Appraising Traditional and Purpose-built Person Fit Statistics’ Power to Detect Cheating

Sanford R. Student
University of Colorado, Boulder

Follow this and additional works at: https://www.ce-jeme.org/journal

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
Available at: https://www.ce-jeme.org/journal/vol3/iss1/3

This Article is brought to you for free and open access by Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊. It has been accepted for inclusion in Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊 by an authorized editor of Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊.
Appraising Traditional and Purpose-built Person Fit Statistics’ Power to Detect Cheating

Cover Page Footnote
I thank Drs. Derek C. Briggs and Benjamin R. Shear (University of Colorado Boulder) for their thoughtful, helpful feedback on this study and Dr. Sandip Sinharay (Educational Testing Service) for useful discussion early in the development of this research.
Appraising Traditional and Purpose-Built Person Fit Statistics’ Power to Detect Cheating

Sanford Student
University of Colorado Boulder

Abstract
Person-fit statistics (PFSs) have been suggested as a tool to detect cheating in large-scale testing, and this study investigates their potential for this application. Most PFSs are equally sensitive to scores that appear spuriously high or spuriously low. Xia and Zheng introduced four PFSs that are meant to be more sensitive to spuriously high scores and therefore may be more appropriate for detecting cheating. Comparing the power of these weighted PFSs against the power of traditional PFSs to detect cheating shows that there is no single best statistic in all or most scenarios, and in most scenarios, most examinees flagged as cheating by person fit analysis did not cheat. Implications for operational use of PFSs to detect cheating are discussed.

Keywords
Person fit; Person fit statistics; Test security; Aberrant responding

1 Introduction
Person-fit statistics (PFSs) are considered important for detecting the presence of test takers whose response patterns are aberrant relative to most examinees (Karabatsos, 2003; Sinharay, 2017). PFSs generally fall into two categories. Parametric PFSs depend upon the results of calibrating a measurement model, typically an Item Response Theory (IRT Embretson & Reise, 2000) model. Nonparametric statistics do not use the results of a measurement model, instead estimating a statistic of interest directly from raw response data. An example of a popular parametric fit statistic is \( l_z \) (Drasgow et al., 1985), which expresses the standardized log-likelihood of a given respondent’s item response pattern relative to estimated IRT parameters. A well-known nonparametric PFS is \( H^T \) (Sijtsma, 1986), which expresses the sum of the covariances between a given examinee’s responses and all other examinees’ responses, divided by the maximum possible sum of those covariances (Linacre, 2012). The interpretation of PFSs varies; \( l_z \) is interpreted relative to a standard normal distribution, with negative values indicating person misfit and positive values indicating overfit, while \( H^T \) falls between -1 and 1, with lower values indicating the deviance of an individual’s responses from those of other respondents, but with no specific distribution from which to derive critical values.

No single PFS has achieved widespread adoption, and there appears to be no consensus on which statistic is best suited to the detection of aberrant response patterns, or even whether the use of such statistics is appropriate at all (Meijer & Tendeiro, 2012; Wainer, 2012). To gain more clarity on which PFS(s) to use in the detection of aberrant responses, Karabatsos (2003) used receiver operating curves (ROCs) to compare the power of 36 PFSs to detect aberrant responding in datasets simulating the responses of examinees to test items. The datasets vary by length, prevalence of aberrant responding, and nature of aberrant responding. While Karabatsos found that the nonparametric \( H^T \) statistic (Sijtsma, 1986) was the most powerful for detecting aberrance, Sinharay (2017) pointed out methodological issues with the calculation of power that skewed Karabatsos’s findings in favor \( H^T \). Sinharay (2017) instead concludes that three other statistics—the parametric \( l_z \) (Drasgow et al., 1985) and \( ECI_{4z} \) (Tatsuoka, 1984) statistics, and the nonparametric \( U_3 \) (van der Flier, 1980)—are equally as powerful as \( H^T \) for detecting aberrant responses.

Both studies investigate PFSs’ power to detect aberrance of two distinct natures: spuriously high scores and spuriously low scores. An examinee’s score is spurious if it is much higher or much lower than the score associated with
that examinee’s “true” ability relative to the tested content. Notably, this can never be observed outside of a simulation, as true ability is always unknown in the real world, irrespective of the measurement model applied (or not). The studies in Karabatsos (2003) and Sinharay (2017) are based upon simulations of a variety of hypothesized sources of aberrance, only some of which produce spuriously high scores. In the context of high-stakes standardized testing, these scores are of greater interest than spuriously low scores for security purposes. The possible use of PFSs in high-stakes standardized testing has been suggested in both the person-fit literature (Sinharay, 2016a), albeit tentatively, and with more enthusiasm in operational test security handbooks (Olson & Fremer, 2013; Wollack & Fremer, 2013). In this context, it is not clear that general-purpose PFSs are best suited to the task of identifying examinees who have cheated.

Theoretically, a PFS that is more sensitive to spuriously high scores than spuriously low scores may be better suited to the detection of cheating, as it should flag more respondents whose deviation from the expected response pattern matches the assumption that cheating will produce higher scores. To this end, the present study considers not only the four PFSs outlined in Sinharay (2017), but also four PFSs derived specifically for the purpose of detecting scores that are spuriously high (Xia & Zheng, 2018). These statistics are all members of the family of PFSs outlined in Snijders (2001) to which $l_1$ and $ECI_4^*$ also belong, as they express a (standardized) weighted sum of the squared residuals of a respondent’s observed item responses given their estimated ability and the estimated item parameters. Notably, $l_1$, $ECI_4^*$, and all other previously conceived PFSs in this family use symmetric weighting functions, meaning that whether a score appears spuriously high or spuriously low, the same amount of overall deviance from the expected response pattern will produce the same value for the statistic. An asymmetric weighting function can produce a statistic that will yield values indicating greater misfit for response patterns that appear to produce spuriously high scores than for response patterns that produce apparently spuriously low ones. To this end, Xia and Zheng introduce $SHa(\lambda)^*$ and $SHb(\beta)^*$, two families of fit statistics that more heavily weight correct responses to items that the respondent had little chance of getting correct given the IRT model parameters. Here, $\lambda$ and $\beta$ are parameters to the two different statistics; in their study, Xia and Zheng consider four total statistics, two from each family. These are $SHa(1/2)^*$, $SHa(1)^*$, $SHb(2)^*$, and $SHb(3)^*$. The specifics of these statistics are available in the original paper.

This study investigates the power of these asymmetrically-weighted PFSs, as well as four traditional, symmetrically-weighted PFSs, to detect cheating in two parts. First, I replicate the simulations performed in Sinharay (2017), then compare the overall average power of the weighted PFSs to detect aberrant examinees compared with the power of four traditional PFSs to do so. These traditional PFSs are $H^T$, $l_1^*$ (Snijders, 2001), $U_3$ and $ECI_4^*$ (Sinharay, 2016a). The second set of simulations is conceptually similar to the crossed design used by Karabatsos (2003) and by Sinharay (2017), but with several departures intended to bring the simulations closer to realistic conditions that might be found on a high-stakes test. These changes are outlined below, but the most notable update is to simulate cheating alongside other types of aberrant responses instead of in isolation. This supports investigations into whether PFSs can detect cheating specifically alongside other aberrant test-taking behaviors, as cheating is of greater interest in a test security context than these other types. Additionally, this study goes beyond relative power to consider the prospect of using any of the investigated statistics for operational test security purposes. Here, the question is not which statistic is most powerful, but rather: are the statistics “good enough” at flagging cheaters? To investigate this question, I leverage positive predictive value (PPV).

