Evaluating the Performance of Person-Fit Detection Methods in Diagnostic Classification Models

Jeffrey C. Hoover[1], W. Jake Thompson[1]

[1]University of Kansas; Accessible Teaching, Learning, and Assessment Systems (ATLAS)

Correspondence concerning this article should be addressed to Jeffrey C. Hoover: jhoover4@ku.edu

Abstract

Person-level model fit (i.e., person-fit) should be evaluated in addition to evidence of test- and item-level model fit because person-fit has implications for our confidence in the student-level inferences made from assessment results. In this manuscript, we compared the Type I error and power rates of two new machine learning based person-fit metrics to the performance of other person-fit metrics. The performance of the person-fit metrics in this study was not adequate, which is in contrast to previously published studies. To better understand this discrepancy, we observationally examined differences in the simulation designs between the current study and previous studies to identify factors that may have contributed to the differences in the findings. We observed differences in classification accuracy, test length, the number of attributes assessed, base rates of attribute mastery, and the method for selecting items for the imposition of person-misfit.

Evaluating the Performance of Person-Fit Detection Methods in Diagnostic Classification Models

Adequate model fit is required for making high quality inferences from assessment results (Han & Johnson, 2019). The psychometric model links the observed responses to estimates of student proficiency. Consequently, model fit is directly related to the model's accuracy in estimating students' proficiency. Practically, adequate model fit increases our confidence in the interpretations of students' proficiency in the assessed constructs based on the assessment results.

Model fit can be assessed at the test-, item- and person-level (Han & Johnson, 2019). Model fit quantifies the consistency between the observed data and the model predicted values at each of these levels (Gu, 2011). Test-level model fit quantifies whether the model fits the data obtained across all items (Sinharay & Almond, 2007). Item-level model fit quantifies whether the model fits the data obtained for each item (Han & Johnson, 2019). Person-level model fit (i.e., person-fit) is the consistency between a student's responses and the model predicted values for that student (Gu, 2011). Person-fit statistics provide a statistical indicator for the aberrance of each student's observed responses by comparing the observed responses to the expected responses.

Providing evidence of adequate test-level model fit is necessary to support the intended uses and interpretations of the assessment results (Han & Johnson, 2019). Item-level model fit evidence also plays a role in supporting the intended uses and interpretations of assessment results. Namely, adequate item-level model fit across all items will translate into adequate test-level model fit, and inadequate item-level model fit indicates specific areas for improvement

that can be used to improve the test-level model fit (Han & Johnson, 2019). Thus, adequate test- and item-level model fit are critical to supporting the intended uses and interpretations for an assessment (Chen et al., 2013).

Person-fit also has implications for the intended uses and interpretations of assessment results in terms of the inferences made for each student (Walker, 2017). For example, evidence of adequate person-fit for a student increases confidence in student-level inferences made from the assessment for that student (e.g., Cui & Roberts, 2013). Conversely, evidence of person-misfit for a student decreases confidence in student-level inferences made from the assessment for that student, even if there is adequate evidence of test- and item-level model fit. Thus, evaluating person-fit evidence is important in supporting the intended uses and inferences of assessment results, and evidence of person-fit should be evaluated in addition to evidence of test- and item-level model fit.

Although machine learning has not been previously used to evaluate person-fit, we can predict whether students demonstrate evidence of person-misfit with machine learning models. Machine learning is an increasingly popular topic, and machine learning is continuously being applied to new areas because of its wide range of applicability (e.g., Kucak et al., 2018). For example, Zhai et al. (2020a, 2020b) published reviews of how machine learning is used in science assessment. More specific to evaluating person-fit evidence, machine learning models should be able to identify students with person-misfit by identifying patterns in the data if the models include predictor variables that are pertinent to person-fit. Thus, machine learning models are a potential method for evaluating person-fit.

To date, person-fit has been an underexplored topic within the context of diagnostic classification models (DCMs; Rupp et al., 2010). Works defining person-fit statistics have been published (e.g., Cui & Leighton, 2009; Cui & Li, 2015), but more work is needed in exploring person-misfit detection for DCMs. Given the role of person-fit evidence in supporting the intended uses and inferences of assessment results, it is important to develop a better understanding of person-misfit detection in DCMs.

In this study, we evaluate the performance of person-fit metrics to detect person-misfit for DCMs. We first present an overview of DCMs, person-fit detection methods for DCMs, and machine learning models that we can use to assess person-fit in DCMs. Then, we present the methods and results for the simulation study evaluating the performance of person-fit metrics to detect person-misfit for DCMs. Finally, we conclude this study with a discussion of the findings and their implications for using machine learning models to detect person-fit in assessments scaled with DCMs.

Diagnostic Classification Models

DCMs (Bradshaw, 2016; Rupp et al., 2010) are growing in popularity for applied and operational uses. Rather than a continuous scale score, DCMs categorically estimate students' proficiency in the assessed latent traits (Rupp et al., 2010). DCMs also estimate the probability that each respondent has mastered each assessed trait and the probability a respondent in each latent class would provide a correct response.

Model Fit in Diagnostic Classification Models

Because DCMs conceptualize the assessed latent traits as discrete rather than continuous, model fit methods for item response theory (IRT) and structural equation modeling

(SEM) that assume a continuous latent trait may be inappropriate for DCMs. For example, IRT and SEM indices for absolute test-level model fit include χ^2, Root-Mean-Square-Error-of-Approximation, Goodness-of-Fit, Adjusted GFI, Root Mean Square Residual, Standardized Root Mean Square Residual, Normed-Fit Index, Comparative Fit Index, and Parsimony Fit Indices (Hooper et al., 2008). However, they are inappropriate to use with DCMs because these indices assume a continuous latent trait (Maydeu-Olivares & Joe, 2005, 2014) and DCMs estimate a categorical latent trait (Rupp et al., 2010). Thus, alternative measures for assessing person-fit in DCMs are needed.

Person-Fit in Diagnostic Classification Models

Although person-fit has been relatively understudied in DCMs relative to the amount of work on test- and item-level fit (e.g., Chen et al., 2013; Sorrel et al., 2017), four person-fit statistics for DCMs have been defined (Cui & Leighton, 2009; Cui & Li, 2015; Liu et al., 2009). Additionally, posterior predictive model checking (PPMC) can also be used to evaluate person-fit in Bayesian DCMs. PPMC provide a statistical indicator for the aberrance of each student's observed responses by comparing the observed responses to responses simulated from the parameter values at each iteration of the posterior distribution.

