Entropy-Based Measures for Person Fit in Item Response Theory

Applied Psychological Measurement

2017, Vol. 41(7) 512–529

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617698945

journals.sagepub.com/home/apm

William R. Dardick¹ and Brandi A. Weiss¹

Abstract

This article introduces three new variants of entropy to detect person misfit (E_i , EM_i , and EMR_i), and provides preliminary evidence that these measures are worthy of further investigation. Previously, entropy has been used as a measure of approximate data–model fit to quantify how well individuals are classified into latent classes, and to quantify the quality of classification and separation between groups in logistic regression models. In the current study, entropy is explored through conceptual examples and Monte Carlo simulation comparing entropy with established measures of person fit in item response theory (IRT) such as I_z , I_z^* , U , and W . Simulation results indicated that EM_i and EMR_i were successfully able to detect aberrant response patterns when comparing contaminated and uncontaminated subgroups of persons. In addition, EM_i and EMR_i performed similarly in showing separation between the contaminated and uncontaminated subgroups. However, EMR_i may be advantageous over other measures when subtests include a small number of items. EM_i and EMR_i are recommended for use as approximate person-fit measures for IRT models. These measures of approximate person fit may be useful in making relative judgments about potential persons whose response patterns do not fit the theoretical model.

Keywords

item response theory, person fit, model fit, IRT fit, entropy

There are well-known desirable properties and advantages of item response theory (IRT) models when the model fits such as invariant item and person parameters (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Rupp & Zumbo, 2006). Even when these properties hold in general for a model, there is still a need to investigate fit of individuals to the model. Person-fit scores are investigated for aberrant conditions as the inferences and claims the authors make about an individual are influenced by how well that individual fits the model used.

The purpose of this article was twofold: (a) to introduce three new measures of entropic person fit (E_i , EM_i , and EMR_i) for detecting contamination in IRT and (b) to provide preliminary evidence that these measures are worthy of further investigation. Specifically, in this article, the authors expand entropy from a measure of model fit in latent class analysis (LCA) to a measure

¹The George Washington University, DC, USA

Corresponding Author:

William Dardick, The George Washington University, 2134 G Street NW, Room 105, Washington, DC 20052, USA.

Email: wdardick@gwu.edu

of person fit in IRT models. They also introduce the idea of calculating a separate value of entropy for persons whose predicted scores (correct/incorrect) do not match their observed scores, and comparing this value of entropy with the total amount of entropy for a given person. This measure calculates the amount of entropic misfit and compares that with total entropy. Previously, entropy has been used in LCA to examine the quality of classification in mixture modeling. In this article, the authors (a) provide a generalized conceptual overview of IRT and person-fit indices, (b) introduce a new variant of entropy to account for misfit of persons in dichotomous IRT models via conceptual examples, (c) present hypothetical data scenarios comparing the entropy-based measures with pre-existing measures of person fit, (d) present results from a Monte Carlo simulation study.

IRT

In IRT, factors, traits, or ability in predicting an individual's performance on a test item are considered. These models are falsifiable in that the fit of the model can be assessed. When the model fits, we have invariant properties of IRT: Ability is not test dependent, and item indices are not group dependent. Focusing on the basic dichotomous model for constructing test instruments, the parametric logistic model (PLM) is considered. The probability of getting an item correct, $P(\theta)$, where item parameters change as a logistic function can be expressed as a 3PLM:

$$P(\theta) = c_i + (1 - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \quad (1)$$

where e is the exponential constant 2.718, and θ is the latent trait of a given person's ability. For a given item i , there are three parameters in the model: a , b , and c , where a is the slope at the inflection point of the model or discrimination, b is the item difficulty, c is the lower asymptote and pseudochance parameter. Other frequently used models can be obtained by setting c to the constant 0 (2PLM), and a to some constant, identical across all items (1PLM).

IRT: Person Fit

Statistical models, which do not perfectly reproduce sample data, are often of interest to researchers. When examining a particular statistical model, researchers turn to measures of model fit to examine how well the global model fits to the actual data. Measures of data-model fit indicate how well a model fits to data, and are essential evidence to support the validity of test scores. Model fit indices frequently utilized in IRT include the likelihood-ratio test, information criteria, (Rupp, 2013), and chi-square goodness-of-fit tests (Swaminathan, Hambleton, & Rogers, 2007). Fit can also be examined at the item and person levels for local assessment (Rupp, 2013; Swaminathan et al., 2007). Much methodological IRT model-fit research has focused on assessing global IRT assumptions (e.g., correct dimensionality, local independence, examinee response independence, global model fit, and item-fit statistics), which aggregate item scores across persons. On the contrary, person-fit statistics provide a numerical measure of how well persons fit to the model by aggregating a person's responses across items and using a statistical method to evaluate fit (Embretson & Reise, 2000; Meijer & Sijtsma, 2001).

Person-fit measures are helpful in detecting misfitting item-score patterns, which may systematically uncover specific behaviors such as guessing, cheating, plotting, sleeping, or other idiosyncratic mechanisms of misfit. Selection criteria for detecting misfit can be considered solely or in combination when one statistic is often not sufficient (Miejer, Niessen, & Tendeiro, 2016). For an in-depth discussion of person fit, the authors refer readers to the seminal work of

Meijer and Sijtsma (2001), comprising an extensive review of person-fit statistics throughout classical test theory and IRT, and Rupp (2013), presenting a more recent review of simulation studies involving person fit.