First, I review the notion of aberrant responding and why cheating differs from other types of aberrant response in the context of high-stakes standardized testing. Next, I review the simulations performed in Karabatsos (2003) and Sinharay (2017), as I both replicate them in this study and base the new simulations in this study on their work. After reviewing these simulations, I explain the new simulations performed for this study, noting both how they differ from Karabatsos’s and Sinharay’s, and why these departures produce simulations more relevant to real-world educational testing scenarios. Then, I briefly review the methodology used to compare different PFSs’ power to detect cheating and investigate how they might perform in an operational setting. I report the results of both sets of simulations, then a discussion section considers the implications of these results for the use of PFSs in test security programs.

$^{1}l_1^*$ and $ECI_4^*$ are asymptotically-corrected versions of the corresponding statistics used in Sinharay’s investigation of PFSs’ power to detect aberrance.
2 Background

2.1 What Is Aberrant Responding?

On an exam, an aberrant response pattern is one that produces a spuriously high or low score relative to a given examinee’s true ability (Karabatsos, 2003). Essentially, an aberrant response pattern results when the process by which an examinee arrives at their item responses does not rely upon that examinee’s ability, but the existence of such response patterns in real data is always hypothetical, as true ability is never known. Thus, investigating such response patterns almost always requires simulation and a theory about what aberrant responding looks like. Karabatsos hypothesizes five sources of aberrant response patterns: cheating, creative responding, random responding, lucky guessing, and responding carelessly. Two of these sources, cheating and lucky guessing, are expected to produce spuriously high scores relative to examinees’ true ability. In contrast, careless and creative responding would produce spuriously low scores. The implications of random guessing hinge on students’ true ability: a student who would typically struggle mightily on a given test might achieve a spuriously high score by responding at random, while a student of high ability would likely produce a spuriously low score via the same strategy. This study pays particular interest to cheating, which is more clearly relevant to policymaking and test security than the other sources of aberrance.

2.2 Why Is Detecting Cheating of Particular Interest?

Test security is often described as a validity issue (Wollack & Fremer, 2013, Xia & Zheng, 2018), and rightly so; if examinees’ scores do not reflect their ability relative to the targeted construct, any subsequent action based on test scores may be targeting the wrong people. This is a particularly important concern when tests have high stakes, such as professional licensure or college admissions. Just as test security is partially a matter of validity, so too; if examinees stand to benefit from high scores. The behaviors modeled in Karabatsos (2003) are all plausible. However, these other behaviors are not considered a test security concern or a reason to invalidate examinees’ scores. If person fit analysis is to be used in test security, the statistic used should be able to differentiate between cheating examinees and examinees whose responses are different from those predicted by the measurement model, but not indicative of a test security issue.

3 Methods

3.1 Replication: Karabatsos (2003)’s Simulations

The logic of leveraging PFSs as a security measure implicitly relies on two assumptions. The first is that there is some possibility, even a very small one, that test security was unknowingly breached, allowing some examinees to gain an advantage unrelated to their true ability to answer the questions on the test. The second is that non-cheating examinees will respond in line with the measurement model used for scoring, so that departures from this model can be considered evidence of possible cheating. However, cheating is far from the only source of aberrant response patterns on tests, even ones with very high stakes where all examinees stand to benefit from high scores. The behaviors modeled in Karabatsos (2003) are all plausible. However, these other behaviors are not considered a test security concern or a reason to invalidate examinees’ scores. If person fit analysis is to be used in test security, the statistic used should be able to differentiate between cheating examinees and examinees whose responses are different from those predicted by the measurement model, but not indicative of a test security issue.

\[
P_{n/j} = \frac{1}{1 + \exp [\delta_j - \theta_n]} \tag{1}
\]

where \(\theta_n\) and \(\delta_j\) are on a common scale (this is not the original parameterization of the model, but is presented...
here for its concise form). Simulation of item responses from this model is straightforward once one has defined the distribution of the ability and difficulty parameters.

Next, five types of aberrance are defined. As noted above, these are: cheating, creative responding, random responding, lucky guessing, and responding carelessly. For each type of aberrance, one must define how the calculation of \( P_{nj1} \) differs from (1).

For cheating respondents, abilities are uniformly distributed from -2 to -0.5, and the probability of a correct response is constrained to 1 for items with difficulty \( \delta_j \geq -1.5 \). This represents a scenario in which less able respondents gain access to correct answers for the hardest items on an exam. The equation to calculate \( P_{nj1} \) for a cheating examinee is therefore

\[
P_{nj1} = I_{[1.5, 2]}(\delta_j) + I_{[-2, -1.49]}(\delta_j) \frac{1}{1 + \exp[\delta_j - \theta_n]} \tag{2}
\]

where \( I_{[x, y]}(z) \) is an indicator variable taking on value 1 if \( x \leq z \leq y \) and 0 otherwise.

For respondents responding creatively, abilities are uniformly distributed from 0.5 to 2, and probability of correct response is constrained to 0 for items with difficulty \( \delta_j \leq -1.5 \). This appears to represent a situation in which highly able respondents read additional meaning into very easy items and respond in a way that does not match the intent of the item. The equation to calculate \( P_{nj1} \) for an examinee responding creatively is therefore

\[
P_{nj1} = I_{[-1.49, 2]}(\delta_j) \frac{1}{1 + \exp[\delta_j - \theta_n]} \tag{3}
\]

This equation produces a value of 0 for any item with \( \delta_j \leq -1.5 \).

For random responding, every item was assigned a probability of a correct response of 0.25, mirroring the process of purely guessing on a multiple-choice item with four options. The equation to calculate \( P_{nj1} \) for an examinee responding at random is therefore

\[
P_{nj1} = 0.25 \tag{4}
\]

For respondents exhibiting so-called “lucky” guessing, abilities are uniformly distributed from -2 to -0.5, and probability of correct response is constrained to 0.25 for items with difficulty \( \delta_j \geq 0.5 \). The equation to calculate \( P_{nj1} \) for a guessing examinee is therefore

\[
P_{nj1} = 0.25 I_{[5.2]}(\delta_j) + I_{[-2, -4.9]}(\delta_j) \frac{1}{1 + \exp[\delta_j - \theta_n]} \tag{5}
\]

For respondents responding carelessly, abilities are uniformly distributed from 0.5 to 2, and probability of correct response is constrained to 0.5 for items with difficulty \( \delta_j \leq -0.5 \). This represents a situation in which more able examinees rush through or misread easy items, reducing their likelihood of responding correctly to those items. The equation to calculate \( P_{nj1} \) for a careless examinee is therefore

\[
P_{nj1} = 0.5 I_{[-2, -5]}(\delta_j) + I_{[-4.9, 2]}(\delta_j) \frac{1}{1 + \exp[\delta_j - \theta_n]} \tag{6}
\]