Person-Fit Statistics

Only four person-fit statistics for DCMs have been defined in the literature: a likelihood ratio test (Liu et al., 2009), the Hierarchy Consistency Index (*HCI*; Cui & Leighton, 2009), the Response Conformity Index (*RCI*; Cui & Li, 2015), and a modified l_z statistic (Cui & Li, 2015).

Likelihood Ratio Test. Liu et al. (2009) used a model-based approach to assess person-fit in DCMs by evaluating whether a DCM that includes a parameter for aberrant responses fits the data better than a DCM without the aberrant response parameter. The underlying assumption of Liu et al. (2009) is that DCMs likely do not fit data with aberrant responses unless an aberrant response parameter is included. Thus, to improve model fit when aberrance is present, Liu et al. (2009) suggest adding a parameter, $\boldsymbol{\rho}$, to represent the probability of a student providing an aberrant response.

After estimating the DCM and the DCM with the aberrant response parameter, Liu et al. (2009) suggested using the likelihood ratio test based on the marginal likelihood to assess person-fit. The likelihood ratio test examines whether a DCM that allows for the probability of a student providing an aberrant response fits the data better than a DCM without the aberrant response parameter while accounting for the added aberrant response parameter.

Hierarchical Consistency Index. The HCI statistic (Cui & Leighton, 2009) was designed to be used with assessments using a hierarchical Q-matrix (e.g., Liu et al., 2017); however, the HCI statistic can be used when an attribute hierarchy is not present. When a student responds correctly to an item measuring an attribute, the HCI statistic assumes that this student should also respond correctly to all the other items measuring the same attribute. Similarly, when a student responds correctly to an item measuring two attributes, the HCI statistic assumes that this student should also respond correctly to all the other items measuring either of those attributes. For example, when a student responds correct to an item measuring Attribute 1 and Attribute 2, the HCI statistic assumes this student should respond correctly to all items

measuring both Attribute 1 and Attribute 2, all items measuring only Attribute 1, and all items measuring only Attribute 2.

Cui and Leighton (2009) defined the *HCI* statistic as

$$HCI = 1 - \frac{2\sum_{i\epsilon S_{cj}}\sum_{h\epsilon S_i} X_{ji}\left(1 - X_{jh}\right)}{\mathrm{N}_j} \quad (3)$$

where S_{cj} is the set of items measuring the attribute(s) that were answered correctly by student *j*, S_i is the set of items measuring any of the attributes measured by item *i*, N_j is the number of comparisons made for student *j*, X_{ji} is the dichotomously scored item response to item *i* for student *j*, and X_{jh} is the dichotomously scored item response to item *h* for student *j*. The *HCI* statistic can range from -1 to 1, with values near 1 indicating better person-fit and values near -1 indicating worse person-fit.

Conceptually, the *HCI* statistic is a function the number of inconsistencies (i.e., the number of times a student answers one item correctly while answering another item that measures any of the assessed attributes incorrectly). The maximum number of inconsistencies results in $HCI = -1$. The minimum number of inconsistencies results in $HCI = 1$. For example, suppose a student completes three items assessing a single attribute, and this student responds correctly to the first two items (i.e., Item 1 and Item 2) and incorrectly to the final item (i.e., Item 3). In this example, the item response for Item 1 is then compared to the item response for Item 2 and Item 3. Similarly, the item response for Item 2 is then compared to the item response for Item 1 and Item 3. Since N_j is the number of comparisons, $N_j = 4$ in this example. When the compared item responses are both correct responses (e.g., Item 1 and Item 2), $X_{ji}\left(1 - X_{jh}\right)$ reduces to $1(1-1) = 1(0) = 0$. In contrast, when the compared item responses

include one correct response and one incorrect response (e.g., Item 1 and Item 3), $X_{ji}(1 - X_{jh})$ reduces to $1(1 - 0) = 1(1) = 1$. For this example, $HCI = 1 - \frac{2(0+1+0+1)}{4}$, which reduces to $1 - \frac{4}{4} = 0$. Thus, the two inconsistent item comparisons (Item 1 compared to Item 3 and Item 2 compared to Item 3) contributed person-misfit, while the consistent item comparisons between Item 1 and Item 2 did not.

Response Conformity Index. The RCI statistic (Cui & Li, 2015) measures the consistency between a student's observed responses and the expected responses given the student's estimated attribute mastery profile and the conditional probabilities of responding correctly given attribute mastery. In other words, masters are expected to respond correctly and non-masters are expected to respond incorrectly at rates specified by the estimated item parameters. While similar to the HCI statistic, the RCI statistic compares the observed and expected responses across all items, rather than just the items that the student responded to correctly.

Cui and Li (2015) defined the *RCI* statistic as

$$RCI = \sum_{i=1}^{I} |\, RCI_{ji}| = \sum_{i=1}^{I} |\ln(-\frac{X_{ji} - \pi_{ji}}{I_{ji} - \pi_{ji}})^{X_{ji}+I_{ji}}| \qquad (4)$$

where π_{ji} is the probability of student *j* providing a correct response to item *i* given the student's mastery classification, I_{ji} is the dichotomously scored expected response from student *j* to item *i* given the student's mastery classification, and X_{ji} is the dichotomously scored observed item response to item *i* for student *j*. I_{ji} takes a value of 1 when the student has mastered all the assessed attributes, and I_{ji} takes a value of 0 when the student has not

mastered all the assessed attributes. The *RCI* statistic ranges from 0 to infinity, with values near 0 indicating better person-fit.

Conceptually, the *RCI* statistic is measuring the unexpectedness of inconsistent responses. Put simply, consistent responses do not contribute to person-misfit as quantified by the *RCI* statistic. When the response is inconsistent (e.g., $X_{ji} \neq I_{ji}$), $ln\left(-\frac{X_{ji}-\pi_{ji}}{I_{ji}-\pi_{ji}}\right)$ reduces to $ln\left(\frac{1-\pi_{ji}}{\pi_{ji}}\right)$ when $X_{ji} = 1$ and $I_{ji} = 0$ and to $ln\left(\frac{\pi_{ji}}{1-\pi_{ji}}\right)$ when $X_{ji} = 0$ and $I_{ji} = 1$. Hence, the degree of unexpectedness is purely a function of the conditional probability of providing a correct response given attribute mastery status. For example, suppose a master responds incorrectly (i.e., $I_{ji} = 1$ and $X_{ji} = 0$) to a relatively easy item for masters (e.g., $\pi_{ji} = .8$). In this example, the unexpectedness of this response would be $\frac{.8}{1-.8} = \frac{.8}{.2} = 4$ and $ln(4) \approx 1.39$. In contrast, now suppose a master responds incorrectly (i.e., $I_{ji} = 1$ and $X_{ji} = 0$) to a relatively more difficult item for masters (e.g., $\pi_{ji} = .55$). For this relatively more difficult item, the unexpectedness of this response would be $\frac{.55}{1-.55} = \frac{.55}{.45} \approx 1.22$ and $ln(1.22) \approx 0.20$. As can be seen from this example, the unexpectedness of the incorrect response from a master to a more difficult item is lower, reflecting how the *RCI* statistic is a function of the conditional probability of responding correctly.