Likelihood-based statistics are the most frequently investigated measures of person fit. Likelihood-based statistics such as l_z (the likelihood statistic; Drasgow, Levine, & Williams, 1985) and l_z^* (Snijders, 2001) assess the likelihood of score patterns, given an IRT model. Other frequently used measures of person fit in IRT are residual based such as U and W (Wright & Masters, 1982; Wright & Stone, 1979). Person-fit statistics such as these vary in the type of misfit and aberrant response patterns they detect and they classify misfit differently. One type of measure that has not yet been explored in the person-fit IRT literature are entropy-based measures, which could assess the amount of separation between groups. Previously, entropy has been used as an approximate measure of data-model fit to quantify how well individuals are classified into latent classes (Celeux & Soromenho, 1996; Henson, Reise, & Kim, 2007), and to quantify the quality of classification and separation between groups in logistic regression models (Weiss & Dardick, 2016). Approximate fit indices are continuous measures used to supplement tests of statistical significance (Hu & Bentler, 1999). An entropy-based person-fit index may not discriminate between persons the same way as likelihood-based and residual-based measures.

Entropy in LCA

In LCA, entropy is sometimes used as a measure of data-model fit. Entropy is a classification-based approach that helps researchers determine the number of latent classes that exist in a population (Celeux & Soromenho, 1996; Clark & Muthén, 2009; Henson et al., 2007; Pastor & Gagne, 2013). In LCA, the entropy index is calculated as

$$\text{Entropy} = 1 - \frac{E}{N \ln K}, \quad (2)$$

where N is the number of people, K is the number of latent classes for a given model, and E_k is defined as

$$E = \sum_{i=1}^N \sum_{k=1}^K (-p_{ik} * \ln p_{ik}), \quad (3)$$

where $E \geq 0$, and p_{ik} represents the conditional posterior probabilities calculated for each observation, i , and denotes the probability of membership in each of the K classes. Entropy in the context of LCA can be interpreted as the discrimination capability for conditional posterior probabilities. Entropy ranges from 0 to 1, where values closer to 1 occur with larger class separation. The number of latent classes is unknown; therefore, researchers compare models with different numbers of classes to determine which model results in the largest value of entropy.

Methodologists have placed increased emphasis on the use of multiple fit indices in IRT along with other fields such as structural equation modeling (SEM), LCA, and logistic regression (Kline, 2016; Schumacker & Lomax, 2015). The authors propose a new variant of entropy to measure IRT person fit. Entropy-based fit statistics can provide unique information that may not be accounted for by the typical group based or likelihood methods, and may be used separately or in conjunction with other statistics to flag aberrant response patterns.

New Variants of Entropic Person Fit

In this section, the authors introduce three new variants of entropic person fit: The first applies entropy from LCA to IRT as a measure of person fit (E_i). In IRT, observed scores for person i on item j are known. Thus, the authors know if their predictions of whether or not a person answered an item correctly are correct, providing them with the unique ability to evaluate the quality of their predictions by comparing predicted scores with observed scores. The second measure they present, EM_i , calculates entropy for those whose predicted outcomes do not match the observed outcome providing us with an entropy value for misfitting persons only. Finally, the third measure of entropy they present, EMR_i , is a ratio representing the proportion of entropic misfit (EM_i) to overall entropy (E_i).

Entropy for Person Fit

Consider scores for n persons on a set of J items fit to a simplistic dichotomous IRT model. In LCA, true class membership is unknown; however, in IRT the dichotomous outcome for each item j is known (i.e., correct vs. incorrect). To calculate entropy for the dichotomous IRT model for a given person, Equation 2 can be adapted to examine person fit across items:

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

where J is the number of items, K is the number of response categories (2 for dichotomous correct/incorrect items), and E_i^* represents the unstandardized entropy of the set of items j for person i and is defined by

$$E_i^* = \sum_{j=1}^J \sum_{k=1}^K (-p_{ijk} * \ln p_{ijk}). \quad (5)$$

$E_i^* \geq 0$, and p_{ijk} represents the predicted probabilities from the PLM calculated across all items for each person, and denotes the probability the person i endorses an item. Entropy, E_i , is calculated for each person across all items, and represents a standardized format of E_i^* in which total entropy is bound between 0 and 1, where higher values are indicative of more distinct separation among categories of response.

Entropy Weighted for Misfit

Because the correctness of item responses is known, the authors have the unique ability to evaluate the quality of their predictions by comparing predicted scores with observed scores. This information can be used to aid in the interpretation of total entropy by knowing the amount of entropic misfit (i.e., for person-item responses knowing how far off predictions were from observed correct/incorrect responses). For example, an undesirable model is one that predicts that person i answered item j incorrectly, when the observed item response is actually correct. Knowing the observed score for person i on item j allows us to calculate entropic misfit for model predictions that do not match observed data. Thus, the authors modify the calculation of entropy to reflect the amount of misfit. EM_i partitions predicted probabilities of a response (0,1) into those that fit (correct classification) and those that do not fit (incorrect classification). Following Equations 3 and 4, we get

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

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

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

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. x_j is a weight and defined as

$$x_j = \begin{cases} 1 & \text{correct classification,} \\ 0 & \text{incorrect classification} \end{cases} \quad (8)$$

A cell-level entropy value is effectively calculated for each item–person combination and averaged across items to obtain a person-level entropy value. Alternatively, one can think of this as an entropy matrix or persons by items where persons who were correctly classified received cell scores of 0 to reflect no misfit, while persons who were incorrectly classified receive cell scores that range asymptotically from 0 to 1, which reflect the amount of misfit for that particular person–item combination.