Once the various types of response have been defined, the next step is to define the different test lengths and proportions of aberrant responding for the crossed dataset simulation design. Here, the three test lengths are 17, 33 and 65 items, and the four proportions of aberrant respondents are 0.05, 0.10, 0.25 and 0.50. Sinharay (2017) and Karabatsos (2003) differ in the size of each dataset: Karabatsos uses a size of 500 while Sinharay uses 10,000. The use of 10,000 examinees per dataset greatly reduces the standard error of subsequent calculations of power, and also aligns more closely with the expected sample size for a large scale test. Here I use 10,000 as well. From here, one can then simulate datasets as follows:

For every test length \( l \), aberrance type \( t \), and proportion of aberrance \( p \):  

1. Assign “true” item difficulties (\( \delta_j \)) for all items (of which there are \( l \)) and “true” abilities (\( \theta_n \)) for all non-aberrant examinees (of whom there are \( 10,000^* (1-p) \)). Both difficulties and abilities are uniformly distributed between -2 and 2.

2. Assign “true” abilities (\( \theta_n \)) for all aberrant examinees (of whom there are \( 10,000^* p \)). The distribution of these abilities depends on \( t \), as described above.

3. Use the Rasch Model (1) to simulate the item responses of all non-aberrant examinees to all items by treating every response of person \( n \) to item \( j \) as a Bernoulli-distributed random variable taking on value 1 with probability \( P_{nj1} \) and 0 otherwise.
4. Use one of Equations (2)–(6) to simulate the item responses of all aberrant examinees to all items by treating every response of person \( n \) to item \( j \) as a Bernoulli-distributed random variable taking on value 1 with probability \( P_{nj1} \) and 0 otherwise.

At this point, one has constructed 60 separate datasets (3 lengths \( 5 \) proportions of aberrant responding \( 5 \) types of aberrant responding) of 10,000 examinees’ item responses. The first set of simulations in this study replicate this approach exactly.

3.2 New Simulations

The second set of simulations in this study modifies the methodology of the first in several key ways, summarized in Table 1 and described in more detail below.

The number of examinees in each dataset was 10,000, as in Sinharay (2017), which produces very low standard errors on the metric used to assess each PFS’s power to detect cheating (discussed below) and more closely resembles the sample sizes found in typical large scale assessment contexts. I again used the Rasch model to simulate non-aberrant responses.

The equations for all types of aberrance except cheating [(3)-(6)] remain identical to those in the original studies. I used a different approach to simulate the effect of cheating on item responses. Karabatsos’s definition of cheating is, “[c]heating (e.g., copying from another examinee) refers to behavior where the examinee unfairly obtains the correct answers on test items that he/she is unable to answer correctly” (Karabatsos, 2003, p. 278). This describes a scenario in which a group of examinees gain unauthorized access to test items before the test is administered, such as from an examinee who has already taken the test. When this happens, there is no reason to believe that they will specifically gain access to only the very hardest items on the test. Instead of simulating cheating by guaranteeing that cheating students answer the most difficult items correctly, I simulated cheating by selecting a random subset of the harder 50% of test items and marking them as compromised—that is, an item \( j \) can only be compromised if \( \delta_j > 0 \). Cheating students were guaranteed to get these items right, but some hard items were not compromised, producing Rasch-generated responses to these items from cheating examinees. Let \( \text{comp} \) represent the subset of items that have been marked compromised. The equation used to calculate a cheating examinee’s responses is

\[
P_{nj1} = I_{j(\text{comp})} + I_{j(\text{non-comp})} \frac{1}{1 + \exp[\delta_j - \theta_n]} \tag{7}
\]

where \( I_{j(\text{comp})} \) indicates that item \( j \) was compromised and \( I_{j(\text{non-comp})} \) indicates that it was not.

These simulations differ from the prior studies in how aberrant respondents are selected from the larger group of examinees. In prior studies, the proportion of aberrant respondents was used to “set aside” some examinees from the larger population, and then the abilities of the aberrant and non-aberrant examinees were simulated from separate uniform distributions. In the present study, I begin by simulating the abilities of all 10,000 examinees from a standard normal distribution. I simulate aberrance after the ability distribution for all 10,000 examinees has been simulated, which guarantees that the overall distribution of ability among all examinees remains normal and is not distorted by the proportion of aberrant examinees. This is of particular concern for the datasets in the original studies where 25% or more of responses are simulated as aberrant.

To simulate cheating, with a proportion of aberrance \( p \), I select 10,000 \( p \) respondents by selecting at random among all examinees with “true” ability below -0.5. This mirrors the original studies’ choice of an upper limit on the distribution of ability among cheaters, though their lower bound of -2 does not make sense to implement when ability follows a standard normal distribution. As a result, these simulations also follow the assumption that high-ability examinees do not cheat. This is a limitation that future studies should address. A similar logic underlies the approaches to simulating aberrance for the remaining types. Creative responders are selected from examinees with ability above 0.5, guessers are selected from examinees with ability below -0.5, careless responders are selected from examinees with ability above 0.5, and random respondents are selected from examinees with ability below -0.5.

The crossed design of the simulated datasets in this study also differs. I do not consider the scenario in which fully half of all responses are aberrant, instead considering the scenario where only 1% are; the notion that fully half of the examinees taking a high-stakes test would respond aberrantly struck me as less likely than a very small proportion doing so. Additionally, the prior studies simulated each type of aberrance in isolation. As the present focus is on cheating, the only type of aberrance
Table 1
Crossed Designs in Karabatsos (2003)/Sinharay (2017) and Current Study

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Original studies</th>
<th>Current study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of total aberrance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• 0.05</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>• 0.10</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>• 0.25</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>• 0.50</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>Test length</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>Aberrance type</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>• Cheating</td>
<td>Cheating only</td>
<td></td>
</tr>
<tr>
<td>• Creative responding</td>
<td>Cheating + creative</td>
<td></td>
</tr>
<tr>
<td>• Random responding</td>
<td>Cheating + random</td>
<td></td>
</tr>
<tr>
<td>• Lucky guessing</td>
<td>Cheating + guessing</td>
<td></td>
</tr>
<tr>
<td>• Responding carelessly</td>
<td>Cheating + careless</td>
<td></td>
</tr>
</tbody>
</table>

simulated in isolation is cheating. Instead of simulating the other types of aberrance in isolation, I simulate cheating alongside one of each of the other four aberrance types. In all scenarios, the proportion of total aberrance is divided evenly among the aberrance types in the scenario. Table 1 summarizes both crossed designs. After simulating the responses of all examinees in each dataset, the next step is to generate estimated item difficulty parameters and examinee ability estimates. As in Sinharay (2017), I used the R (R Core Team, 2020) package ltm (Rizopoulos, 2006) for item difficulty estimation via marginal maximum likelihood and estimated examinees’ ability via maximum likelihood using the irtoys package (Partchev, 2017).