l_z Statistic. Cui and Li (2015) extend the $\boldsymbol{l_0}$ and $\boldsymbol{l_z}$ statistics from the IRT context to DCMs. In contrast to the form of the Levine and Rubin (1979) $\boldsymbol{l_0}$ statistic, Cui and Li (2015) use natural logarithm properties to define $\boldsymbol{l_0}$ as

$$l_0 = \ln \prod_{i=1}^{I} [\, \pi_{ji}^{X_{ji}} \left(1 - \pi_{ji}\right)^{1-X_{ji}}] \qquad (5)$$

where X_{ji} is a dichotomously scored response to item *i* by student *j* and π_{ji} is the conditional probability of student *j* providing a correct response to item *i* given the student's estimated attribute mastery status. The expected value of l_0 is determined by the number of items a student took and their corresponding conditional probabilities. The expected value can be calculated using

$$E_{l_0} = \sum_{i=1}^{I} [\pi_{ji} \ln(\pi_{ji}) + (1 - \pi_{ji}) \ln(1 - \pi_{ji})] \quad (6)$$

and the variance of l_0 can be calculated using

$$Var_{l_0} = \sum_{i=1}^{I} \pi_{ji} (1 - \pi_{ji}) \ln(\frac{\pi_{ji}}{1 - \pi_{ji}}) \quad (7)$$

The observed l_0 statistic is then compared to the expected value and variance by calculating a *z*-score. This is the l_z statistic, which follows a standard normal distribution, with numbers closer to zero indicating better person-fit. Theoretically derived cut points of -1.96 and 1.96 can be set for detecting aberrant responses with $\alpha = .05$, using the properties of a standard normal distribution.

Posterior Predictive Model Checking

PPMCs are a method for evaluating absolute model fit in Bayesian models (Gelman et al., 2013). PPMCs simulate data based on the parameter values from each iteration of the posterior distribution and compare the simulated data to the observed data. The rationale is that if the model has good fit, data simulated from the posterior distribution will be similar to the observed data. Thus, the similarity of the observed data relative to the simulated data is indicative of how well the model fit the data.

To evaluate the similarity between the observed and simulated data, a discrepancy statistic is selected to quantify a pertinent aspect of the data, and the discrepancy statistic for the observed data is compared to the discrepancy statistic from each iteration of the posterior distribution. The effectiveness of PPMC relies on selecting a discrepancy statistic that captures features of the data that are reflective of model fit.

The PPMC in this study used two discrepancy statistics. The first PPMC used the difference between the proportion of correct responses and the expected percent correct given estimated attribute mastery status as the discrepancy statistic. For brevity, we will refer to this discrepancy statistic as the raw difference. The equation for the raw difference PPMC is provided in Equation 8

$$diff = \hat{\pi} - \pi_0 \quad (8)$$

where $\hat{\pi}$ is the proportion of correct responses for a student and π_0 is the expected percent correct given estimated attribute mastery status.

The second PPMC used the number of correct responses as the discrepancy statistic. This discrepancy statistic simply calculated the number of correct responses for each student.

As previously mentioned, the similarity between the observed and simulated data are indicative of model fit. For example, there is adequate person-fit based on the second PPMC when the observed number of correct responses falls towards the center of the distribution of the number of correct responses from the data simulated using the posterior distribution parameter values. The results of PPMC are often quantified as posterior predictive *p* values (*ppp* values), where the *ppp* value is the percentile for the observed discrepancy statistic in the

distribution of the simulated discrepancy statistics. Cut-points of .025 and .975 are used in this study to flag person-misfit when *ppp* values are in the tails of the distribution.

Machine Learning

Machine learning is an exploratory statistical modeling approach that identifies trends within the data to better predict the outcome variable (Hoover, 2022). Machine learning models can be either supervised or unsupervised (e.g., Jordan & Mitchell, 2015). In supervised machine learning models, the outcome variable is known, which allows for optimizing the model for predicting the outcome variable (Bell, 2015). In unsupervised machine learning models, the models do not make use of the outcome variable in optimizing the model (Jain et al., 1999). In the case of the person-fit, students' true person-fit statuses are latent variables, meaning that we cannot be certain whether a student is truly an aberrant responder or not. Because of this, unsupervised machine learning models are necessary for identifying students with person-misfit. Consequently, we focused on unsupervised methods in this topic guide.

Machine learning models can be used to assess person-fit, even though little work has been done in this area. To date, only Zhu et al. (2022) has conducted work using machine learning models to assess person-fit in DCMs. Notably, however, the Zhu et al. study makes multiple assumptions and design choices that limit its generalizability. More specifically, Zhu et al. train a supervised neural network using simulated data to assess person-fit. Consequently, the generalizability of this approach is dependent on the ability to adequately simulate data that is realistic to operational data. To highlight such difficulties, the occurrence of person-misfit is a complex process that can occur in a variety of forms and be affected by a wide range of factors. As such, it may be incredibly difficult to simulate person-misfit with a generating

model that adequately maps on to how person-misfit occurs in an operational setting. Along similar lines, researchers have previously discussed the difficulty of accurately simulating data to reflect realistic cheating behaviors (e.g., Cui & Li, 2015; Meijer & Sijtsma, 1995). Thus, additional work is needed to establish a generalizable machine learning based approach for assessing person-fit in DCMs without making the machine learning models dependent on the quality of a data simulation process.

The majority of unsupervised machine learning models use clustering algorithms that organize data into similar groups (e.g., Alloghani et al., 2020), although non-clustering unsupervised algorithms such as principle components analysis exist. Clustering algorithms use predictor variables to group similar students together. Common clustering unsupervised machine learning models include *k*-means clustering, hierarchical clustering, and latent class analysis (LCA). We used three machine learning approaches in this study: *k*-means clustering, LCA, and boosted LCA.

