A Study of Person Fit Based on Residual Statistics in Polytomously Scored Tests

Authors: Tong Hao, Yu Xiaofeng, Qin Chunying, Peng Yafeng, Zhong Xiaoyuan, Yu Xiaofeng

Date: 2022-04-06T17:30:59+00:00

Abstract

This study proposes a personal fit statistic R for polytomous scoring items to investigate its performance in detecting six common aberrant response patterns (cheating, guessing, random responding, carelessness, innovative responding, and mixed aberrance), and compares it with the standardized log-likelihood statistic lzp. The results indicate that: (1) when the proportion of aberrant responses is low and the aberrant response types are cheating and guessing, the detection rate of R is significantly higher than that of lzp; (2) the detection rates of both statistics increase with test length and the degree of aberrance; (3) under certain conditions, the detection effects of R and lzp are comparable. Empirical data analysis further illustrates the application method and procedure of the R statistic, and the results also demonstrate that the R statistic possesses promising application prospects.

Full Text

Preamble

A Residual-Based Person-Fit Statistic for Polytomous Scoring Tests

Hao Tong[1], *Xiaofeng Yu*[1], *Chunying Qin*[2], *Yafeng Peng*[1], *Xiaoyuan Zhong*[1]
([1] School of Psychology, Jiangxi Normal University, Nanchang, 330022, China)
([2] School of Mathematics and Information Science, Nanchang Normal University, Nanchang, 330032, China)

Abstract

This paper proposes a person-fit statistic, denoted as , for polytomously scored items and examines its performance in detecting six common types of aberrant response patterns: cheating, guessing, random responding, carelessness, creative responding, and mixed aberrant behaviors, comparing it against the

standardized log-likelihood statistic . The results demonstrate that: (1) when the proportion of aberrant responses is low and the aberrant behavior involves cheating or guessing, achieves significantly higher detection rates than ; (2) as test length and the severity of aberrant behavior increase, the detection rates of both statistics improve; and (3) under certain conditions, the detection performance of approaches that of . An empirical data analysis further illustrates the application procedure of the statistic and confirms its promising practical utility.

Keywords: polytomous scoring, item response theory, person-fit statistic, aberrant behavior detection, graded response model

Introduction

Educational and psychological measurement aims to assess individuals' latent traits to inform and guide their future development. These traits may encompass domain-specific knowledge and skills or psychological attributes such as attitudes and emotions. To this end, researchers extensively employ tests and questionnaires as measurement instruments. However, during actual test administration, various extraneous factors inevitably influence examinees' responses, thereby threatening test validity. Examples include cheating and low test motivation. Unlike random error in the response process, these factors often introduce systematic error into measurement data, a phenomenon referred to as aberrant responding.

Classical test theory (CTT) and item response theory (IRT) are commonly used frameworks for estimating examinee ability levels, with their accuracy depending on the degree of model-data fit (Hotaka & Maeda, 2017). If aberrant response data are directly used for analysis and calculation, the resulting estimates will be substantially biased, rendering any conclusions and subsequent decisions (such as personnel selection or placement) invalid. This contamination seriously undermines numerous aspects of testing, including parameter estimation (Oshima, 1994; Schnipke, 1996; Shao et al., 2016), equating (Wollack et al., 2003), and reliability and validity (Gulliksen, 1950, p. 236; Lu & Sireci, 2007), ultimately compromising test fairness and accuracy (Buchanan & Smith, 1999; Glas & Dagohoy, 2007; Huang et al., 2015). Numerous studies have reported the prevalence of aberrant responding among test-takers, with proportions ranging from 3.5% to 50% and typically around 20% (Curran et al., 2010; Meade & Craig, 2012; Rupp, 2013; Meade, 2016; Shao et al., 2016; Yu & Cheng, 2019). Consequently, identifying examinees with aberrant response patterns has long been a central concern in educational and psychological measurement (Schnipke & Scrams, 1997).