3.3 Selection of PFSs for Comparison

Karabatsos’s 2003 original study compared the power of 36 PFSs, while Sinharay’s 2017 reconsideration of Karabatsos’s work considered four of these 36: $H_T$, $U_3$, $l_z$, and $ECI4_z$. These comprise both parametric and nonparametric statistics that have, in studies other than Karabatsos’s, been shown to have similar power. Here, I consider eight statistics. The first two are $H_T$ and $U_3$, as in Sinharay (2017). However, rather than use $l_z$ and $ECI4_z$, I use $l_z^*$ (Snijders, 2001) and $\zeta_2^*$ (Sinharay, 2016a). Because $\zeta_2^*$ is an asymptotic correction of $ECI4_z$, for clarity, I instead refer to $\zeta_2^*$ as $ECI4_z^*$. In both cases, the * indicates the asymptotically normal corrected version of the statistic, derived using Snijders 2001’s asymptotic correction. Prior work has demonstrated that the uncorrected $l_z$ has less power to detect aberrance than anticipated when its theoretical null distribution is derived using true ability parameters (Meijer & Sijtsma, 2001), and Sinharay (2016a) demonstrated that $ECI4_z^*$ is generally superior in power to detect aberrance to its uncorrected counterpart. This study therefore investigates whether the asymptotic corrections of $l_z$ and $ECI4_z$ provide any meaningful improvement in power to detect aberrance compared to their uncorrected counterparts. Snijders (2001) notes that the asymptotic corrections he outlines produce the least benefit for the Rasch model compared to more complex IRT models such as the two- or three-parameter logistic models, but the corrected statistics should perform at least as well as their uncorrected counterparts when the Rasch model is employed.

In addition to $H_T$, $U_3$, $l_z$, and $ECI4_z$, I also consider four PFSs that were proposed specifically to detect spuriously high scores. These are Xia and Zheng (2018)’s $SHA(\lambda)^*$ and $SHb(\beta)^*$, each with two different values of their respective parameter as Xia & Zheng suggest, producing the four separate PFSs: $SHA(1/2)^*$, $SHA(1)^*$, $SHb(2)^*$ and $SHb(3)^*$. In all cases, the statistics follow an asymptotically standard normal distribution, where high values of the index indicate response patterns producing scores suspected of being spuriously high (it is important to remember that PFSs, and IRT as a whole, cannot ever tell us true ability levels directly, and so cannot ever tell us with certainty if
a score is truly “spurious.”). As described above, all four of these PFSs are conceptually similar to $l_z^2$ and $ECI4^z$, but with an asymmetric weighting function (see Magis et al., 2012 for a didactic explanation of the role of weighting functions in the calculation of $l_z^2$); these statistics are all part of the family of PFSs described in Snijders (2001).

Because these indices’ power to detect spuriously high scores should be greater than their power to detect low scores, they appear to be better suited to the purpose of employing PFSs in test security. This practice is described in several guides to test security (Kim et al., 2017; Olson & Fremer, 2013; Wollack & Fremer, 2013) but its validity is still debated due to the lack of research on the connection between PFSs and proposed sources of aberrance (Meijer & Tendeiro, 2012; Wainer, 2012). If weighted PFSs outperform their traditional counterparts, then this helps narrow in on how one might use PFSs most productively in test security. However, in investigating these statistics’ use in detecting simulated cheating, Xia and Zheng (2018) also found inflated Type I error rates under certain conditions. Type I errors are concerning for two reasons. First, if PFSs are to be used as even one method of many to detect cheating by examinees, Type I errors are crucial, as they equate to a false accusation of cheating. At its worst, an inflated Type I error rate means many false positives relative to each true positive, calling into question the efficiency of using PFSs in test security. Second, on the test development side, the use of PFSs to initiate investigations into compromised items assumes that the PFSs provide at least some reason to be concerned that items are compromised. If PFSs tell us little or nothing about which examinees responded aberrantly, then using them for any purpose related to test security appears ill-advised.

I calculated $H^z$, $U_3$, and $l_z^2$ using the PerFit R package (Tendeiro et al., 2016); $SHA(1/2)^z$, $SHA(1)^z$, $SHb(2)^z$ and $SHb(3)^z$ using R code based upon Magis et al. (2012) as recommended in Xia and Zheng (2018); and $ECI4^z$ using code written by the author in accordance with Sinharay (2016a). Data management and figure creation was conducted using the tidyverse (Wickham et al., 2019). Key portions of the R code used in this study are available in the appendix to this study. I also refer readers to Sinharay (2017), which contains a link to a Github repository containing much of the code used in that study.

### 3.4 Comparing PFSs With AUROCs and PPV

Receiver operating characteristic (ROC) curves (Hanley & McNeil, 1982) can be used to compare continuous statistics intended to dichotomously classify subjects by comparing the true positive and false positive rates produced when different values of the statistic are used as a cutoff to classify subjects. An example ROC for a relatively powerful statistic can be found in Figure 1. A statistic with no power would produce an ROC that resembles a straight diagonal line where true and false positive rate are equal.

**Figure 1**

*Example ROC curve*

Interpretation and comparison of ROCs is based upon the area under the ROC (in keeping with Sinharay [2017]. I refer to this as AUROC; it is also known as AUC or $a$). A very powerful statistic will produce an AUROC near 1, as the true positive rate will be nearly 1 at almost all values of the false positive rate. In contrast, a statistic with low power will produce an AUROC near, or even below, 0.5. An AUROC of 0.5 indicates zero accuracy—the statistic is no better at flagging true positives than flagging false ones. An AUROC below 0.5 indicates that the statistic does have power if classifications are inverted. Sinharay (2017) concluded that the four PFSs considered in his study were approximately equally powerful for detecting aberrance by calculating AUROCs for each statistic for each of the 60 simulated datasets and finding that the AUROCs of the various statistics were generally very close when summarized over different test lengths. Moreover, the overall AUROCs reported by Sinharay (2017) are close to one another and fairly high (0.83-0.84 for all four statistics).

This study also uses AUROCs to compare the power of PFSs to detect specific examinee behavior. As described above, this study contains two sets of simulations. First, I replicate Karabatsos (2003)/Sinharay (2017)’s methodology, then I conduct new simulations in which cheating is mixed with other simulated behaviors. This requires two different sets of AUROC analyses. For the
replication, I use Sinharay's (2017) methodology exactly, meaning that AUROCs are used to express how well different statistics can distinguish among aberrant and non-aberrant response patterns. For the simulations with mixed aberrance types, instead of using each PFS to classify examinees as aberrant or non-aberrant, I use each PFS to classify examinees as cheating or non-cheating and base AUROCs on that instead. Here, a true positive occurs when the PFS flags an examinee whose responses were simulated as cheating from equation (7), while a false positive occurs when the PFS flags as cheating an examinee whose responses were produced by any other type of response, including other types of aberrance. For all AUROC-based analysis, I used the ROCR package (Sing et al., 2005).