K-Means Clustering. We conducted *k*-means clustering with two latent classes (i.e., $\boldsymbol{k = 2}$) to flag students with person-misfit. Two latent classes were chosen based on the definition of the outcome variable (i.e., person-fit), where one class represents non-aberrant responders (i.e., those with adequate person-fit) and the other class represents aberrant responders (i.e., those with person-misfit).

We used five predictor variables in the *k*-means clustering model. These predictor variables included the three person-fit statistics (*HCI*, *RCI*, l_z) and the *ppp* values for the two PPMC.

In *k*-means clustering, class membership is determined by placing each student into the nearest of the *k* clusters. The clusters are based on the values of the five predictor variables. Practically, this means that each student is placed into the cluster where the predictor variables for the student are most similar to the values for the cluster.

The cluster labels were added *post hoc* in this *k*-means clustering model. The cluster containing fewer students was labelled as the aberrant responders class based on the definition of aberrant responding where aberrance implies a deviation from the usual response pattern. Thus, the majority of students are expected to be non-aberrant responders, which supports labeling the smaller of the two clusters as aberrant responders.

Latent Class Analysis. We conducted LCA with two latent classes to flag students with person-misfit. The two latent classes were again chosen because of the definition of person-fit, where one class represents aberrant responders and the other class represents non-aberrant responders.

We used five predictor variables in the LCA. These predictor variables included the three person-fit statistics (*HCI*, *RCI*, l_z) and the *ppp* values for the two PPMC.

Similar to DCMs, latent class analysis produces probability estimates that each student is a member of each class. Class membership can then be determined by the largest probability of class membership. This means that students were assigned to the class with the probability estimate greater than .50. For example, a student with a .55 probability of being a member of class 1 and a .45 probability of being a member of class 2 would be classified as a member of class 1, since .55 is greater than .50.

The latent class labels were added *post hoc* in this exploratory LCA. The latent class containing fewer students was labelled as the aberrant responders class based on the definition of aberrant responding where aberrance implies a deviation from the usual response pattern. Thus, the majority of students are expected to be non-aberrant responders, which supports labeling the smaller of the two classes as aberrant responders.

Boosted Latent Class Analysis. Machine learning models can quickly become complex, and this complexity can have implications for the computational time and requirements of the model (Al-Jarrah et al., 2015). Researchers have created boosted machine learning algorithms, which consist of many simpler models, to circumvent these concerns (Schapire, 2003). The simpler models use fewer predictor variables, which reduces their complexity and subsequently the computational demands (Roe et al., 2020). However, to make accurate predictions, multiple simpler models need to be estimated (Schapire, 2003). The underlying rationale is that estimating many simple models may be more efficient than estimating a single, incredibly complex model. Thus, boosted LCA entails estimating multiple LCA models, where each LCA model uses relatively few predictor variables, and all of the LCA models contribute to the classification from the boosted LCA.

We used boosted LCA with two latent classes to flag students with person-misfit, where each boosted LCA consisted of 10 LCA. Many of the modeling approaches for the boosted LCA were similar to those for the LCA, since boosted LCA consists of multiple simpler LCA models. The two latent classes were again chosen because of the definition of person-fit, where one class represents aberrant responders and the other class represents non-aberrant responders.

For each LCA model within the boosted LCA model, one, two, or three predictor variables were randomly selected from the available predictor variables. The possible predictor variables for the boosted LCA included the three person-fit statistics (*HCI*, *RCI*, l_z) and the *ppp* values for the two PPMC.

Within each LCA model in the boosted LCA model, class membership was determined by the largest probability of class membership, meaning students were assigned to the class with the probability estimate greater than .50. As with the approach for LCA, the latent class containing fewer students in each LCA of the boosted LCA model was labelled as the aberrant responders class. Since each LCA produced a classification for each student, each student was classified 10 times within the boosted LCA. Each of the 10 LCA models classified each of the students as having adequate or person-misfit. When six or more of the classifications indicated that a student demonstrated evidence of person-misfit, the student was classified as having person-misfit by the boosted LCA. Otherwise, the student was classified as having adequate person-fit by the boosted latent class analysis. The threshold of six was chosen so that the level of evidence required to classify students as showing evidence of person-misfit was high, with the hope that this would avoid elevated Type I error rates.

Evaluating the Machine Learning Models. To compare the performance of the k-means clustering, LCA, and boosted LCA models to the person-fit statistics and the PPMC, we evaluated the performance of the three machine learning models using the Type I error and power rates within each condition.

Objectives

In this study, we report the results of a simulation study designed to evaluate the performance of the two new machine learning based metrics to detect person-misfit for DCMs relative to the HCI, RCI, l_z, and PPMC metrics.

Methods

In this study, we manipulated three factors: the number of assessed attributes, the minimum number of items per attribute, and the proportion of students with misfit. The levels for the number of assessed attributes were two and three. The levels for the minimum number of items per attribute were three and 10. The levels for the proportion of students with misfit were 0%, 10%, and 20%. This resulted in 12 total conditions in this simulation. We simulated 100 repetitions simulated for each condition.

Data Simulation

In each repetition, we simulated 1,000 students. We chose to simulate 1,000 students based on the recommendation of Sen and Cohen (2020), who recommended sample sizes of at least 1,000 to obtain precise parameter estimates from DCMs.

In each condition, the test length (*T*) was the product of the minimum number of items per attribute (*I*) and the number of attributes (*A*), with $T = I$ x A. The first half of the items in each condition formed an identify matrix. The second half of items also formed an identity matrix, but each item had a 50% chance of measuring a second attribute. Items were constrained to measure no more than two attributes. Thus, each attribute is measured by at least the minimum number items per attribute.

Dichotomous attribute mastery status for each student was determined using the base rate, between-attribute correlation, and a random number. The base rate for mastering each attribute was .50, the between-attribute correlation was drawn from $U(.00, .80)$, and a random number was drawn from $U(0,1)$. Base rate determines the proportion of the students that have mastered the given attribute. The between-attribute correlation determines the strength of relationship between the attributes, with higher between-attribute correlations indicating mastery of one attribute is more likely to be associated with mastery of the other attribute. We then aggregated each student's dichotomous attribute mastery statuses to form an attribute mastery profile.

We randomly drew the item intercept parameters from $U(-2.2, -1.4)$, which corresponds to $U(.1, .2)$ on the probability scale. For items measuring a single attribute, we randomly drew the item main effect parameters from $N(3,0.25)$. For items measuring two attributes, we randomly drew the item main effect parameters from $N(3,0.5)$. We constrained all of the item main effect parameters to be positive to uphold the assumption of monotonicity. For items measuring two attributes, we randomly drew the item interaction parameters from $N(1,0.17)$. We constrained the item interaction parameters to be greater than negative one times the small of the two item main effect parameters to ensure that masters of both assessed attributes have a higher conditional probability of responding correctly than masters of only one of the assessed attributes. We chose the distributions for the item main effect and interaction parameters based on the simulations conducted by Johnson and Sinharay (2018).