To address this issue, researchers have developed person-fit statistics (PFS) to extract information from response data and identify aberrant individuals. Meijer and Sijtsma (2001) classified PFS into two categories: (1) nonparametric PFS, which compute nonparametric statistics using observed response data or

compare individuals to groups; and (2) parametric PFS, which establish model assumptions, estimate parameters from data, construct fit statistics, and compare them to the distribution under the null hypothesis that the examinee belongs to the normal population.

Person-fit research traces back to the 1940s, when Guttman (1944, 1950) integrated observed responses, estimated abilities, and item cutting points to determine response appropriateness, laying the foundation for subsequent work. Later researchers proposed nonparametric PFS such as point-biserial and biserial correlations (Donlon & Fischer, 1968). While convenient to compute, these methods typically fail to extract deeper information from data and yield ambiguous interpretations. In contrast, parametric PFS rely on rigorous mathematical models (e.g., Rasch models) to quantify aberrance and provide clear conclusions based on established criteria. Parametric PFS can be further divided into residual-based and likelihood-based approaches. The former focuses on residuals between observed and expected responses, applying appropriate weighting schemes—such as the and statistics proposed by Wright and Stone (1979) and Wright and Masters (1982), which use the inverse of average conditional variance and total test variance as weights, respectively. The latter constructs statistics based on response likelihood, exemplified by the log-likelihood index $_0$ (Levine & Rubin, 1979) and the standardized log-likelihood index (Drasgow et al., 1985), which overcame distributional limitations of $_0$.

Current person-fit research predominantly focuses on dichotomous (0-1) scoring contexts. However, polytomous data are common in educational and psychological assessments, such as constructed-response items in mixed-format tests or Likert-type scales in psychological questionnaires. Compared to dichotomous items, polytomous items provide more information and achieve equivalent measurement precision with fewer items (van der Ark, 2001). Existing parametric PFS for polytomous data include the standardized weighted mean square residual statistic (Wright & Masters, 1982) and the polytomous standardized log-likelihood index (Drasgow et al., 1985). However, these statistics have limitations: (1) Rogers and Hattie (1987) noted that is insensitive to classifying aberrant response patterns, and our pilot study confirmed its poor detection capability, limiting its practical utility; (2) while exhibits good distributional properties and overall detection performance, researchers are particularly concerned with low-ability examinees showing unexpectedly high performance, as high-stakes test outcomes carry significant consequences, and such patterns often indicate serious test security breaches like item preknowledge or cheating. In contrast, high-ability examinees showing unexpectedly low performance typically result from individual factors like fatigue or rapid guessing and pose less threat. Therefore, a polytomous PFS sensitive to aberrantly high performance is urgently needed.

2.1 Polytomous IRT Models

Various IRT models have been developed for polytomous data, including the graded response model (GRM; Samejima, 1969) and the partial credit model (PCM; Masters, 1982). This study employs GRM, though findings can be generalized to other polytomous models like PCM, allowing researchers to select models appropriate for their applications.

In GRM, each item has one discrimination parameter and multiple difficulty parameters. Let denote an examinee's response to item , with representing the maximum score for item , such that {0,1,…, }. Let represent the examinee's latent trait level, the item discrimination, and = (1, 2,…,) the vector of difficulty parameters. For each difficulty threshold , the probability of obtaining a score of or higher is denoted as ∗( ) and modeled using the two-parameter logistic function:

$$P_{jk}^{*}(\theta) = \frac{1}{1 + e^{-1.702a_j(\theta - b_{jk})}}$$

Since ∗( ) represents the probability of achieving score or above, to obtain the probability for each specific score category, we subtract adjacent cumulative probabilities:

$$P_{jk}(\theta) = P_{jk-1}^{*}(\theta) - P_{jk}^{*}(\theta)$$

For the lowest category, the probability is:

$$P_{j1}(\theta) = 1 - P_{j1}^{*}(\theta)$$

Consider an item with maximum score 4, yielding five response categories (0,1,2,3,4). The probabilities for each category are:

$$\begin{aligned}
P(X_j = 0|\theta) &= 1 - \frac{1}{1 + e^{-1.702a_j(\theta - b_{j1})}} \\
P(X_j = 1|\theta) &= \frac{1}{1 + e^{-1.702a_j(\theta - b_{j1})}} - \frac{1}{1 + e^{-1.702a_j(\theta - b_{j2})}} \\
P(X_j = 2|\theta) &= \frac{1}{1 + e^{-1.702a_j(\theta - b_{j2})}} - \frac{1}{1 + e^{-1.702a_j(\theta - b_{j3})}} \\
P(X_j = 3|\theta) &= \frac{1}{1 + e^{-1.702a_j(\theta - b_{j3})}} - \frac{1}{1 + e^{-1.702a_j(\theta - b_{j4})}} \\
P(X_j = 4|\theta) &= \frac{1}{1 + e^{-1.702a_j(\theta - b_{j4})}}
\end{aligned}$$

2.2 The Statistic

Wright and Masters (1982) proposed a standardized weighted mean square residual statistic :

$$\upsilon = \frac{\sum_{j=1}^{M} \frac{[X_j - E(X_j|\theta)]^2}{\sum_{k=0}^{K}(k-E(X_j|\theta))^2 P_{jk}(\theta)}}{M}$$

where is the number of items, is the maximum score, is the observed score on item , ( ) is the probability of obtaining score on item for an examinee with ability , and (|) is the expected score.

Our pilot study revealed that lacks sufficient power under aberrant responding conditions (see Appendix A), so we excluded it from subsequent simulation comparisons.

2.3 The Statistic

The log-likelihood statistic $_0$ was originally developed for dichotomous items (Levine & Rubin, 1979). Drasgow et al. (1985) standardized $_0$ to create the standardized log-likelihood statistic and derived its polytomous extension . For an examinee's response pattern $= \{ _1, _2, \ldots, \}$, the log-likelihood function is:

$$l_p = \ln[L(\theta)] = \sum_{j=1}^{M}\sum_{k=0}^{K} d(X_j = k) \ln P_{jk}(\theta)$$

where ( ) is the likelihood function and ($\cdot$) is an indicator function equal to 1 when the condition is satisfied and 0 otherwise. Standardizing yields:

$$l_{zp} = \frac{l_p - E(l_p)}{\sqrt{Var(l_p)}}$$

The expectation and variance of are:

$$E(l_p) = \sum_{j=1}^{M}\sum_{k=0}^{K} P_{jk}(\theta) \ln P_{jk}(\theta)$$

$$Var(l_p) = \sum_{j=1}^{M}\left[\sum_{k=0}^{K} P_{jk}(\theta)[\ln P_{jk}(\theta)]^2 - \left(\sum_{k=0}^{K} P_{jk}(\theta) \ln P_{jk}(\theta)\right)^2\right]$$

The standardized exhibits good distributional properties (asymptotically normal) and superior detection performance across various conditions (Karabatsos, 2003), leading Nering (1995) to deem it "the most promising PFS."

3. Development of a Weighted Residual-Based PFS for Polytomous Items

Our simulation experiments comparing existing PFS revealed that residual-based statistics demonstrate advantages in detecting aberrant high-ability performance (cheating, lucky guessing), consistent with Karabatsos (2003). To explain this phenomenon, we examined the construction logic of different PFS under polytomous scoring and found that residual-based statistics are inherently sensitive to higher score categories. Building on this insight, we extended residual-based parametric statistics within the polytomous IRT framework to develop a person-fit statistic sensitive to aberrant high performance.