For the 1 or 2 PLMs, a predictive probability at or above .5 indicates that a person is more likely to endorse an item with a response of 1 (correct) rather than 0 (incorrect). Predicted probabilities close to .5 will result in entropy values near 0 because we are less confident in our predictions. Predicted probabilities of 0 or 1 will result in entropy values closer to 1 because we are more confident in the predicted group membership. In this description, the authors used the predicted probability of .5 as the cut-point in determining misfit. EM_i , however, uses the weight $1 - x_j$, where x_j is defined as correct classification, which will result in calculating values of entropy only for incorrect classifications. It can be interpreted as when entropy increases, classification decreases. Thus, smaller values of EM_i are more desirable and indicative of better quality predictions.

Entropy Misfit Ratio

Finally, to compare misfit among alternative levels of entropy, a ratio of entropic misfit is presented, represented here by EMR_i , which is calculated by taking a ratio of entropy misfit, EM_i , to total entropy, E_i , thus representing the relative strength of fit that is also misfit. In mathematical formula, entropic misfit is calculated as

$$EMR_i = \frac{EM_i}{E_i}. \quad (9)$$

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 thus indicative of less misfit (i.e., fewer incorrect predictions) and are thus more desirable. This ratio may enhance interpretation as misfit is comparable with different item– θ combinations on different tests. In the next section, three hypothetical

scenarios are presented to exemplify how the variant(s) of entropic misfit (E_i , EM_i , and EMR_i) capture unique information not contained in likelihood measures of person fit in IRT.

Hypothetical Scenarios for Entropy

Three scenarios are presented in Table 1 based on data from Embretson and Reise (2000). Each scenario contains 15 examinees, each with a different possible pattern of responses on six items. The following constants were set in all scenarios: discrimination = 1; difficulty (left to right) of $-2, -1, -0.5, 0.5, 1.0, 2.0$; θ estimates differ between the scenarios. The likelihood statistic, l_z , is included as it is one of the most commonly utilized measures of person fit (Rupp, 2013). Smaller values of l_z are more desirable as they are indicative of better fit.

In Scenario 1 (Table 1), $\theta = .96$ for all persons, but as is evident, entropic misfit was greatest for the first examinee who had incorrect responses for four of six items; only the two most difficult items were correctly endorsed. Furthermore, we can see that the EMR_i portioned misfit was not captured by l_z . In particular, identical l_z values are observed for examinees 7-8-9 ($l_z = -1.57$), whereas EMR_i differentiates between these three examinees (Examinee 7 being largest EMR_i of the three, surpassing the value of Person 6). On the contrary, Examinees 8, 10, and 11 have very different l_z values but similar EMR_i . These results indicate that EMR_i ranks misfit differently than l_z and may change dispersion. Differentiation in misfit can also be seen with Examinee 7 being classified higher than Examinee 6. This may reflect some of the disparities in selecting the easiest and most difficult item over an alternative form of misfit (this finding is explored further in Scenario 3). In Scenario 2, θ was changed to $+.96$. Similar comparisons between l_z and EMR_i can be seen; however, the two measures rank ordered some examinees differently and changed spread of items.

For Scenario 3, θ was randomly generated $n(0,1)$ and differs across examinees to represent local person fit, where the six items may be selected from a larger test, or alternatively discriminations in the 2PLM vary (Embretson & Reise, 2000). Examinees 1 and 2 had similar values of l_z (-4.33 and -4.45 , respectively), while their EMR_i values were very different (0.96 and 0.54 , respectively).

Consider the following errors made by Examinees 1 and 2: Examinee 1, predicted probabilities suggest five errors (Items 1-3 predicted correct; Items 5 and 6 incorrect); Examinee 2, only two errors were made (those endorsed as 1). Based on l_z , the authors conclude that these two equally misfit, but based on entropy, they could conclude Examinee 2 to be substantially less misfit than Examinee 1.

Several illustrations of response patterns that provide entropic misfit, EM_i and EMR_i values of 0 or close to 0, exist in the three scenarios: For example, referring to Scenario 1 (Table 1) for Examinee 15, the predicted responses for all six items were correctly classified based on the predicted probabilities, resulting in EM_i and EMR_i of 0 for this person. For Examinee 14, the predicted responses for five of six items were correctly classified based on the predicted probabilities; thus, this examinee has EM_i and EMR_i close to but not equal to 0.

For EM_i , the maximum amount of misfit that can occur is the total entropy in the model for that examinee. For example, in Scenario 2, Examinee 1 has an EM_i value of .28; this is the maximum amount of misfit that can occur for Examinee 1. However, because EMR_i is a ratio of how much entropy is misfit, a value of 1 is obtained for EMR_i .