Imposing Person-Misfit

Students were randomly assigned to have person-misfit in accordance with the proportion of misfit present. Students were randomly assigned to have one of four types of misfit. Each type of misfit was equally likely. The types of misfit were:

- **Creative responders**. This type of misfit is defined as true masters who interpret the item in a non-standard manner, leading to an incorrect response. More specifically, creative responders respond incorrectly to all items measuring the first attribute even if they were masters of this attribute.
- **Random responders**. With this type of person-misfit, students respond at random to all the completed items with a 25% probability of responding correctly. We chose this 25% probability of responding correctly based on the probability of randomly guessing the correct answer to a multiple choice item with four response options.
- **Sleepers**. In this condition, students responded to the first 33% of items incorrectly as a representation of situations where students may miss the initial items on an assessment due to anxiety rather than due to a lack of proficiency. The first 33% of items were chosen because the minimum number of items per attribute in this study is three, and imposing misfit on the first 33% of items would affect at least one item for each attribute.
- **Fatigued responders**. Students who were "fatigued" are those who responded to the last 33% of items incorrectly as a representation of situations where students may miss the last items on an assessment due to mental fatigue from the assessment rather than due to a lack of proficiency. The last 33% of items were again chosen because the

minimum number of items per attribute in this study is three, so imposing misfit on the last 33% of items would affect at least one item for each attribute.

Consider a student who completed six items with no misfit. The hypothetical Q-matrix for these six items is presented in Table 1. This student might have a simulated response pattern across items of [0, 1, 1, 0, 1, 0]. When this student is assigned to have person-misfit, this response pattern would result in the following response patterns:

- creative: [0, **0**, **0**, 0, 1, 0]
- sleeping: [0, **0**, 1, 0, 1, 0]
- fatigue: [0, 1, 1, 0, **0**, 0]

For students with creative, sleeping, or fatigue patterns, the bolded responses indicate where an incorrect response was imposed due to person-misfit. The random response pattern was excluded from this example because the random response pattern is less prescriptive.

Table 1

Hypothetical Q-Matrix

Attribute 1	Attribute 2
1	0
0	1
1	0
0	1
1	1
1	1

Person-misfit can be operationalized as spuriously high (i.e., students overperform given their true attribute mastery profiles) or spuriously low scores (i.e., students underperform given their true attribute mastery profiles). Although spuriously high scores are possible in practice, the types of person-misfit imposed in this study exclusively produced spuriously low scores

(e.g., Liu et al., 2009). Flagging person-misfit for spuriously high scores was not emphasized in this study because there are at least two types of mechanisms leading to spuriously high scores. The first type of mechanism is behaviors or circumstances that lead to improved performance through randomness and factors associated with the assessment. Examples of this type of mechanism might include a student correctly guessing item responses at an unexpectedly high rate (e.g., a random responder with a higher base rate of responding correctly) or a student being cued to the correct response for one item on the assessment based on another item on the assessment. These sorts of spuriously high person-misfit should conceivably be flagged similarly to the spuriously low person-misfit examined in this study given the relationship between how the types of person-misfit are imposed. The second type of mechanism is behaviors or circumstances that lead to improved performance contingent upon external factors. One example of this would be a student receiving assistance from a teacher or peer. As pointed out by Meijer and Sijtsma (1995) and reiterated by Cui and Leighton (2009), this type of spuriously high person-misfit is problematic for simulations because the level of aberrance in the resulting responses depends on the frequency and quality of these behaviors, which is not easily introduced into a simulation study. Consequently, this type of person-misfit was not the focus of this study. Thus, this study focused exclusively on spuriously low person-misfit.

Model Estimation

For each simulated data set, we estimated a log-linear cognitive diagnosis model (LCDM; Henson et al., 2009). We chose the LCDM because it is a general DCM that subsumes many other subtypes of DCMs, and thus supports the generalizability of our findings. From each model, the item parameters (i.e., the conditional probabilities of masters and non-masters

providing a correct response) and the estimated mastery probability for each student were used to estimate the person-fit metrics. In this study, we dichotomized the estimated mastery probabilities for each student using a cut-point of .50.

Flagging Person-Misfit

In this study, we flagged students for person-misfit using the *HCI*, *RCI*, and l_z person-fit statistics but not the likelihood ratio test. Liu et al. (2009) found that the likelihood ratio test did not detect spuriously low scores as well as it detected spuriously high scores. Thus, we did not use the likelihood ratio test to assess person-fit in this study.

The *HCI*, *RCI*, and l_z person-fit statistics were calculated for each student in each condition. While the l_z statistic follows a standard normal distribution, which allows for statistically flagging students demonstrating person-misfit, the *HCI* and *RCI* statistics do not follow a known distribution. Thus, cut-points for flagging person-misfit with the *HCI* and *RCI* statistics have to be determined for use within this study.

Using the process reported by Cui and Li (2015) and Cui and Leighton (2009), who followed established approaches for determining cut-points for person-fit statistics (e.g., Seo & Weiss, 2013; van Krimpen-Stoop & Meijer, 2002), we included a condition with no person-misfit to ascertain the empirical distributions of the *HCI* and *RCI* statistics when there was adequate person-fit for all students. The *HCI* cut-points were determined for each response pattern, and the *RCI* cut-point was determined for each attribute mastery profile. These cut-points were determined independently for the conditions based on the number of measured attributes and the minimum number of items per attribute. By knowing the empirical distributions of the *HCI* and *RCI* statistics when no person-misfit was present, we can identify the cut-points for flagging

the most extreme five percent of the *HCI* and *RCI* statistics for detecting person-misfit. More specifically, in each repetition of the condition with no person-misfit, we identified the cut-point for the most extreme five percent of the *HCI* and *RCI* statistics, and we then averaged those repetition-specific thresholds across the 100 repetitions within the condition to establish our *HCI* and *RCI* cut-points for the other conditions with the same number of measured attributes and minimum number of items per attribute in this study.

For the l_z person-fit statistic, thresholds of -1.96 and 1.96 were used to flag students showing evidence of an aberrant response pattern, since the l_z statistic has a standard normal distribution.

