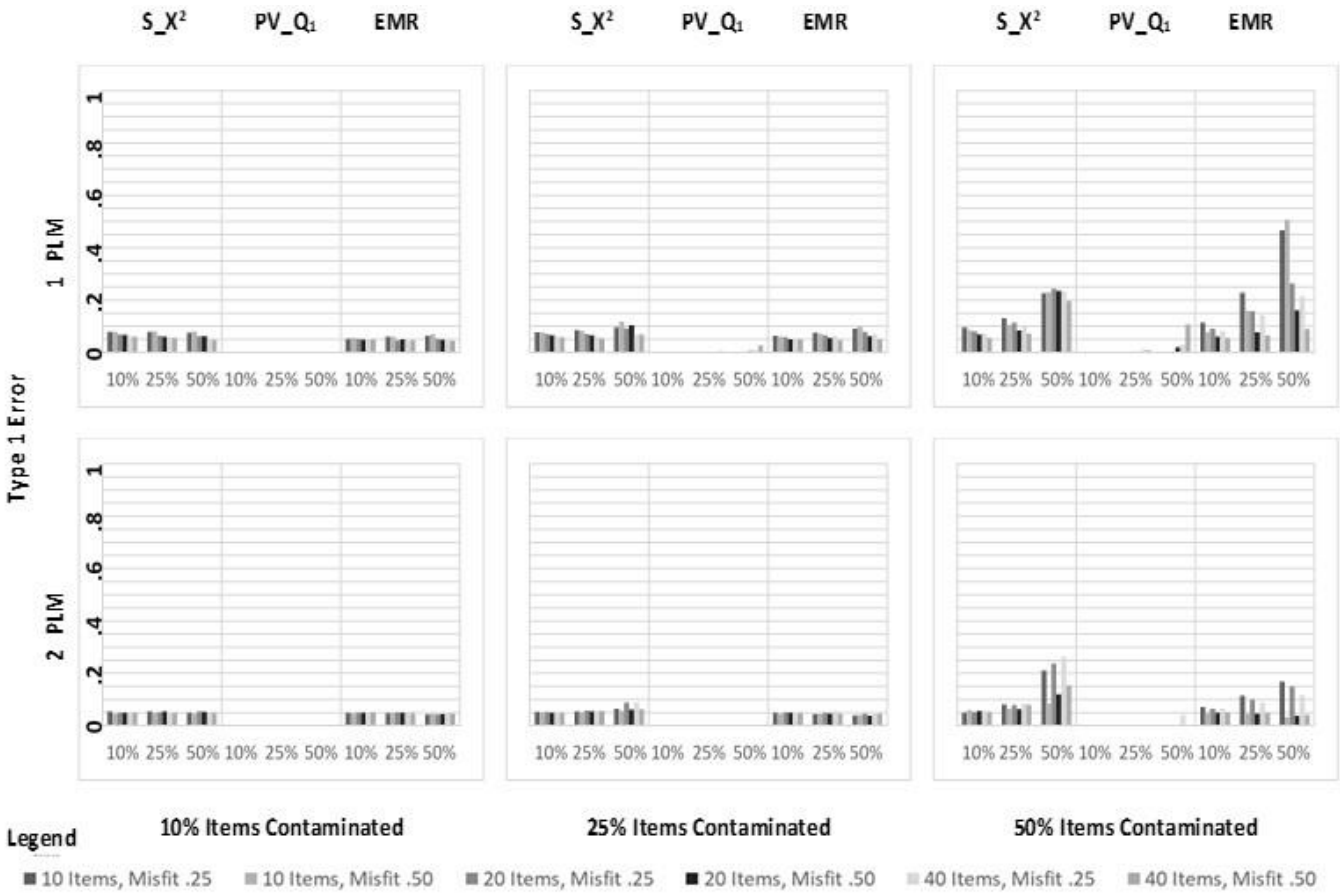

Figure 3. Type 1 error for S-X^2, PV-Q_1, and *EMR*; note: within each panel there are nine sets of six bars; the legend represents what each bar represents; the *x*-axis in each panel represents person contamination