To compare the power of the eight PFSs in both sets of simulations, I use several aggregations of the individual AUROCs, of which there are sixty. First, I compare the overall “average” AUROC for each PFS. As Sinharay (2017) notes, a meta-analytic method such as the Dersimonian-Laird algorithm (DerSimonian & Laird, 1986) will produce the best estimates of the average AUROC, since each AUROC is subject to measurement error and has a standard error. I used the Dersimonian-Laird algorithm as well, as implemented in the metafor R package (Viechtbauer, 2010). Next, I compare the average AUROCs across each of the dimensions making up the crossed design of the study: over each test length, over each type of aberrance, and over each proportion of aberrance. This provides some insight into whether certain PFSs exhibit higher power in certain scenarios. For all three dimensions, I use the Dersimonian-Laird algorithm to produce an average AUROC for each value of the conditions—for example, average AUROCs for each statistic for 17-item tests, 33-item tests, and 65-item tests.

For length only, for both the replication and new simulations, I also produce an additional comparison, following Sinharay (2017)’s corrected approach to aggregating over all tests of a given length to produce a single AUROC. The approach is as follows, and is identical to that used by Sinharay (2017):

1. For a given test length, combine all 10,000-response datasets into a single dataset of 200,000 examinees. Keep track of who is aberrant/cheating and who is not.

2. Estimate a single set of item difficulty parameters from the combined dataset using marginal maximum likelihood.

3. Estimate a single set of examinee ability parameters from the combined dataset using maximum likelihood.

4. Compute each PFS for all examinees in the combined dataset.

5. Calculate AUROCs for each PFS based upon the values in (4).

This approach produces a single AUROC per PFS that corresponds to the power of that PFS to detect aberrance/cheating when many types of aberrance are present at once. For each test length, there are 20 datasets of that test length, within which every type of aberrance is present. The combined 200,000-examinee dataset therefore contains every type of aberrance at once. The power of each statistic to detect cheating within this large dataset with many types of aberrance likely corresponds closely to circumstances under which actual test security investigators operate, where examinees taking a large-scale test exhibit a variety of aberrance types and the investigator does not know the types or amounts of aberrance present (Sinharay, 2017). This scenario is not comprehensive, though, as different tests probably produce different amounts and types of aberrance (for example, a professional licensure exam versus a K-12 accountability test). Moreover, there is no analogous way to aggregate to a single AUROC for the other dimensions of the crossed design. Aggregations using the Dersimonian-Laird algorithm therefore help provide a more comprehensive picture of each PFS’s power to detect cheating/aberrance.

Finally, for any simulations that involve cheating alongside another aberrance type, I conclude by considering the number of examinees that each statistic flags and the percentage of the flagged examinees who were actually cheating when a specific critical value is used to distinguish flagged from non-flagged examinees, known as positive predictive value (PPV). Sinharay (2017) includes a related but distinct analysis looking at what percentage of aberrant examinees were flagged for each test length, which is the true positive rate or sensitivity. Similarly, one can calculate specificity/true negative rate, the proportion of non-cheating examinees who were not flagged. While both of these statistics are widely used, they often do not tell the full story when it comes to the value of a given statistic for making useful predictions. It is my contention that what matters most operationally is the proportion of the flagged examinees who actually cheated, which is what PPV expresses. If a PFS flags a very high proportion of
cheating examinees, but also many more times non-cheaters than cheaters, then the value of person fit analysis in such a scenario appears to be minimal. In such a scenario, the PFS would have high sensitivity and could also have very high specificity. For example, consider a scenario of 10,000 examinees where 100 have cheated, corresponding to the 1% aberrance condition in the second set of simulations. A PFS might flag 95 cheaters as cheating, leading to 95% sensitivity. Now, imagine that the PFS also flags 900 non-cheaters. Specificity is thus 91%. PPV, however, is about 10%. While not being flagged in such a scenario corresponds to a very low probability of having cheated, there is also a fairly low probability that a given flagged examinee truly cheated. Thus, in this scenario, being flagged as cheating by the given PFS is only very weak evidence of cheating. However, when cheating is more prevalent, PPV likely tells a similar story to sensitivity and specificity. I report PPV in this study because it appears to correspond most closely to the types of judgments for which PFSs are used.

4 Findings

4.1 Results: Original Simulations

This section presents results of replicating the simulations of Karabatsos (2003) and Sinharay (2017), then fitting the eight PFSs specified in this study. Overall AUROCs calculated using the Dersimonian-Laird algorithm are found in Table 2. The AUROCs for $H^T$ and $U_3$ are almost identical to those found in Sinharay (2017). $l^*$ and $ECI^*$, asymptotically-corrected versions of the parametric statistics used in Sinharay (2017), produce average AUROCs that are again nearly identical to those found in Sinharay (2017); as the original study reports AUROCs to two decimal places, it is unknown exactly how close the AUROCs found here are to those found in his study, but it is clear that if the asymptotic corrections yield any additional power, it is a very small amount, and that $H^T$, $l^*$ and $ECI^*$ are of essentially equal power. This is not entirely unexpected given that item responses were simulated from the Rasch model. Turning to the weighted PFSs, we can see that $SHa(1/2)^*$ and $SHa(1)^*$ appear to be of approximately equal power to the traditional PFSs, while $SHb(2)^*$ and especially $SHb(3)^*$ appear less powerful; no weighted PFS outperforms the traditional PFSs.

Recall that the AUROCs here represent power to detect aberrance writ-large. Because most datasets in the original simulation contain no cheating, it does not make sense to try to calculate AUROCs that represent power to detect cheating specifically. However, as described above, one can combine all of the datasets of a given test length to produce a single very large dataset that contains all types of aberrance. From these, one can then produce AUROCs either for detecting all aberrance or for detecting cheating specifically. Results of doing so are found in Table 3. Results indicate that the traditional PFSs are roughly equally powered to detect both aberrance writ large and cheating specifically, and that this power is fairly strong, in the 0.91-0.93 range for tests of all three lengths. However,
for this large dataset with many types of aberrance present at once, \( SHa(1/2)^* \) and \( SHa(1)^* \) outperform the traditional statistics when trying to detect cheating specifically. Still, as shown in Table 4, Table 5, and Table 6, the average AUROC for the original 10,000-examinee datasets was roughly the same for the traditional PFSs and for \( SHa(1/2)^* \), while \( SHa(1)^* \) appears slightly underpowered in most scenarios compared to the traditional PFSs; \( SHb(2)^* \) and \( SHb(3)^* \) remain noticeably underpowered for most scenarios. This more closely mirrors the findings in Table 2, which also aggregates over AUROC calculated separately for each scenario. Recall, however, that these results are for detecting aberrance broadly, not cheating specifically, so they are less relevant to test security than the findings in Table 3.

It appears that \( SHa(1/2)^* \) is roughly as powerful as \( H^T \), \( U_3 \), \( I_2 \) and \( ECI4 \) in scenarios with 10,000 examinees and one type of aberrance, with \( SHa(1)^* \) appearing slightly underpowered. However, for a very large set of examinees with all types of aberrance at once, \( SHa(1/2)^* \) and \( SHa(1)^* \) both outperform the traditional statistics. Because the combined datasets that produced the results in Table 3 correspond more closely to a real test security investigation than the individual 10,000-item datasets per Sinharay (2017), this appears to be evidence that \( SHa(1/2)^* \) and \( SHa(1)^* \) may be more appropriate for some test security applications (large examinee pool, many types of aberrance) but slightly worse for smaller examinee pools where cheating is the only type of aberrance (see Table 5). However, as described above, most of the datasets in the original simulation contain no cheating, so they cannot be used to investigate power to detect cheating; this can only be done for the datasets that aggregate over test lengths, and each dataset has a fixed proportion of aberrance as well as proportion of cheating. This motivated the revised simulations whose results are described in the next section.