For the PPMC, students are flagged for person-misfit when their *ppp* values are less than .025 or greater than .975. The cut-points are designed to flag students with observed discrepancy statistics that fall in either tail of the distribution of discrepancy statistics from the simulated data.

For the *k*-means clustering and LCA models, students are flagged for person-misfit when they are assigned to the smaller of the two latent classes. As previously described, two latent classes (aberrant responders and non-aberrant responders) were used based on the definition of person-fit, and the smaller latent class was labelled as aberrant responders since aberrance implies a deviation from the usual response pattern.

For the boosted LCA, students are flagged for person-misfit when they are labelled as aberrant responders by six or more of the LCAs in the boosted LCA. Similar to the *k*-means clustering and LCA models, two latent classes were used for each LCA in the boosted LCA, and the smaller latent class was labelled as aberrant responders.

In this study, we calculated Type I error and power rates for each person-fit index within each condition. The Type I error rate is the proportion of students flagged for person-misfit when the students were not randomly assigned to have person-misfit. The power rate is the proportion of students flagged for person-misfit when they were randomly assigned to have person-misfit.

Follow-Up Analyses

We calculated the classification accuracy of the estimated LCDMs for those with and without imposed person-level misfit as follow-up analyses for this study.

Results

We estimated an LCDM and applied the person-fit detection methods for each of the 1,200 simulated repetitions across the 12 conditions. The Type I error and power rates for the person-fit detection methods are presented by condition in Table 2 and Table 3, respectively. The performance of the person-fit detection methods was suboptimal. The Type I error rates for many of the person-fit detection methods were elevated (*HCI*, number correct PPMC, LCA, *k*-means clustering), slightly elevated for the *RCI* statistic, and well controlled for the l_z statistic and the distance PPMC. However, the power rate followed a similar pattern. For example, the Type I error rate was well controlled for the l_z statistic, but the power rate for the l_z statistic was less than .20 in each condition, which indicates a pattern of underflagging. The observed power rates were too low for operational use for all of the person-fit detection, with the exception of the power of *k*-means clustering in some of the conditions measuring three attributes with a minimum of 10 items per attribute. However, the Type I error of the *k*-means clustering method was extremely elevated for these conditions. Thus, none of the applied

person-fit detection methods demonstrated acceptable Type I error and power rates. Potential reasons for poor performance are described in the Discussion section.

We present the attribute- and class-level classification accuracy by condition in Table 4. For students with good person-fit, the attribute-level classification accuracy was high with estimates ranging from .91 to .97, and the class-level classification accuracy was also high with estimates ranging from .77 to .94. For students with person-misfit, the attribute- and class-level classification accuracy was much lower with attribute-level classification accuracy ranging from .73 to .78 and class-level classification accuracy ranging from .48 to .62.

The follow-up analyses indicated that a significant proportion of students were misclassified when person-misfit was present. For example, the attribute-level classification accuracy dropped by approximately .20 when person-misfit was present compared to when students had good person-fit. Similarly, the class-level classification accuracy dropped by approximately .30 when person-misfit was present compared to when students had good person-fit. The classification accuracy when students had good person-fit was similar to classification accuracy estimates reported by Shan and Wang (2020). It is possible that poor classification accuracy obscured evidence of person-misfit, which subsequently led to poor person-fit detection.

Table 2

Type I Error Rates, by Condition

Attributes	Minimum items per attribute	Proportion misfit	HCI	RCI	l_z	Number correct PPMC	Distance PPMC	LCA	K-means clustering	Boosted LCA
2	3	0	.39	.07	.02	.23	.02	.04	.18	.18
2	3	10	.42	.07	.02	.22	.02	.05	.17	.00
2	3	20	.43	.07	.02	.22	.02	.05	.18	.18
2	10	0	.52	.07	.05	.35	.04	.39	.07	.00
2	10	10	.50	.06	.04	.35	.04	.36	.06	.12
2	10	20	.50	.06	.04	.35	.04	.35	.07	.07
3	3	0	.41	.08	.02	.29	.02	.14	.09	.19
3	3	10	.42	.08	.02	.29	.02	.14	.09	.09
3	3	20	.43	.07	.02	.29	.02	.13	.09	.13
3	10	0	.48	.08	.04	.40	.04	.64	.02	.21
3	10	10	.37	.07	.04	.40	.04	.63	.02	.26
3	10	20	.37	.06	.03	.39	.04	.61	.02	.21

Table 3

Power Rates, by Condition

Attributes	Minimum items per attribute	Proportion misfit	HCI	RCI	l_z	Number correct PPMC	Distance PPMC	LCA	K-means clustering	Boosted LCA
2	3	10	.41	.10	.05	.04	.03	.07	.24	.00
2	3	20	.40	.10	.05	.03	.04	.08	.24	.25
2	10	10	.50	.14	.14	.10	.07	.49	.10	.20
2	10	20	.50	.12	.13	.10	.06	.47	.11	.11
3	3	10	.42	.14	.06	.07	.04	.22	.13	.13
3	3	20	.42	.11	.06	.07	.03	.20	.13	.19
3	10	10	.34	.18	.16	.13	.08	.70	.04	.40
3	10	20	.33	.15	.14	.13	.08	.69	.04	.31

Table 4

Attribute- and Class-Level Classification Accuracy, by Condition

				Students with good person-fit		Students with person-misfit	
Condition	Attributes	Minimum items per attribute	Proportion misfit	Attribute classification accuracy	Class classification accuracy	Attribute classification accuracy	Class classification accuracy
1	2	3	0	.91	.84	---	---
2	2	3	10	.91	.83	.73	.57
3	2	3	20	.91	.83	.74	.58
4	2	10	0	.97	.94	---	---
5	2	10	10	.97	.94	.76	.60
6	2	10	20	.97	.94	.77	.62
7	3	3	0	.92	.77	---	---
8	3	3	10	.91	.77	.75	.48
9	3	3	20	.92	.77	.75	.49
10	3	10	0	.97	.92	---	---
11	3	10	10	.97	.91	.78	.54
12	3	10	20	.97	.91	.78	.54
Total	---	---	---	.94	.86	.76	.55

Discussion

In this study, we compared the performance of the new machine learning based person-fit metrics to performance of the HCI, RCI, l_z, and PPMC person-fit metrics. Generally, the performance of the person-fit indices was not adequate. While the Type I error rates were controlled in some cases, none of the indices demonstrated adequate power. The observed power rates were below previously reported power rates for detecting person-misfit (e.g., Cui & Li, 2015).