Discussion

Three measures of entropy (E_j, EM_j, and EMR_j) were introduced as measures of item-fit for IRT. EM_j incorporates misfit for an item, but that misfit is not relative to the total amount of departure of a predicted probability from a threshold. EMR_j incorporates both pieces of information in a ratio. Two simulation studies were conducted to investigate the distribution, empirical power, and Type I error rate of these entropy measures, and compared these statistics with traditionally used measures of χ^2, G^2, S-X^2, and PV-Q_1. Data were simulated to mirror issues with unidimensional misfit. However, the nature of many simulations, as was the case here, is to generate a mixture or multivariate data and estimate using a unidimensional model to determine impact.

In the first simulation study, the thresholds (Table 1) were useful to make decisions regarding type I error. Baseline models were examined in which all participants were generated to fit the appropriate IRT model and cut points assumed model-data fit. In applied settings researchers may find it useful to think of *EMRj* values as a type of item-level outlier detection. That is, examining the most egregious entropy values may help researchers determine whether or not an item fits the data. For example, consider a 1 PLM with twenty-five items and 1200 participants, and *EMRj* values ranging from .10 to .24, Mean = .15(.03), with the three largest values being .240, .184, and .176. Recall, larger values of *EMRj* are reflective of more misfit and are thus less desirable. Using Table 1, one would interpolate a threshold value to be between .185 and .203. Instead of using a threshold value and attempting to interpolate (perhaps coming up with a value marginally over .185), or running a mini-simulation using 25 items and 1200 participants, it would be useful to consider the most extreme values and use entropy measures as one element of the investigation of the item(s). In this example, a researcher might either select a predetermined number of extreme items to investigate, say 2 or 3 rounding up for values of 5 and 10 percent of the number of items, or alternatively examine items that are more clearly outliers. In the first case we could select a rule for the number of items to investigate, perhaps the largest 3 values, or we may flag an item with *EMRj* value of .240 as a potentially misfitting outlier. The value added by using *EMRj* is that low misfit values have clearer separation between than those with a larger misfit ratio. More specifically, persons predicted to have an *EMRj* value of 1 or 0 are likely to be classified correctly whereas misclassified persons are typically near the threshold for classification (e.g. threshold of .185 for classification correct and predicted value is .190). Considering

significance testing has fallen out of favor amongst many in the statistical and measurement fields (Trafimow & Marks, 2015; Wasserstein et al., 2019), we might suggest a similar form of outlier detection for other measures of the item-fit.

In the second simulation study empirical power and empirical pseudo-Type I error rate were compared for S-X^2, PV-Q_1, and EMR_j across several types of contamination (2 model types, 3 test lengths, 3 percentages of items contaminated, 3 percentages of persons contaminated, and 2 types of contamination). Previous research presents conflicting recommendations for the use of χ^2 and G^2 as measures of item-fit. Thissen and Steinberg (1997) state χ^2 and G^2 should only be used when there are fewer than five or six items whereas Cochran (1952) found χ^2 and G^2 yield incorrect *p*-values with short test lengths. Other researchers suggested χ^2 and G^2 are only useful for small test lengths (i.e., 10-20 items; Chon et al., 2010; Maydeu-Olivares et al., 2011; Zimowski et al., 2003). In the current study, χ^2 and G^2 resulted in extreme empirical power values (high for small number of items; low for large number of items) and highly inflated Type I error rates, especially in the 10-item conditions. Based on findings in the current study, we do not recommend χ^2 and G^2 for use for 10-item test lengths.

Although EMR_j may be useful for tests with a smaller number of items, it is important to note values were not completely invariant across test length. This is likely because longer tests have slightly more accurate parameter estimates, thus predicted response probabilities also become more accurate, therefore test length has an indirect impact on the thresholds through IRT model parameter estimates. While entropy may be useful for a small number of items, researchers should note that the accuracy of parameter estimates improve as test length increases. This finding is similar to trends observed with sample size and is expected with all fit statistics and all models.