Method

Simulation Study Design

The purpose of this simulation study was to show that the new variants of entropy can be used as measures of person fit. More specifically, the authors sought to determine if entropy (E_i),

Table 1. Three Hypothetical Scenarios of Entropy Values for 15 Examinees on Six items.

Examinee	Observed response pattern ^a	Scenario 1: $\theta = -.96$						Scenario 2: $\theta = -.96$						Scenario 3: $\theta = \text{RANDOM } N(0,1)$					
		θ	I_z	E_i	EM_i	EMR_i	θ	I_z	E_i	EM_i	EMR_i	θ	I_z	E_i	EM_i	EMR_i			
1	000011	-.96	-4.10	0.28	0.22	0.80	.96	-5.71	0.28	0.28	1.00	-.08	-4.33	0.23	0.22	0.96			
2	000101	-.96	-3.68	0.28	0.20	0.71	.96	-5.29	0.28	0.28	0.98	-2.14	-4.45	0.47	0.26	0.54			
3	001001	-.96	-2.83	0.28	0.15	0.55	.96	-4.45	0.28	0.23	0.82	2.06	-6.77	0.46	0.35	0.77			
4	000110	-.96	-2.83	0.28	0.16	0.55	.96	-4.45	0.28	0.25	0.88	-.31	-2.90	0.23	0.14	0.59			
5	010001	-.96	-2.41	0.28	0.15	0.53	.96	-4.03	0.28	0.20	0.73	-.76	-2.38	0.26	0.15	0.58			
6	001010	-.96	-1.99	0.28	0.11	0.40	.96	-3.61	0.28	0.20	0.72	1.27	-4.23	0.32	0.24	0.74			
7	100001	-.96	-1.57	0.28	0.12	0.42	.96	-3.19	0.28	0.16	0.58	-.08	-1.75	0.23	0.11	0.49			
8	001100	-.96	-1.57	0.28	0.09	0.30	.96	-3.19	0.28	0.20	0.70	-.61	-1.53	0.25	0.08	0.34			
9	010010	-.96	-1.57	0.28	0.11	0.37	.96	-3.19	0.28	0.18	0.63	.41	-2.28	0.24	0.13	0.55			
10	010100	-.96	-1.15	0.28	0.08	0.28	.96	-2.77	0.28	0.17	0.60	-.32	-1.18	0.23	0.08	0.35			
11	100010	-.96	-0.73	0.28	0.08	0.27	.96	-2.35	0.28	0.13	0.47	-.08	-4.33	0.23	0.22	0.96			
12	100100	-.96	-0.31	0.28	0.05	0.18	.96	-1.93	0.28	0.13	0.45	-2.14	-4.45	0.47	0.26	0.54			
13	011000	-.96	-0.31	0.28	0.03	0.12	.96	-1.93	0.28	0.13	0.45	2.06	-6.77	0.46	0.35	0.77			
14	101000	-.96	0.53	0.28	0.01	0.02	.96	-1.08	0.28	0.08	0.29	-.31	-2.90	0.23	0.14	0.59			
15	110000	-.96	0.95	0.28	0.00	0.00	.96	-0.66	0.28	0.06	0.20	-.76	-2.38	0.26	0.15	0.58			

Source. This example is extended from Embretson and Reise (2000).

Note. I_z = the likelihood statistic, E_i = total entropy, EM_i = misfit entropy, EMR_i = the ratio of misfit to total entropy (calculated as EM_i/E_i).

^aObserved response pattern on six items where discrimination = 1 for all items, difficulty for items is -2, -1, -0.5, 0.5, 1.0, 2.0 (left to right), where scores of 0 = incorrect and 1 = correct.

entropy misfit (EM_i), or entropy ratio (EMR_i) were able to detect aberrant behavior when a subset of contaminated items were present. In addition to examining E_i , EM_i , and EMR_i ability to detect aberrant behavior, four well-established statistics, U , W , I_z , and I_z^* , were calculated.

Constants. The Monte Carlo method was used to generate a test with 40 items and 1,000 simulees representing person's responses. Data were simulated using a 1PLM model with random noise around the measures and incorporated misfit with random guessing. Item difficulties were generated from a random normal distribution $N(0,1)$ with a slope set to 1.

For the uncontaminated subtest, all simulated participants were generated to have an ability (θ) score from a random distribution $N(0,1)$ to generate a probability of response scores generated to conform to a 1PLM with some error. The probability was then used for each item-person interaction to generate a 0,1 response by using the generated probability and comparing it with a random probability drawn from a uniform distribution. If the random probability was less than the 1PLM probability, the generated response was correct, a response of 1; otherwise, it is 0 for incorrect.

Variables. Three variables were considered in the simulation design: the percentage of items contaminated, the percentage of persons contaminated, and the guessing parameter. While the overall test length of 40 items remained constant, the subtest lengths were varied to contain different percentages of contaminated items, representing 10%, 25%, and 50% contamination. Correspondingly, the uncontaminated and contaminated subtests included the following number of items: 36 and 4 (10% contamination), 30 and 10 (25% contamination), and 20 and 20 (50% contamination), respectively.