The follow-up analysis indicated imposing person-misfit had consequences for classification accuracy. When there was adequate person-fit, the classification accuracy was consistent with previous studies (e.g., Shan & Wang, 2020). However, the classification accuracy was considerably lower when person-misfit was present. This suggests that the poor classification accuracy may have led to downstream consequences for evaluating person-fit. The *RCI* statistic, the l_z statistic, and the raw difference PPMC incorporate the conditional probability of providing a correct response given estimated mastery status, so the values for these statistics change when the estimated mastery status changes. It is possible that a given response pattern would lead a master to be flagged for person-misfit but a non-master would not be flagged. Thus, the imposed person-fit in this study that led to spuriously low scores may have led true masters to be misclassified as non-masters who were not flagged for person-misfit.

The performance of the machine learning models was likely affected as a downstream consequence of the performance of the other person-fit indices. Ultimately, machine learning models need predictors that are related to the outcome variable, and stronger relationships

usually lead to better performance. Because the person-fit statistics and the PPMC were not effective in flagging person-misfit, these indices were ineffective as predictors in the machine learning models. Consequently, the suboptimal performance of the machine learning models is unsurprising, given the poor performance of the other person-fit detection methods. However, the performance of the machine learning based person-fit metrics is not entirely dependent on the performance of the other person-fit detection methods. It is possible that improved but still inadequate performance of the other person-fit detection methods could lead to adequate performance for the machine learning based person-fit metrics.

Understanding Poor Person-Fit Detection

To better understand our findings, we examined differences between the current study and previous studies to identify factors that may facilitate person-misfit detection. We identified multiple differences in the simulation designs (and hence operational contexts) that may affect person-misfit detection. We observationally identified these differences by comparing the simulation designs between the current study and previous studies.

Classification Accuracy

The existing person-fit statistics assume high classification accuracy. More specifically, the *RCI* and l_z person-fit statistics incorporate information from the mastery status when evaluating person-fit, meaning that changes in estimated mastery status affect the value of the person-fit statistic. This can be detrimental to the accurate detection of person-misfit. Thus, effective person-misfit detection with the existing person-fit statistics may rely on the students being classified accurately even in the presence of person-misfit.

Test Length

We found test length (or the minimum number of items per attribute as it was conceptualized in this study) to be related to person-misfit detection. As an example, Cui and Li (2015) used conditions with 20 and 40 items to measure three and six attributes in their simulation study. They found power increased within each type of person-misfit as test length increased. In our study, we used shorter test lengths. For example, consider our two-attribute condition with a minimum of three items per attribute (six total items). When sleeping person-misfit was imposed on the first two items, students only had four items remaining to demonstrate attribute mastery. For students who were simulated as true masters of the attributes, the remaining four items may have been insufficient to demonstrate attribute mastery. Thus, the imposed person-misfit may have led to these masters being classified as non-masters, and the observed response patterns for these true masters may not be aberrant for a non-master.

Number of Attributes Assessed

We also found the number of assessed attributes to be related to detecting person-misfit. By increasing the number of multi-attribute items, students who are masters of all the assessed attributes can use the multi-attribute items to demonstrate attribute mastery. For example, suppose person-misfit is only imposed on single-attribute items (e.g., creative misfit). By including a significant number of multi-attribute items, students who have mastered multiple attributes can demonstrate attribute mastery by responding correctly to these multi-attribute items.

In our study, we included fewer multi-attribute items than Cui and Li (2015). For example, in the conditions measuring two attributes with a minimum of 10 items per attribute in our study, there were 15 single-attribute items and five two-attribute items on average. With creative misfit, there were only five multi-attribute items on average that would allow for demonstrating mastery of Attribute 1. Thus, there were limited opportunities for masters of Attribute 1 and Attribute 2 to demonstrate mastery of Attribute 1, and there were no opportunities for masters of only Attribute 1 to demonstrate mastery of Attribute 1.

Base Rate of Attribute Mastery

We found the base rate of attribute mastery to be indirectly related to effectively detecting person-misfit. The base rate of attribute mastery determines the number of students who have mastered each attribute and ultimately the number of students who have mastered multiple attributes. As previously mentioned, students who have mastered multiple attributes have additional opportunities to demonstrate attribute mastery in the presence of person-misfit.

To demonstrate, Cui and Li (2015) used a .80 base rate of attribute mastery in their simulation. This implies approximately 90% of their simulated students in their simulations measuring three attributes had mastered multiple attributes. In our study, we simulated students with a .50 base rate of attribute mastery. For the conditions measuring two attributes, this implies only 25% of our simulated students had mastered multiple attributes. For the conditions measuring three attributes, this implies only half of our simulated students had mastered multiple attributes.

Item Selection for Imposing Person-Misfit

As one final note for assessment (and simulation) design as it pertains to person-fit detection, the effectiveness of person-misfit detection depends on which items are selected for misfit to be imposed. For example, in both the current study and previous studies, the single attribute items tended to be the first items on the assessment. As a result, imposing sleeping misfit tended to impose person-misfit on single-attribute items, and imposing fatigue misfit tended to impose person-misfit on multi-attribute items. Systematically imposing misfit on a specific type of item in this manner can have downstream consequences for demonstrating attribute mastery by affecting students' opportunities to demonstrate attribute mastery.

The Combination of These Factors

The intersection of these observations leads to questions about the effectiveness of the existing person-fit statistics to identify person-misfit. Although we did not set out to identify issues with flagging person-misfit using the existing person-fit statistics, understanding our suboptimal results has highlighted factors that appear to have contributed to success in previous studies while contributing to the suboptimal results in our study. Previous studies appear to have been conducted to imitate ideal estimation conditions. However, the current study imitated reasonable but less than ideal estimation conditions, and we were generally unable to identify students with person-misfit.

Our *post hoc* observations raise questions pertaining to how robust the existing person-fit statistics are to reasonable but less than ideal estimation circumstances. For example, previous studies have used relatively large numbers of items to measure multiple attributes with many multi-attribute items, and these studies have tended to simulate data using high

base rates of attribute mastery. The cumulative effect of each of these design choices may be significantly contributing to accurate person-misfit detection. For instance, high base rates of attribute mastery allow for the majority of students to be masters of multiple attributes and consequently have a higher probability of responding correctly to multi-attribute items, which supports high classification accuracy. However, would lower base rates of attribute mastery be problematic for accurately detecting person-misfit even if the all the other factors were held constant? It seems possible that only adjusting one of these factors may not affect person-misfit detection, but our results appear to indicate that adjusting a sufficient number of these factors negatively impacted person-misfit detection.