### 4.2 Results: Revised Simulations

This section presents findings based upon the revised set of simulations in which cheating is simulated alongside another type of aberrance. Results indicate that the power of the weighted PFSs \( SHa(1/2)^* \), \( SHa(1)^* \), \( SHb(2)^* \) and \( SHb(3)^* \) to detect cheating examinees specifically is generally inferior to the power of the traditional PFSs, which exhibit nearly identical power to detect cheating examinees. As shown in Table 7, \( SHa(1/2)^* \)’s overall average AUROC came closest to that of the traditional PFSs, with \( SHb(3)^* \)’s falling lowest; AUROCs are lower than in the original simulations for all statistics, but fall more for the weighted PFSs. Figure 2 shows that the distributions of the AUROCs also differ substantially. In particular, AUROCs for \( SHa(1/2)^* \) and \( SHa(1)^* \) are distributed more widely than for the traditional PFSs, and are skewed lower, producing lower average AUROCs despite similar medians. This means that in some scenarios, these PFSs exhibit exceptionally poor power relative to the traditional PFSs, even as they perform similarly or better in others. I explore this further below.

**Figure 2**

**Distribution of AUROCs, Revised Simulations**

Turning to the AUROCs for the 200,000-examinee combined datasets for each test length (Table 8), results are similar to those found when using data generated in the original simulations, with \( SHa(1/2)^* \) and \( SHa(1)^* \) exhibiting slightly higher power to detect cheating than the traditional PFSs, though this gap all but disappears for longer test lengths. \( SHa(1/2)^* \) and \( SHa(1)^* \) also appear to detect general aberrance with about the same power as the traditional statistics. As in the original simulations, \( SHb(2)^* \) and \( SHb(3)^* \) are the two least powerful statistics in all scenarios.

The average AUROCS in Table 7 are largely consistent across the other levels of aggregation, as shown in Tables 9-11, though in some scenarios, the performance of \( SHa(1/2)^* \) was on par with or slightly better than that of the traditional PFSs. However, \( SHa(1/2)^* \) was less powerful than the traditional PFSs with a high (25%) rate of aberrance, while \( SHb(2)^* \) appears to be more powerful with this rate of aberrance. Overall, it remains the case that
Table 4

AUROCs for Detecting Aberrance by Test Length, Original Simulations

<table>
<thead>
<tr>
<th>Length</th>
<th>( l^*_z )</th>
<th>( H_T )</th>
<th>( U_3 )</th>
<th>( ECI^*_z )</th>
<th>( SHa(1/2)^* )</th>
<th>( SHa(1)^* )</th>
<th>( SHb(2)^* )</th>
<th>( SHb(3)^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>0.80</td>
<td>0.81</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.79</td>
<td>0.77</td>
<td>0.72</td>
</tr>
<tr>
<td>33</td>
<td>0.84</td>
<td>0.85</td>
<td>0.84</td>
<td>0.83</td>
<td>0.84</td>
<td>0.83</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td>65</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>0.86</td>
<td>0.87</td>
<td>0.86</td>
<td>0.83</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 5

AUROCs for Detecting Aberrance by Aberrance Type, Original Simulations

<table>
<thead>
<tr>
<th>Type</th>
<th>( l^*_z )</th>
<th>( H_T )</th>
<th>( U_3 )</th>
<th>( ECI^*_z )</th>
<th>( SHa(1/2)^* )</th>
<th>( SHa(1)^* )</th>
<th>( SHb(2)^* )</th>
<th>( SHb(3)^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheating</td>
<td>0.87</td>
<td>0.88</td>
<td>0.86</td>
<td>0.87</td>
<td>0.85</td>
<td>0.81</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>Creative</td>
<td>0.70</td>
<td>0.67</td>
<td>0.71</td>
<td>0.70</td>
<td>0.69</td>
<td>0.67</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>Guessing</td>
<td>0.88</td>
<td>0.89</td>
<td>0.87</td>
<td>0.88</td>
<td>0.86</td>
<td>0.81</td>
<td>0.83</td>
<td>0.76</td>
</tr>
<tr>
<td>Careless</td>
<td>0.73</td>
<td>0.69</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.72</td>
<td>0.68</td>
<td>0.64</td>
</tr>
<tr>
<td>Random</td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.86</td>
<td>0.84</td>
<td>0.81</td>
<td>0.80</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 6

AUROCs for Detecting Aberrance by Percent Aberrant, Original Simulations

<table>
<thead>
<tr>
<th>Percent</th>
<th>( l^*_z )</th>
<th>( H_T )</th>
<th>( U_3 )</th>
<th>( ECI^*_z )</th>
<th>( SHa(1/2)^* )</th>
<th>( SHa(1)^* )</th>
<th>( SHb(2)^* )</th>
<th>( SHb(3)^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>10%</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.92</td>
<td>0.90</td>
<td>0.84</td>
</tr>
<tr>
<td>25%</td>
<td>0.87</td>
<td>0.89</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>0.84</td>
<td>0.81</td>
<td>0.73</td>
</tr>
<tr>
<td>50%</td>
<td>0.59</td>
<td>0.60</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>0.61</td>
<td>0.60</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Table 7

Overall AUROCs Eight Person-Fit Statistics for Detecting Cheating

<table>
<thead>
<tr>
<th>( l^*_z )</th>
<th>( H_T )</th>
<th>( U_3 )</th>
<th>( ECI^*_z )</th>
<th>( SHa(1/2)^* )</th>
<th>( SHa(1)^* )</th>
<th>( SHb(2)^* )</th>
<th>( SHb(3)^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.81</td>
<td>0.80</td>
<td>0.81</td>
<td>0.81</td>
<td>0.79</td>
<td>0.76</td>
<td>0.77</td>
<td>0.71</td>
</tr>
</tbody>
</table>

the weighted PFSs tended to be underpowered compared to the traditional PFSs.

As Figure 3 shows, the relative power of the PFSs does differ somewhat by test length, but the distributions of the AUROCs for the weighted PFSs follows a fairly consistent pattern. AUROCs for \( SHb(3)^* \) are tightly dispersed and low; they are slightly higher and more spread out for \( SHb(2)^* \), but the distribution is skewed toward lower AUROCs. The 75th percentile of the AUROCs for \( SHa(1/2)^* \) and \( SHa(1)^* \) are actually higher for 17- and 33-item tests, but for all test lengths, these statistics produce the most widely dispersed AUROCs. While they perform very well for certain datasets, this performance is not consistent.