The core idea of residual-based fit statistics is quantifying discrepancies between observed and ideal response patterns. An ideal pattern is generated by strictly following the response function's probability distribution given model parameters (ability, difficulty, discrimination), where examinees are more likely to obtain low scores on relatively difficult items and high scores on relatively easy items. Misfitting patterns necessarily deviate substantially from ideal patterns (Meijer & Sijtsma, 2001), providing a basis for aberrance detection. To compute residuals accurately, we use expected scores to represent ideal responses. Let be the observed response and (|) the expected score for item ; the item-level residual is = - (|), which can be decomposed as:

$$r_j = X_j - E(X_j|\theta) = \sum_{k=0}^{K} k \cdot d(X_j = k) - \sum_{k=0}^{K} k \cdot P_{jk}(\theta) = \sum_{k=0}^{K} k \cdot [d(X_j = k) - P_{jk}(\theta)]$$

Here, is the maximum score for item , and ($\cdot$) is the indicator function. Equation (10) shows that each score category's discrepancy is weighted by , meaning higher scores receive greater weight. Under normal conditions, responses cluster around the highest-probability categories, yielding small residual sums. For aberrant responses, residuals become substantially larger, particularly for aberrant high performance where high score categories amplify residual contributions. Thus, residual-based statistics are inherently sensitive to aberrant high-ability patterns.

However, residuals alone cannot capture potential differences across items. For the same examinee, different item parameters produce different score probability distributions, so equal residuals across items should not be treated identically. Following Snijders (2001), a common approach applies a weight function ( ) that integrates item and examinee characteristics to give residuals more appropriate contributions. The weight function amplifies "suspicious" components

while attenuating relatively normal ones to maximize aberrance sensitivity. For instance, when an examinee's score approximates the expected score, the weight should be small to reduce normal data's contribution; conversely, when scores deviate markedly from expectations, the aberrant impact should be amplified. Drawing on standardized residuals used by Masters and Wright (1997) and weighted residuals by Yu and Cheng (2019), we define the weight function as:

$$w_j(\theta) = \sum_{k=0}^{K} d(X_j = k) \cdot P_{jk}(\theta)$$

This weight depends on the theoretical probability of the observed score. Higher probabilities indicate more normal responding consistent with the score distribution, while lower probabilities suggest greater aberrance. Consequently, ( ) is small under normal conditions and large under aberrant ones.

Furthermore, weighted residual statistics accumulate across all items to reflect overall aberrance. Since examinees may exhibit mixed aberrant behaviors (e.g., cheating on some items while speeding through others), residuals from high-ability and low-ability aberrant items can cancel each other, reducing detection power. To prevent this and maximize detection effectiveness, we accumulate absolute residuals. In practice, the true ability is unknown and must be replaced by its estimate ̂. For aberrant examinees, ̂ and (̂) deviate in the same direction, necessarily reducing computed residuals below theoretical values. For example, a low-ability examinee (= -2) who cheats may have an overestimated ability (̂ = 1). On a compromised item where the observed score = , since (̂) > (|), we have [- (̂)] < [- (|)]. This attenuates detection effectiveness. Therefore, we use absolute values to accumulate residuals. The statistic is defined as:

$$R = \sum_{j=1}^{M} |r_j| \cdot w_j(\theta) = \sum_{j=1}^{M} \left| \sum_{k=0}^{K} k \cdot [d(X_j = k) - P_{jk}(\theta)] \right| \cdot \sum_{k=0}^{K} d(X_j = k) \cdot P_{jk}(\theta)$$

Since expected scores are typically non-integer, even perfectly normal examinees produce non-zero residual sums. Ideally, remains small, while aberrant patterns yield significantly larger values. This forms the basis for PFS-based aberrance detection: after obtaining 's null distribution (under the null hypothesis), we set cutoffs at different Type I error rates to determine whether an examinee's falls in the acceptance region. As accumulates absolute weighted residuals, larger values indicate poorer fit, so hypothesis testing uses the right-tail probability. Note that when observed scores contain aberrant data, joint estimation of item and person parameters can produce a "masking effect" (Fung, 1993; Yuan & Zhong, 2008), where aberrant data influence parameter estimation and reduce detectability. This would hinder comparison between and .