Misfit can be examined at the item level, or it can be varied at the person level. The percentage of simulated persons contaminated within the subtests was also varied to be 10, 25, and 50. Random responses act as aberrant responses, and are thus considered a source of misfit. This form of guessing could not only represent "random response" to a subsection but it also fits well with "fatigued responding," "content knowledge across subparts of the assessment," and possible a few other forms of aberrant behavior (Rupp, 2013).

The contaminated persons within the contaminated subtest were modified by two types of guessing behavior: 25% and 50%. The first type of guessing behavior represents a random response when guessing is more likely to lead to a correct response than using than the probability that the respondent knows the answer. This was intended to replicate a realistic scenario in which persons were simulated with a random guessing response of .25 when the probability of answering the item correctly was lower than .25. The second condition can represent simulated participants' ability to reduce a question down to two responses when distractors are not well constructed and guess randomly, thus increasing chances of a correct response, or it could represent a guessing response to an alternate choice question with only two responses. Similar but stronger contamination than the first condition, persons were simulated with random guessing responses of .50 when the probability of answering the item correctly was lower than .50. The uncontaminated portion of simulated persons within the contaminated subtest was simulated in the same manner as for the uncontaminated subtest.

Summary. In summary, the simulation study was comprised of 18 conditions in a fully crossed $3 \times 3 \times 2$ design: three contaminated subtest lengths (10%, 25%, and 50%), three contaminated proportion of simulated persons (10%, 25%, and 50%), and two types of contaminated responses (25% and 50% representative guessing). Within each of the 18 conditions, two groups were established for comparison: the contaminated and uncontaminated subgroups. Separation between these two subgroups was examined for the 40-item test, and for the contaminated and

Table 2. Average Spearman's Rank Correlation Coefficient Matrix.

Fit statistics	E_i	EM_i	EMR_i	U	W	I_z	I_z^*
E_i	1.00						
EM_i	-.03	1.00					
EMR_i	-.49	.83	1.00				
U	-.32	.86	.84	1.00			
W	-.40	.83	.86	.86	1.00		
I_z	-.02	-.90	-.79	-.82	-.88	1.00	
I_z^*	.00	-.92	-.77	-.86	-.89	.98	1.00

Note. E_i = entropy; EM_i = entropy misfit; EMR_i = entropy misfit ratio.

uncontaminated subtests. Overall, there were 1,000 replications for each condition on the 40-item test, each with 1,000 simulated persons.

Data were generated and analyzed using SAS 9.4 (2013) software supported by “Calling Functions in the R Language” (SAS Institute, 2013) to incorporate R (R Core Team, 2016). The item and person parameters were estimated using the IRT procedure in SAS. The default estimation method for item parameters in SAS is the gradient-based convergence criterion for the quasi-Newton algorithm (SAS Institute, 2013). Maximum-likelihood (ML) person parameters were estimated using SAS. E_i , EM_i , and EMR_i , U , and W were calculated using SAS/IML 13.2 software, while the R-package PerFit (Tendeiro, 2016) was used to calculate I_z and I_z^* .

Results

This study was comprised of 18 conditions, 1,000 replications for each condition on a 40-item test, each with 1,000 simulated persons. Descriptive statistics for fit indices are presented in Figures 1 through 6 and Tables A1 through A4; Spearman's rank correlation coefficients among person-fit statistics are provided in Table 2. Results for the descriptive statistics are presented separately for the entire 40-item test (Figures 1-2, Table A1), the uncontaminated subtest (Figures 3-4, Table A2), and the contaminated subtest (Figures 5-6, Table A3). Descriptive statistics were averaged across the 1,000 repetitions for uncontaminated persons and contaminated persons separately. For example, in the conditions in which 10% of persons were contaminated, results are presented separately for the 900 uncontaminated persons and the 100 contaminated persons.

While the tables (located in the online appendix) are useful for viewing exact numerical values and stability across conditions, the figures are useful to graphically examine separation between the two subgroups (uncontaminated and contaminated persons). Each figure contains multiple blocks of person-fit statistics. Figures 1, 3, and 5 contain four blocks of statistics: U (upper left), W (upper right), I_z (lower left), and I_z^* (lower right). Figures 2, 4, and 6 contain three blocks of the entropy-based person-fit measures: E_i (upper left), EM_i (upper right), EMR_i (lower left). Within each block, there are 3×3 cells that contain the percentage of persons contaminated (10% = 100 persons, 25% = 250 persons, and 50% = 500 persons) across the graph and the percentage of items contaminated (10% = four items, 25% = 10 items, 50% = 20 items) down the graph. The two levels of guessing are displayed within each cell on the x axis: 25% and 50%. Within each cell are vertical box plots with the mean emphasized in the center as a circle and the median as a line bar. Each cell compares the uncontaminated subset of persons (dark gray) with the contaminated subsets of persons (light gray). The y axis for each cell represents the values of the person-fit statistics.

Figure 1. U , W , I_z , and I_z^* measures across all 40 items.

The authors caution readers to keep in mind that as an indirect consequence of varying the percentage of contaminated items, the number of items in each subtest may vary. For example, when 10% of items were contaminated, 40 items were on the total test, 36 items were on the uncontaminated subtest, and four items were on the contaminated subtest.