For the 2PLM, S-X^2 and PV-Q_1 had low empirical power in all conditions. Orlando and Thissen (2000) had similar low empirical power in estimating the 2PLM on 3PLM generated data. PV-Q_1 had lower than desired Type I error in most conditions consistent with previous literature (Chalmers & Ng, 2017). In contrast, EMR_j resulted in adequate empirical power rates in many of the conditions, although most notably when 50% of people are contaminated, and Type I error rates across most conditions were accurate, and in some conditions performed well across all test lengths. When extreme amount of contamination was present (i.e., 50% of items) high rejection rates were found for EMR for the uncontaminated items. In these conditions, EMR_j had more statistical power than S-X^2 and G^2. Thus, this finding brings us to two conclusions. First, perhaps a more conservative cut-point should be selected to find a better balance between Type I error and statistical

power. Second, in additional to item-level fit, these measures can help determine if questionable model-level fit is present. More specifically, researchers should be cautioned when many warnings of misfitting items are present, they should revisit model-level fit and not just assume they have rogue items.

Future Research and Limitations

The focus was on the dichotomous, unidimensional, correct/incorrect 1PLM and 2PLM models. A next step is to investigate EMR_j for other types of models (e.g., polytomous, multidimensional, ordered, or partial-credit). Furthermore, sample size was not varied, although it was quite large (108+6 conditions) given the manipulated factors. Varying sample size to even 2 values would have resulted in a fully-crossed design with twice the number of conditions. Given that sample size was not varied, it is not recommended to use the thresholds derived in study 1 as a hard cut-point. Given the bias analysis of the study (i.e., bias was < .006 for item difficulty and discrimination, on average), larger sample sizes are not likely to be impacted. Similarly, these results cannot be readily generalized to data conditions not explored in the current study (e.g., theta distributions and item parameter distributions).

The χ^2, G^2, S-X^2, and PV-Q_1 statistics were used as comparison measures for reasons described earlier in this paper. The χ^2 and G^2 measures rely on binning ability estimates that are based on model dependent arbitrary cut-points, whereas Orlando and Thissen's (2000) S-X^2 and S-G^2 statistics use observed test scores (summated scores) instead of theta ability estimates creating a theoretical dissonance in that a researcher believes in a latent variable ability estimate and falls back on a classical test theory observed-score ability estimate. Future research may aim to compare EMR_j to other measures of item-fit in IRT such as Stone's (2000) χ^2* and G^2* statistics, and the posterior predictive model checking (Swaminathan et al., 2007). Although previous research found Stone's χ^2* and G^2* to have inflated Type I error rates associated with these pseudo-observed scores and noted that it took 40 minutes to calculate these measures for a single dataset for 40-items and 4,000 persons (Chon et al., 2010). OUTFIT and INFIT are also residual-based measures sometimes used for item-fit. These measures follow an approximate χ^2 distribution and only assign one individual per bin, but each individual/bin is weighted differently. The goal of the current study was to introduce the entropy variants as measures of item-fit in IRT; the goal was not to compare all measures of item-fit. Similarly, it may be of interest to focus on identifying a one-size-fits-all threshold for EMR_j.

Similarly, it may be beneficial to consider test length. Although a test length of 10 seems extreme for the current practitioner, advanced technology, new psychometric methods, shorter tests, embedded in systems (such as schools) which act as both summative and formative assessment (Dardick & Choi, 2016) are on our horizon and it is valuable to include such test lengths in current simulation studies.

An advantage to calculating entropy-based measures is the capability of adapting entropy misfit for measures of person-, item- and model-fit in IRT models. Future research should investigate model-level fit of *EMR* and consider how levels of person-, item- and model-fit collective interact toward the understanding of a model. There are numerous methods to incorporate misfit and calculate power in simulation studies. Misfit was incorporated by varying the percent of contaminated items, percent of contaminated persons, and the type of contamination via response patterns (25% versus 50%). The focus was to create aberrant conditions at the item level of misfit. This had the added benefit that it permitted us to calculate both power for aberrant items and Type I error rates for correctly generated items within each condition. There are other methods of incorporating misfit, however. For example, some researchers use a population-generating model but analyze the data using a different model. Future research may examine different methods to incorporate misfit.