Future Research

At a minimum, an area for future research pertains to the robustness of the existing DCM person-fit statistics. More specifically, are these person-fit statistics able to accurately identify person-misfit under suboptimal but realistic estimation conditions? For example, base rates of attribute mastery approaching .80 may lead to a sizable portion of students mastering multiple attributes, but lower base rates may be common in practice, especially in formative or through year assessment where additional learning is expected. Given the findings of the current study, it remains to be shown at what point the existing person-fit statistics become ineffective at detecting person-misfit.

Another area for future research includes detecting person-misfit when the imposed person-misfit leads to a change in the estimated mastery statuses. As previously mentioned, many of the simulated conditions in previous studies appear to have imposed person-misfit without altering the estimated mastery classifications. It is possible, however, for person-misfit

to lead to a change in mastery classification where the changed mastery status may better align with the observed responses. Because some existing person-fit statistics incorporate the conditional probability of a correct response given a student's estimated mastery status, such changes are fundamental to the identification of students with person-misfit. Additional methods and/or person-fit statistics may be needed to overcome this dependency.

Conclusion

The person-fit detection methods used in this study were not particularly successful in accurately identifying students with person-fit. The methods evaluated in this study demonstrated elevated Type I errors and suboptimally low power. Given the poor performance of the person-fit statistics and PPMC, it is unsurprising that the machine learning approaches were not able to accurately identify students with person-misfit. It is possible that machine learning approaches for person-fit may be effective if better predictors of person-fit can be established. To better understand our results, we noted discrepancies between our study and previous studies examining person-misfit detection in DCM-based assessment, which allowed us to identify areas for future research that may improve person-fit evaluation in DCM-based assessments.

References

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, *2*(3), 87–93. https://doi.org/10.1016/j.bdr.2015.04.001

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. In M. Berry, A. Mohamed, & B. Yap (Eds.), *Supervised and unsupervised learning for data science* (pp. 264–323). Springer, Cham. https://doi.org/10.1007/978-3-030-22475-2_1

Bell, J. (2015). *Machine learning: Hands-on for developers and technical professionals*. John Wiley & Sons, Inc.

Bradshaw, L. (2016). Diagnostic classification models. In André A. Rupp & J. P. Leighton (Eds.), *Handbook of Cognition and Assessment* (pp. 297–327). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch13

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute evaluation in cognitive diagnosis momdeling: Relative and absolute fit evaluation in CDM. *Journal of Educational Measurement*, *50*(2), 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*(4), 429–449. https://doi.org/10.1111/j.1745-3984.2009.00091.x

Cui, Y., & Li, J. (2015). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement*, *39*(3), 223–238. https://doi.org/10.1177/0146621614557272

Cui, Y., & Roberts, M. R. (2013). Validating student score inferences with person-fit statistic and verbal reports: A person-fit study for cognitive diagnosis assessment. *Educational Measurement: Issues and Practice*, *32*(1), 34–42. https://doi.org/10.1111/emip.12003

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman; Hall/CRC.

Gu, Z. (2011). *Maximizing the potential of multiple-choice items for cognitive diagnostic assessment* [PhD thesis]. University of Toronto.

Han, Z., & Johnson, M. S. (2019). Global- and item-level model fit indices. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models*. Springer Nature. https://doi.org/10.1007/978-3-030-05584-4_17

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210. https://doi.org/10.1007/S11336-008-9089-5

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structure equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Hoover, J. C. (2022). Machine learning. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (2nd ed., pp. 842–847). SAGE Publications. https://doi.org/10.4135/9781071812082.n314

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for diagnostic assessments. *Journal of Educational Measurement*, *55*(4), 635–664. https://doi.org/10.1111/jedm.12196

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kucak, D., Jurici, V., & Dambic, G. (2018). Machine learning in education – a survey of current research trends [Conference paper]. Proceedings of the 29th DAAAM International Symposium, Vienna, Austria. https://doi.org/10.2507/29th.daaam.proceedings.059

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269–290. https://doi.org/10.3102/10769986004004269

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement*, *77*(2), 220–240. https://doi.org/10.1177/0013164416645636

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*(8), 579–598. https://doi.org/https://doi.org/10.1177/0146621609331960

Maydeu-Olivares, A. and Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n^ contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

Maydeu-Olivares, A. and Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

Meijer, Rob R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8*, 261–272. https://doi.org/10.1207/s15324818ame0803_5

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8*, 261–272. https://doi.org/10.1207/s15324818ame0803_5

Roe, K. D., Jawa, V., Zhang, X., Chute, C. G., Epstein, J. A., Matelsky, J., Shpitser, I., & Taylor, C. O. (2020). Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance. *PLoS ONE*, *15*(4), e0231300. https://doi.org/10.1371/journal.pone.0231300

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear estimation and classification: Lecture notes in statistics* (pp. 149–171). Springer.

Sen, S., & Cohen, A. S. (2020). Sample size requirements for applying diagnostic classification models. *Frontiers in Psychology*, *11*, 621251. https://doi.org/10.3389/fpsyg.2020.621251

Seo, D. G., & Weiss, D. J. (2013). Lz person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement*, *73*(6), 994–1016. https://doi.org/https://doi.org/10.1177/0013164413497015

Shan, N., & Wang, X. (2020). Cognitive diagnosis modeling incorporating item-level missing data mechanism. *Frontiers in Psychology*, *11*, 3231. https://doi.org/https://doi.org/10.3389/fpsyg.2020.564707

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnosis models. *Educational and Psychological Measurement*, *67*(2), 239–257. https://doi.org/10.1177/0013164406292025

Sorrel, M. A., Abad, F. J., Olea, J., Torre, J. de la, & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis models. *Applied Psychological Measurement*, *41*(8), 614–631. https://doi.org/https://doi.org/10.1177/0146621617707510

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement*, *26*(2), 164–180. https://doi.org/https://doi.org/10.1177/01421602026002004

Walker, A. A. (2017). Why education practitioners and stakeholders should care about person-fit in educational assessments. *Harvard Educational Review*, *87*(3), 426–443. https://doi.org/10.1111/jedm.12196

Zawacki-Richter, O., Marin, V. I., Bond, M., & Gouverneur. F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*(39). https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020a). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430-1459. https://doi.org/10.1002/tea.21658

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020b). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111-151. https://doi.org/10.1080/03057267.2020.1735757

Zhu, Z., Arthur, D., & Chang, H.-H. (2022). A new person-fit method based on machine learning in CDM in education. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 616–637. https://doi.org/10.1111/bmsp.12270