Therefore, our simulation experiments compare their performance under known item parameters.

4. Simulation Study

The simulation study aims to evaluate 's detection capability for aberrant examinee groups, requiring a strong benchmark to reveal 's advantages. Given 's excellent overall performance and widespread attention (de la Torre & Deng, 2008; Sinharay, 2016), we selected it as the comparison target. Three studies were conducted: Study 1 simulated normal examinee groups under the null hypothesis to obtain statistical distributions of and across different test lengths, providing a foundation for Study 2; Study 2 simulated various aberrant testing scenarios to evaluate false alarm and detection rates; Study 3 applied to empirical data.

Using PFS for fit assessment requires obtaining null distributions. Even for standardized statistics like , caution is warranted. Sinharay (2016) found that (ˆ) does not follow an asymptotic standard normal distribution, and van Krimpen-Stoop and Meijer (2002) showed that for short tests (e.g., 20-30 items), 's distribution is negatively skewed. Since ability estimates ˆ are used in practice, cannot be assumed standard normal, and standard normal cutoffs should not be directly applied. To explore and compare the aberrant detection capabilities of and in polytomous tests, we designed the following simulation studies: Study 1 derives the distributions and critical values for and ; Study 2 computes and compares their false alarm and detection rates.

4.1 Study 1: Distributions and Critical Values of and

This study used GRM with five score categories, following numerous polytomous studies (Chen et al., 2010; Cheng et al., 2012; Dodd et al., 1995; Emons, 2008; Li & Ding, 2018; Sinharay, 2016). Four test lengths were constructed: 20, 40, 60, and 80 items, representing short, medium, longer, and long tests. Item difficulty parameters were sampled from a standard normal distribution and sorted in ascending order; discrimination parameters were sampled from a lognormal distribution with mean 0 and standard deviation 1 (Xiong et al., 2020; Xiong et al., 2018). For each test length, 10,000 examinees with ability parameters ~ N(0,1) were simulated, and observed scores were generated based on GRM.

When computing and distributions, we used known item parameters and estimated ability values ˆ. This reflects real testing situations where items are carefully developed and pretested, yielding calibrated parameters (Shao et al., 2016; Sinharay, 2016). With known item parameters, Type I error rates and power are unaffected by the number of examinees in the sample (Sinharay, 2016). To enhance generalizability, we used expected a posteriori (EAP) estimation for ˆ:

$$\hat{\theta}_{EAP} = \frac{\int \theta L(\theta) f(\theta) d\theta}{\int L(\theta) f(\theta) d\theta}$$

where ( ) is the likelihood function and ( ) is the prior density. EAP incorporates group prior information, improving estimation accuracy over maximum likelihood estimation (MLE). When prior information is vague, partially informative priors or MLE can be used.

Both and reflect fit unidirectionally: smaller indicates poorer fit, while larger indicates poorer fit. Therefore, for Type I error rates of 1%, 2.5%, and 5%, we selected the 1st, 2.5th, and 5th percentiles of 's distribution and the 99th, 97.5th, and 95th percentiles of 's distribution as critical values.

Figures 1 and 2 present the empirical distributions of and . Normality tests and skewness calculations show that is positively skewed while is negatively skewed. Table 1 displays the empirical critical values at various significance levels across test lengths. The critical values are similar across lengths due to its standardized construction, whereas 's critical values increase with test length because accumulates non-negative weighted residuals—more items produce larger values.