Conclusion

There are advantages when using EMR_j as a measure of item-fit. First, it doesn't rely on binning ability estimates that are based on model dependent arbitrary cut-points as the residual-based measures do. Second, it may discriminate between items differently than the frequently utilized residual-based measures. Third, it takes minimal computing time in comparison to other proposed measures (e.g., χ^2* and G^2*). Fourth, item- and person-fit can simultaneously be evaluated within a dataset. Fifth, it has adequate empirical power and accurate empirical pseudo-Type I error rate across many conditions when used with empirically derived cut-points presented for the purpose of this study, and these values are equivalent or superior to those found with χ^2, G^2, S-X^2, and PV-Q_1 in the same condition. Of caution, *EMR* was evaluated using a threshold whereas the chi-square-based measures utilized significance testing. Sixth, it may be useful for short tests, unlike χ^2 and G^2, which were found to have highly inflated Type I error rates for 10-item tests in the current study. Finally, it can also be used as an approximate measure of fit that exists on a

continuum and doesn't rely on dichotomous hypothesis testing procedures like other traditional measures of item-fit (e.g., χ^2, G^2, S-X^2, S-G^2, χ^2*, and G^2*).

Even though the current study presents cut points, EMR_j is still a continuous measure of approximate item-fit and should not be thought of as dichotomous hypothesis tests. Also consider the recommendation of Dardick and Weiss (2017) that EMR_j is also useful on a relative scale (i.e., not solely in relation to a cut-point) for model comparison.

There is not a single superior measure of item-fit for all data scenarios, thus methodologists rely on a variety of fit statistics to evaluate fit and misfit. As with all item-fit statistics, these measures to flag items to be further investigated. Thus, EMR_j may be particularly useful when used during test development phases for exploratory item revision. In considering EMR_j as a measure of item-fit, note its potential given the reasonable measures of empirical power and empirical pseudo-Type I error rate, and further emphasize the utility it brings to model fit. Methods of fit are often based on χ^2 statistics, however, EMR_j is derived differently than these measures providing additional information that may work well along side of other measures whose mathematical origin are based on χ^2 statistic. $EMR_{j,}$ was introduced as a measure of item-fit in IRT, which captures the amount of uncertainty associated with predicted response options (e.g., correct vs incorrect).

Similar to a forensic investigation, poorly fitting items give us cause for suspicion to investigate using additional empirical methods to discover the reason(s) why. One philosophy for flagging items for misfit is to use multiple measures that target different types of information about the model and use them in conjunction with one another to make determinations regarding the quality of an item, thus providing nonoverlapping information. For example, it would be useful to use a chi-squared (S-X^2 or PV-Q_1) or other residual measure in conjunction with EMR_j to confirm misfit, as different measures may detect different types of contamination and help tell a story of the conspicuous behavior of the item under investigation. Researchers may find EMR_j useful for item-fit or item selection/revision in IRT when used in conjunction with already existing residual-based analyses.

References

Ames, A., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice, 34*(3), 39-48. doi: 10.1111/emip.12067

Bock, R. D. (1960). *Methods and applications of optimal scaling* (Research memorandum 25). Chapel Hill, NC: University of North Carolina Psychometric Laboratory.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*, 195-212. doi: 10.1007/BF01246098

Chalmers, P. (2019). *mirt: Multidimensional item response theory* [R package, version 1.31]. Retrieved from https://cran.r-project.org/package=mirt

Chalmers R. P., & Ng V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement, 41*(5), 372-387. doi: 10.1177/0146621617692079

Chon, K. H., Lee, W., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement, 47*(3), 318-338. doi: 10.1111/j.1745-3984.2010.00116.x

Clark, S. & Muthén, B. (2009). *Relating latent class analysis results to variables not included in the analysis* [Unpublished manuscript]. Retrieved from https://www.statmodel.com/download/relatinglca.pdf