The 40-Item Test

Descriptive statistics for the seven person-fit statistics on the 40-item test are shown separately for the two subgroups in Table A1. Figure 1 conveys the same information graphically for U , W , I_z , and I_z^* , while Figure 2 shows graphical descriptive statistics for E_i , EM_i , and EMR_i . For all seven fit statistics, separation between the contaminated and uncontaminated subgroups increased as the percentage of guessing and the percentage of contaminated items increase. The greatest separation between the two subgroups occurred in conditions with 50% contaminated items (20 items). For the contaminated persons, as the percentage of contaminated persons increased, the standard deviation decreased, as expected because the number of people represented by that number increased.

U and W had a similar pattern of descriptive statistics, while I_z and I_z^* exhibited similar patterns of descriptive statistics across all conditions. Of note, U and W values around 1.0 are desirable, while I_z and I_z^* values greater than or equal to 0 are desirable. Thus, U , W , I_z , and I_z^* convey similar information, even though U and W are in the opposite direction of I_z and I_z^* . E_i did not separate between subgroups as well as the other fit indices. Of the three entropy measures, E_i was the least stable across the 18 simulated conditions. As contamination (both persons and items) increased, EM_i values decreased, implying that there was some instability in values across the conditions. EMR_i performed similarly to the other fit indices; however, it was

Figure 2. E_i , EM_i , EMR_i across all 40 items.
Note. E_i = entropy; EM_i = entropy misfit; EMR_i = entropy misfit ratio.

advantageous, in that it yielded stable values across simulated conditions. For example, when comparing EMR_i with l_z^* in the contaminated persons subgroup both statistics separated well; however, for the uncontaminated subgroup EMR_i yielded more stable fit statistic values across all conditions in comparison with l_z^* . Of note, the SD for E_i , EM_i , and EMR_i remained stable across all simulated conditions.

The Uncontaminated Subtest

Descriptive statistics for the seven person-fit statistics on the uncontaminated subtest are shown separately for the two subgroups in Table A2 of the online appendix. Figures 3 and 4 convey the same information graphically; U , W , l_z , and l_z^* are shown in the former figure, while E_i , EM_i , and EMR_i are shown in the latter figure.

The expected values for contaminated and uncontaminated subgroups do not differ on the uncontaminated subtest. Consistent with this expectation, EMR_i , l_z , and l_z^* did not detect differences between the contaminated and uncontaminated subgroups on the uncontaminated subtest. Interestingly, E_i and EM_i (and U and W to a lesser extent) detected some amount of contamination even in the uncontaminated subtest (i.e., inflated false alarm rate for these measures). This inconsistency was more likely to occur in the extreme conditions (i.e., as percentage of guessing, percentage of persons contaminated, and percentage of contaminated items increased). E_i was the most sensitive to false alarm rate.

Figure 3. U , W , I_z , and I_z^* measures across the uncontaminated subset of items.

The Contaminated Subtest

Descriptive statistics for the seven person-fit statistics on the uncontaminated subtest are shown separately for the two subgroups in Table A3 of the online appendix. Figures 5 and 6 convey the same information graphically; U , W , I_z , and I_z^* are shown in the former figure, while E_i , EM_i , and EMR_i are shown in the latter figure.

The authors expected to observe the greatest separation between subgroups on the contaminated subtest. Specifically, on the contaminated subtest, they expected fit indices to yield desirable values for uncontaminated persons and poor values for contaminated persons. All measures separated well between subgroups across all conditions. However, separation was most noticeable when guessing was 50%. For example, in Figure 6, EMR_i shows the most extreme separation observed at the 50% guessing conditions, even when only 10% of persons and 10% of items were contaminated.

Desirable person-fit indices should also yield consistent values across all conditions for uncontaminated persons on contaminated items. EM_i yielded the most consistent results with values ranging from .022 to .029 (range of .007) across conditions, indicating that it was not sensitive to any variables of the simulation study, and consequently was not sensitive to the number of items on the subtest. EMR_i was the most stable across percentage of persons contaminated and percentage of guessing; however, it yielded slightly more desirable values as percentage of items contaminated increased (see Figure 6). In comparison, for uncontaminated persons on contaminated items, I_z and I_z^* fluctuated more across variables in the study than EM_i and EMR_i .

As expected, person-fit indices for the contaminated persons on the contaminated subtest were the least desirable values in comparison with uncontaminated persons on either subtest and contaminated persons on the uncontaminated subtest. EM_i and EMR_i values for

Figure 4. E , EM , EMR across the uncontaminated subset of items.

Figure 5. U , W , L_z , and L_z^* measures across the contaminated subset of items.

Figure 6. E_i , EM_i , EMR_i across the contaminated subset of items.
Note. E_i = entropy; EM_i = entropy misfit; EMR_i = entropy misfit ratio.

contaminated persons on the contaminated subtest increased when the chance of guessing increased from 25% to 50%. This finding indicates that persons who were able to narrow response options from four to two were more likely to be flagged as engaging in misfitting behavior as indicated by the increased EM_i and consequently EMR_i values.