Comparing each statistic’s power across different combinations of aberrance in Figure 4 and Table 10 is also revealing. When the nature of the aberrance among examinees is cheating only, cheating plus guessing, or cheating plus random responding, the power of all statistics except \( SHb(3)^* \) is at least 0.8. However, power drops substantially, below 0.74, for all statistics when cheating happens alongside careless responding or creative responding. It appears that careless and creative responses are likelier to be flagged as misfitting alongside cheating.
when compared to guessing or random responding. This seems to make sense in terms of the different types of aberrance. Careless and creative responding produce spuriously low scores for high examinees on easy items. These examinees’ correct responses to hard items then appear aberrant, even though those responses are the ones that actually reflect the examinees’ “normal” response process. That said, \( SHa(1/2)^* \) and especially \( SHa(1)^* \) produce much wider distributions of AUROCs for the scenarios where cheating occurs alongside guessing and random responding than the other PFSs. This means that these statistics may not produce consistent classifications even in the scenarios in which other PFSs perform best.

Finally, the picture becomes even more complicated when AUROCs are broken out by the proportion of aberrant respondents in Figure 5/Table 11. While \( SHb(2)^* \) and \( SHb(3)^* \) produce the lowest AUROCs at 1%, 5% and 10% aberrance, they produce the highest AUROCs at 25%, while \( SHa(1/2)^* \) and especially \( SHa(1)^* \) lose power precipitously at 25% aberrance. Overall, there is no single “best” PFS among the weighted statistics.

### 4.3 PPV Analysis

It is not necessarily clear from the AUROCs alone what the implications of these findings are for a process that involves trying to identify cheaters using PFSs. There are two reasons for this. First, the AUROC does not itself contain information about the proportion of true positives among all positives, meaning that a statistic might flag far more non-cheating examinees than cheating examinees even if its AUROC is relatively high. Second, the AUROC is a summary over all possible critical values (cutoff values to delineate flagged from non-flagged examinees) of a given statistic, making it difficult to translate from an AUROC into, for example, a raw number of true or false positives in a given scenario. This requires identifying a single critical value.

To try to make clear the implications of the AUROCs observed here, I summarized the average number of total (true and false) positives (out of 10,000 total examinees) that result from the use of each statistic at each proportion of aberrant respondents, as well as PPV, the percentage of flagged examinees who were truly cheating. Here, I use a specific critical value for each statistic. Most of the statistics follow an asymptotically standard normal null distribution: \( SHa(1/2)^* \), \( SHa(1)^* \), \( SHb(2)^* \) and \( SHb(3)^* \) all share this null distribution, as do \( l^*_z \) and \( ECI^*_z \). For these statistics, I set the critical value to 1.64 or -1.64, depending on whether high or low values of the statistic indicate aberrance. This corresponds to an \( \alpha \) level (type I error rate when no aberrance is present) of 0.05. For \( H^T \) and \( U_3 \), I first standardized the statistics so that equivalent critical values could be used. It is important to note that I standardized across all examinees, cheating or otherwise, in line with

<table>
<thead>
<tr>
<th>Target</th>
<th>Items</th>
<th>( l^*_z )</th>
<th>( H^T )</th>
<th>( U_3 )</th>
<th>( ECI^*_z )</th>
<th>( SHa(1/2)^* )</th>
<th>( SHa(1)^* )</th>
<th>( SHb(2)^* )</th>
<th>( SHb(3)^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheating</td>
<td>17</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.95</td>
<td>0.95</td>
<td>0.79</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.92</td>
<td>0.92</td>
<td>0.81</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
<td>0.90</td>
<td>0.82</td>
</tr>
<tr>
<td>Aberrance</td>
<td>17</td>
<td>0.87</td>
<td>0.86</td>
<td>0.86</td>
<td>0.87</td>
<td>0.87</td>
<td>0.86</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.91</td>
<td>0.89</td>
<td>0.84</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>0.91</td>
<td>0.85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Length</th>
<th>( l^*_z )</th>
<th>( H^T )</th>
<th>( U_3 )</th>
<th>( ECI^*_z )</th>
<th>( SHa(1/2)^* )</th>
<th>( SHa(1)^* )</th>
<th>( SHb(2)^* )</th>
<th>( SHb(3)^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.75</td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>33</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>0.80</td>
<td>0.78</td>
<td>0.75</td>
<td>0.75</td>
<td>0.69</td>
</tr>
<tr>
<td>65</td>
<td>0.85</td>
<td>0.84</td>
<td>0.85</td>
<td>0.85</td>
<td>0.82</td>
<td>0.79</td>
<td>0.83</td>
<td>0.77</td>
</tr>
</tbody>
</table>
a scenario in which an investigator does not know which examinees have cheated or otherwise responded aberrantly. This may be a disadvantage of using nonparametric PFSs in practice.

Table 12 contains the results of this analysis. For each statistic, across each percentage of examinees modeled as aberrant, I present the mean number of flagged examinees (out of 10,000) and PPV when a critical value corresponding to an \( \alpha \) level of 0.05 is used to separate flagged examinees from non-flagged. We can see that for every statistic and percent of aberrant respondents, the average percentage of flagged examinees who were truly cheating is at or below 30.8%. When only 1% of respondents are aberrant, the fewest examinees tend to be flagged, but PPV is also lowest; when more respondents are aberrant, both the number flagged and the proportion of flagged examinees who cheated are higher. PPV remains low, though. At high amounts of aberrance, the statistics that produce the most true positives among the examinees they flag are \( SHb(2)^* \) and \( SHb(3)^* \), which AUROC-based analyses showed to generally be the least powerful statistics in cases with less aberrance.
I also completed the same analysis using the same cutoff for the 200,000-examinee datasets produced when following Sinharay (2017)’s methodology for aggregating over test length. Here, I compared results from the datasets from the original simulations (replication of Karabatsos [2003]/Sinharay [2017]) to results from the datasets in the new simulations introduced in this study. Results are presented in Table 13. There is a clear difference in the statistics’ value in detecting cheating, expressed by PPV, between the original simulation conditions and the revised conditions. For the revised simulations, PPV is generally far lower, especially on the shortest test length. However, relative to the other PFSs, $SHb(2)^*$ and $SHb(3)^*$ perform far worse under the conditions of the original simulation for longer test lengths. Meanwhile we can see that on average, $SHA(1/2)^*$ and $SHA(1)^*$ perform best on longer tests for both combined datasets, but that their superior performance on 17-item tests under conditions from Sinharay (2017) does not hold up under the new conditions introduced in this study. This underscores the inconsistent relationship between the power of specific PFSs to detect cheating and the generating conditions of the data. Ultimately, there is no clear winner in Table 13 but $SHb(2)^*$ and $SHb(3)^*$ produce the lowest average PPV values under most conditions.

5 Discussion

This study finds that for relatively small (10,000 examinees) datasets, PFSs constructed specifically to detect spuriously high scores are rarely any better at flagging cheating than traditional PFSs, and are often worse for this task. I am aware of two reasons why this might be the case. First, it may be the case that the specific statistics investigated here are too specific in the type of aberrance they detect most effectively:

$SHb(2)^*$ (i.e., $\beta = 2$) and $SHb(3)^*$ (i.e., $\beta = 3$) have relatively large weights on the items with $P_r(\hat{\theta}) < .3$ compared with $SHA(1/2)^*$ and $SHA(1)^*$. The authors expect that $SHb(2)^*$
and $SHb(3)^*$ can have greater sensitivity to detect spuriously high scores on items with $0.1 < P_n(\hat{\theta}) < 0.3$, although they may lose some power for the extremely difficult items with $P_n(\hat{\theta}) < 0.1$ compared with $SHa(1/2)^*$ and $SHa(1)^*$ (Xia & Zheng, 2018, p. 349).