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics, 23*(3), 315-345. doi: 10.1214/aoms/1177729380

Dardick, W. R., & Choi, J. (2016) Teacher empowered assessment: Assessment for the 21st century. *Journal of Applied Educational and Policy Research. 2*(2), 87-98. Retrieved from https://journals.uncc.edu/jaepr/article/view/464

Dardick, W. R., & Weiss, B. A. (2017). Entropy-based measures for person fit in item response theory. *Applied Psychological Measurement. 44*(7), 512-529. doi: 10.1177/0146621617698945

De Ayala, R. J. (2019). Item response theory and Rasch modeling. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (2nd edition, pp. 145-163). New York: Routledge. doi: 10.4324/9781315755649-11

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67-86. doi: 10.1111/j.2044-8317.1985.tb00817.x

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*(1), 101-119. doi: 10.1037/1082-989X.10.1.101

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49. doi: 10.1111/emip.12111

Fox, J. (2015). *Applied regression analysis and generalized linear models* (3rd edition). Los Angeles, CA: Sage Publications, Inc.

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika, 78*(3), 417-440. doi: 10.1007/s11336-012-9305-1

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic. doi: 10.1007/978-94-017-1988-9

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling, 14*(2), 202-226. doi: 10.1080/10705510709336744

Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society: Series B (Methodological), 30*(3), 582-598. doi: 10.1111/j.2517-6161.1968.tb00759.x

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. doi: 10.1080/10705519909540118

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298. doi: 10.1207/S15324818AME1604_2

Maydeu-Olivares, A., Cai, L., & Hernandez, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling, 18*(3), 333-356. doi: 10.1080/10705511.2011.581993

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*(1), 49-57. doi: 10.1177/014662168500900105

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196. doi: 10.1007/BF02294457

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64. doi: 10.1177/01466216000241003

Pastor, D. A., & Gagné, P. (2013). Mean and covariance structure mixture models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd edition, pp. 343-393). Charlotte, NC: Information Age Publishing.

Ramaswamy, V., Desarbo, W., Reibstein, D., & Robinson, W. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12*(1), 103-124. doi: 10.1287/mksc.12.1.103

Rupp, A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55*(1), 3-38.

SAS Institute Inc. (2017). *SAS/STAT 14.2 user's guide*. Cary, NC: SAS Institute Inc. Retrieved from https://documentation.sas.com/?cdcId=statcdc&cdcVersion=14.2

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*(2), 429-449. doi: 10.1348/000711005X66888

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*(1), 58-75. doi: 10.1111/j.1745-3984.2000.tb01076.x

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*(4), 331-352. doi: 10.1111/j.1745-3984.2003.tb01150.x

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (pp. 683-718). Amsterdam: Elsevier. doi: 10.1016/S0169-7161(06)26021-8

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-66). New York: Springer. doi: 10.1007/978-1-4757-2691-6_3

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2. doi: 10.1080/01973533.2015.1012991

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer. doi: 10.1007/978-1-4757-2691-6

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational and Behavioral Statistics, 12*(4), 339-368. 10.3102/10769986012004339

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician, 73*(Sup1), 1-19. doi: 10.1080/00031305.2019.1583913

Weiss, B. A., & Dardick, W. R. (2016). An entropy-based measure for assessing fuzziness in logistic regression. *Educational and Psychological Measurement, 76*(6), 986-1004. doi: 10.1177/0013164415623820

Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*(1), 23-48. doi: 10.1177/001316446902900102

Wright, B., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245-262. doi: 10.1177/014662168100500212

Zhang, B., & Walker, C. (2008). Impact of missing data on person-model fit and person trait estimation. *Applied Psychological Measurement, 32*(6), 466-479. doi: 10.1177/0146621607307692

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). *BILOG-MG: Multi group IRT analysis and test maintenance for binary items* [Computer software, Version 3.0]. Lincolnwood, IL: Scientific Software International.