Spearman's Rank Correlation Coefficients

Table 2 contains the Spearman rank correlation coefficients between the three entropy measures and the four pre-existing measures of person fit. The average SD s for these correlation coefficients were very small, and are thus not shown here; they ranged from .02 to .04 for correlations with E_i but were about .01 for all other combinations of measures. The standard errors (i.e., SD s around the means) were even smaller, ranging from .01 to .03 for E_i with other measures, and were $< .01$ between all other combinations of measures. Given the small amount of variance around these correlations, the authors collapsed across the 18 conditions.

The strongest average correlation observed was between the two likelihood-based statistics (I_z and I_z^* ; $r = .98$), indicating that the two measures rank ordered examinees similarly and contained redundant information. The correlation between U and W was also strong ($r = .87$). E_i was weakly correlated with all measures. EM_i and EMR_i were strongly correlated with the likelihood-based and residual-based measures in the expected directions ($|.77| < r < |.92|$), providing convergent validity evidence. However, EM_i and EMR_i were not perfectly correlated with the four pre-existing measures of person fit, demonstrating that they have nonoverlapping variance, and thus providing discriminant validity evidence for these measures.

Discussion

The purpose of this article was twofold: (a) to introduce three new measures of entropic person fit (E_i , EM_i , and EMR_i) for detecting contamination in IRT, and (b) to provide preliminary evidence that these measures are worthy of further investigation. The current article illustrated that the entropy misfit (EM_i) and entropic misfit ratio (EMR_i) successfully detected aberrant conditions when comparing contaminated and uncontaminated subgroups of persons. In comparison with the four established measures of person fit in IRT (I_z , I_z^* , U , and W), EM_i and EMR_i performed similarly in showing separation between the contaminated and uncontaminated subgroups. EMR_i yielded stable results across simulation conditions, whereas other measures such as U , W , I_z , and I_z^* showed unstable results across some conditions. EMR_i may be advantageous when the number of items on a subtest is small. The unadjusted entropy measure (E_i) was not useful in predicting aberrant condition but may be picking up different elements of aberration as change can readily be found as contamination increases.

As is desirable, I_z , I_z^* , and EMR_i showed no noticeable difference between contaminated and uncontaminated subgroups in the uncontaminated condition. However, E_i and EM_i (and to a lesser degree in the extreme contaminated conditions U and W) detected some amount of contamination even in the uncontaminated subtest, indicating that these measures may have a slight inflated false alarm rate. An inflated false alarm occurs when the response patterns of uncontaminated examinees are incorrectly flagged by a person-fit index. One explanation for this inflated false alarm rate is that the measures may be sensitive to the global person disturbance. More specifically, this may be an artifact of the misfitting data affecting the parameter estimates. The authors remind readers to be cautious of interpreting parameters based on incorrect models or misfitting data. Interpretation of any parameter estimates obtained from sample data is based on the assumption that one has the correct model, and that all persons fit to that model. The inflated false alarm rate merits further investigation.

As expected, simulated persons who were able to narrow response options down from four to two were more likely to be flagged as engaging in misfitting behavior as indicated by the increased EM_i and consequently EMR_i values. This makes sense, if a person can narrow response options from four to two, then we can expect that person to answer that item correctly. However if he or she answers the item incorrectly, then that person's response will be flagged as misfitting from expectations, thus resulting in a higher (i.e., less desirable) EMR_i value.

Entropy (E_i) without any adjustment did not accurately detect aberrant conditions when comparing the two subgroups for contaminated and uncontaminated persons. This is unsurprising as entropy determines strength of separation in classification, not accuracy (i.e., E_i values are larger for persons from the point of classification regardless of whether they are correctly classified). For example, a person who has extreme predicted probabilities near 0 and 1 and is correctly classified (e.g., predicted to answer correctly and observed to answer correctly) will have a higher value of E_i than a person who has predicted probabilities close to the classification point but was incorrectly classified. This is one of the difficulties with classification interpretations using entropy in LCA (i.e., entropy in LCA determines the strength of separation, but the correctness of classification predictions is unknown, thus misfit cannot be adjusted for). While E_i may be useful to detect separation, it does incorporate the correctness of classification into the measure. When entropy was adjusted to incorporate misfit, the measure improved in terms of separation and stability. Unlike in LCA models, in IRT the outcome of a person's response (correct vs. incorrect) is known, and thus observed and predicted outcomes can be compared and misfit can be adjusted for. Using this knowledge, entropic misfit is a penalty for making incorrect predictions. The penalty is small for values close to the classification point

(e.g., predicted probabilities of .51 for being in the incorrect class) and large for values from the classification point (e.g., predicted probabilities of .90 for incorrect responses).

The authors hypothesized that the entropy measures would have both convergent and discriminant validity evidence. This hypothesis was supported, in that EM and EMR were strongly correlated with the other measures in the expected directions (I_z , I_z^* , U , and W , convergent validity) but had unique variance (discriminant validity). The variation in rank ordering of aberrant responses between these measures is an indication that the entropy-based measures discriminated between persons differently than the likelihood- and residual-based statistics, and may be useful in conjunction with person-fit indices currently used in IRT models. Of note, the average correlation between I_z and I_z^* was very strong ($r = .98$), indicating that the two measures have some redundancy in how they rank order examinees by aberrant behavior.