These properties are at odds with the scenarios simulated in this study, and potentially with the circumstances under which cheating occurs in the real world. In this study, examinees of a relatively wide range of true ability ($\theta < -0.5$) were simulated as cheating, and items on which examinees cheated were selected at random from a relatively wide range of difficulty ($\delta > 0$). This means that cheating respondents’ probability of a correct response on a given item had they not cheated will vary widely based on the examinee’s ability and the item’s difficulty. What Xia and Zheng (2018) describe above gives reason to believe that no one statistic of the four they propose is likely to be uniformly most powerful even within a single dataset, much less over all sixty simulated datasets.

Additionally, some types of non-cheating aberrant responding—in this study, careless and creative responding—are likely to appear to produce spuriously high scores, even though they are actually producing spuriously low scores. Both of these types of aberrance involve examinees of high ability responding to easy items in a way that makes their probability of a correct response on those easy items lower than it “normally” would be. This means that such respondents will have
lower scores than expected on easy items, and their ability estimates will be deflated as a result. Because the PFSs intended to flag spuriously high scores place more weight on correct answers to items that should be difficult for a given examinee given their ability estimate, and because the examinees being considered here will have their ability estimates deflated, their correct responses to hard items will appear unlikely and the PFSs will flag them as aberrant even though their responses to hard items reflect those examinees’ underlying ability. If the intent is to flag cheating specifically, this is a problem.

However, when datasets are combined to produce a large dataset (200,000 examinees) with all types of aberrance present at once, $SHa(1/2)^*$ and $SHa(1)^*$ are the most powerful statistics in the study. It appears to be the case that in some scenarios, $SHa(1/2)^*$ and $SHa(1)^*$ provide an advantage in the detection of cheating, but in others, they provide a disadvantage. This further muddies the task of trying to select a single “best” PFS for all test security applications and reinforces the importance of understanding the expected power of PFSs relative to the anticipated behavior of test examinees. Still, a statistic is not necessarily appropriate for test security use just because it produces high AUROCs, which is why this study also considered raw numbers of flagged examinees at a sensible critical value and the proportion of those flagged examinees who truly cheated, PPV. Results indicated that depending on an investigator’s goals and the properties of a given test, some of which (such as the true amount and nature of aberrance) may not be known at the time of investigation, different statistics will be most appropriate and useful.

Looking beyond relative power provided some evidence of potential validity issues in a test security program incorporating analysis of PFSs. In general, PFSs appear likely to flag far more non-cheating respondents than cheating respondents as aberrant, though this is again highly contingent upon the behavior of examinees, test length, and sample size. In the revised simulations, with $\alpha = 0.05$ and 1% true aberrance, for every statistic, 93% or more of the flagged examinees were not cheating. In light of this finding, there are some applications in test security where it appears that the value of person fit analysis as supplementary evidence to identify cheating examinees is minimal. If as few as one in fifty flagged examinees will have actually cheated, then PFSs contribute very little to the process of finding convergent evidence pointing to a security violation (because even being flagged is associated with only a very low probability of having cheated). However, if one in three flagged examinees have cheated, as found in some scenarios where cheating is more prevalent, then PFSs appear to provide some value as supplementary evidence after all. It is therefore important to try to anticipate how PFSs will function given the characteristics of a particular test (length, anticipated examinee behavior, etc.) before using them operationally in test security. The trouble, of course, is that the true proportion of cheating examinees is essentially never known, certainly not ahead of time, and this study shows that the meaning of being flagged by person fit analysis can differ substantially depending on this unknowable fact.

Test-specific simulations may help somewhat in this task, but can only go so far because it is impossible under almost all circumstances to know with certainty whether a given examinee cheated, guessed, responded carelessly or otherwise deviated from the normal response process. The assumptions made in any test-specific simulations need to be made explicit and organizations should be prepared to discard PFSs in their security analyses if evidence emerges that these assumptions were incorrect. There are many other sources of evidence available for investigators of test security problems with stronger theoretical connections to cheating (Cizek & Wollack, 2017). While PFSs are easy to calculate, the present study indicates that their value is highly contingent upon unknowable aspects of examinee behavior.

This study does, naturally, have a number of limitations and extensions that further work can address. The most pressing of these are the nature of aberrance simulated, the types of test items, and the IRT model used. Future studies should attempt to produce more realistic simulations of cheating—as noted, it is unlikely that only lower-performing students cheat, and detecting cheating among students of all abilities is likely even harder than detecting it in the scenarios simulated here. Extensions to this study should also consider polytomous, clustered, and other complex item types. Many PFSs are flexible enough for use with a variety of item types, and work in this area is ongoing (Sinharay, 2016b); further studies will hopefully provide more insight into how PFSs should or should not be used on tests with different item types.

Finally, this study simulated examinees’ responses using the Rasch model. Other models, such as the three-parameter logistic (3PL), may warrant investigation; Sinharay (2017) finds lower (around 0.64) AUROCS and inconsistent performance across different datasets when the 3PL model is used for simulation and parameter
estimation when compared to the Rasch model, so this study likely represents a best-case scenario relative to the same types of aberrance modeled using the 3PL. However, in contrast, Drasgow (1982) found that when the 3PL and 1PL/Rasch models were used model the same item response dataset containing aberrant response patterns interspersed with real data, detection under the 3PL was slightly more effective. Notably, Drasgow used sensitivity (discussed previously) and false alarm rate (proportion of non-aberrant respondents flagged as aberrant) to evaluate effectiveness. The present study demonstrated that different PFSs can look more or less effective depending on evaluation criteria, so differences in Drasgow’s and Sinharay (2017)’s findings may be a matter of evaluation criteria, simulation conditions, or both. Additionally, as the present study demonstrates, effective detection of aberrance and effective detection of cheating specifically can differ, and prior studies using the 3PL have focused on detection of aberrance writ-large. Finally, Hong et al. (2020) demonstrate that IRT model misspecification can often lead to inflated Type I error rates, which are related to the sometimes very low PPV values found in this study. Misspecification is a concern in this study for the “guessing” aberrance type, as the Rasch model does not include a lower asymptote that can capture guessing, while the 3PL does. Notably, however, Figure 4 and Table 10 show that guessing did not induce poor performance relative to the other aberrance conditions; cheating was hardest to detect when mixed with creative or careless responding, as discussed above. It would be valuable to learn if this holds under other IRT models. Certainly, the findings of this and prior studies bring into focus the need to more thoroughly explore the use of person fit analysis under the 3PL. The use of the 3PL will also probably bring into sharper focus the pitfalls of using asymptotically uncorrected standardized PFSs such as $l_z$ and $ECI_4$, because prior studies have found greater benefits from correction under IRT models more complex than the Rasch (Sinharay, 2016a; Snijders, 2001).

References


Meijer, R. R., & Tendeiro, J. N. (2012). The use of the $l_z$ and $l_z$ person-fit statistics and problems derived from