Future Research

The fit indices in the investigation are on different scales, thus making it difficult to directly compare the measures. The purpose of the current investigation was to determine if entropy measures merited further investigation as a person-fit measure but not to establish a cut-point for entropy measures. Entropy has been used in LCA for over 20 years (Celeux & Soromenho, 1996) as a measure of approximate model fit; to date, no cut-points have been identified for entropy in the latent literature, thus no Type I error and no statistical power have been presented for entropy in the literature. As Kline (2016) stated, identifying a cut-point for approximate fit indices is nontrivial because the values of the statistic do not directly correspond with the seriousness of the type of specification error. In the future, the authors intend to investigate entropic measures further to establish reasonable cut-points for the entropy and the extended misfit measures, EM_i and EMR_i .

In introducing entropic misfit in this article, a basic 1PLM IRT model was used. However, these measures can easily be extended to more complex IRT models. For the set of typical dichotomous IRT models (1-3 PLM) and their basic extensions (e.g., 4 PLM), the mathematics of entropy does not change as the entropy measures are based on the probability of classification p_{jk} for a dichotomous 0,1 outcome. However, future simulation studies may wish to investigate how the three entropy-based person-fit statistics function when using different IRT models.

The number of items on a test or subtest is an important variable in the stability of person-fit measures. Entropy measures, EM_i and EMR_i , may be less sensitive to the number of items, and thus may be beneficial to person fit when the number of items on a test/subtest is small. The current study indirectly investigated the impact number of items on a test/subtest had on person-fit indices by varying the percentage of items that were contaminated. Future studies should investigate this by directly varying the number of items on a subtest.

The current article focused on person fit, in particular the misfit of persons to a model. The formulas presented show how entropy is calculated at the cell level; thus, a matrix of entropy values at every person-by-item combination is obtained. To calculate entropy at the person level in the current study, the authors summed across the item-level cells for each person; to calculate entropic misfit, they summed across the item-level cells for each person only for misfitting items. Although the focus of this article was on person fit, one could also sum down the person-level cells for each item to obtain an entropy-based item-fit measure. Similarly, entropic misfit could be obtained at the item level if only the person-level cells for misfitting persons for each item were summed down. Finally, to calculate entropy at the test level one could sum across and down the matrix to combine all cell-level values of entropy to obtain a measure of data-model fit. Entropy could then be used to compare models at the test level. In the current study,

the authors choose to focus on entropy as a measure of person fit. However, in future studies they will extend their three measures of entropy to the item- and test-level in more depth.

Conclusion and Recommendations

Methodologist utilizes various person-fit statistics, often in combination with each other, to explain types and rates of aberrant responses. The variants of entropy presented in this article can provide researchers assistance detecting person misfit. More importantly, the entropy-based measures the authors propose potentially offer detection of unique variance in comparison with established measures of person misfit such as I_z , I_z^* , U , and W . Entropic misfit, EMR_i , utilizes model prediction errors differently than likelihood-based methods. This article sets the reasoning for the authors' further investigation of entropy and how it may be of value in IRT models. In general, these measures of entropy work for all dichotomous models and can easily be extended to polytomous models as well.

Traditionally, the purpose of person-fit indices is to flag persons whose response patterns do not fit the model. Previously existing measures of person fit in IRT (such as I_z , I_z^* , U , and W) provide researchers with a statistical significance test to do this. On the contrary, the authors recommend EM_i and EMR_i for use as approximate person-fit indices. Approximate fit indices should not be confused with statistical significance tests. They are not to be used as dichotomous (reject vs. fail-to-reject) decision points, do not incorporate sampling error, and should be used as continuous measures indicating how persons fit to data (Kline, 2016). Fit indices are continuous measures used to supplement tests of statistical significance (Hu & Bentler, 1999). Furthermore, the authors caution readers to use EM_i and EMR_i as relative indices rather than absolute. That is, these measures are useful in identifying person misfit within a given population, sample, and model.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online appendices are available at <http://journals.sagepub.com/doi/suppl/10.1177/0146621617698945>

References

- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13, 195-212.
- Clark, S., & Muthén, B. (2009). *Relating latent class analysis results to variables not included in the analysis*. Retrieved from <http://www.statmodel.com/download/relatinglca.pdf>
- 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, 67-86.
- Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic.

- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- 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, 202-226.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Miejer, R., Niessen, A., & Tendeiro, J. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, 23, 52-62.
- Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107-135.
- Pastor, D. A., & Gagne, P. (2013). Mean and covariance structure mixture models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 343-394). Charlotte, NC: Information Age.
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- 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, 3-38.
- Rupp, A., & Zumbo, B. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66, 63-84.
- SAS Institute. (2013). *SAS 9.4 language reference concepts*. Cary, NC: Author.
- Schumacker, R., & Lomax, R. (2015). *A beginner's guide to structural equation modeling* (4th ed.). New York, NY: Routledge.
- Snijders, T. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66, 331-342.
- Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26: Psychometrics* (pp. 683-718). Amsterdam, The Netherlands: Elsevier.
- Tendeiro. (2016). PerFit: Person Fit. R package Version: 1.4.1. Retrieved from <https://cran.r-project.org/web/packages/PerFit/PerFit.pdf>
- Weiss, B. A., & Dardick, W. R. (2016). An entropy-based measure for assessing fuzziness in logistic regression. *Educational and Psychological Measurement*, 76, 986-1004.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.