4.2 Study 2: Detection Rates of and

Using the PFS distributions and critical values from Study 1, Study 2 incorporated examinees affected by aberrant factors. Aberrant behaviors manifest in various ways, including cheating, item preknowledge, and low motivation (Meijer & Sijtsma, 2001; Rupp, 2013). Following Karabatsos (2003) and Doval and Delicado (2020), we simulated six polytomous aberrant conditions: (1) cheating, (2) lucky guessing, (3) random responding, (4) carelessness, (5) creative responding, and (6) mixed types. Definitions and operationalizations appear in Table 2, which can be summarized as: (1) low-ability examinees showing high performance (cheating, guessing); (2) high-ability examinees showing low performance (carelessness, creativity); and (3) pervasive random responding. Note that these simulated types do not exhaust real-world possibilities but represent the three categories above, so interpretations should be cautious.

The number of aberrant items was = × , where aberrance severity took values 0.1, 0.25, and 0.5 (low, moderate, high). This created 4 (test lengths) × 6 (aberrant types) × 3 (severity levels) = 72 conditions, each with 3,000 simulated examinees, replicated 100 times. Following prevalence rates in the literature (Curran et al., 2010; Meade & Craig, 2012; Rupp, 2013; Meade, 2016; Shao et al., 2016; Yu & Cheng, 2019), we set the aberrant examinee proportion at 20%.

Since our primary goal was comparing detection capability, we examined performance under known item parameters, ensuring aberrant examinees wouldn't affect parameter estimates and allowing PFS to better reflect ability-response

misfit. An examinee was flagged as aberrant if their PFS exceeded the critical value (< or >). Detection rate was defined as the proportion of flagged aberrant examinees, while false alarm rate was the proportion of normal examinees incorrectly flagged. With known item parameters, false alarm rates closely matched the nominal Type I error rates. Results appear in Tables 3-6.

Tables 3-6 show that actual false alarm rates closely approximate nominal Type I error rates because critical values were derived from normal examinee distributions, where extreme cases naturally exceed cutoffs at the specified rates. Several patterns emerge:

First, for test length 20 with low aberrance (10% of items affected), 's detection power exceeds 's across all aberrant types by an average of 17.6%. At moderate or high aberrance (25% and 50% of items), the two statistics perform similarly. For test length 40 with low aberrance, outperforms by 11.2% on average; at moderate and high aberrance, performance is comparable. For test length 60 with low aberrance, 's advantage is 8.5%; for length 80, it's 5.7%. Thus, as test length increases, 's advantage over at low aberrance diminishes. At moderate and high aberrance, detection rates are nearly identical, with occasionally slightly superior.

Second, the statistics show differential detection rates across aberrant types. Both perform well (typically >90%) for creative responding and cheating, though is notably weaker than for cheating at low aberrance. As aberrance increases, slightly outperforms , suggesting that while is sensitive to aberrant high performance, it is less robust to aberrance severity changes. Detection rates for other types are relatively lower because cheating and creative responding are extreme and easier to detect, whereas other types are more ambiguous. Both statistics show similar trends: detection rates increase with the proportion of aberrant items and with test length.

We also compared the area under the ROC curve (AUC) as a comprehensive performance metric independent of fixed thresholds. AUC values closer to 1 indicate better detection. Table 7 presents AUC results, and Figures 3-4 show differences in detection rates and AUC between and for test length 20 (trends were consistent across lengths).

Figure 3 reveals that at low aberrance, 's detection rate exceeds 's in nearly all conditions, particularly for lucky guessing where dominates. As aberrance increases, gains advantage, reflecting 's greater susceptibility to masking effects. At longer test lengths (60-80 items), 's detection rate falls slightly below 's for mixed aberrant types. At shorter lengths (20-40 items), outperforms for lucky guessing, random responding, and carelessness, while performance is similar or slightly lower for other types.

For AUC (Figure 4), except for random responding at moderate-high aberrance where 's AUC is lower, 's AUC is generally higher than or equal to 's across conditions. Since AUC represents overall performance independent of Type I

error settings, this indicates 's superior comprehensive detection, particularly for lucky guessing—a classic aberrant high-performance pattern.

5. Empirical Study

To demonstrate 's practical application, we analyzed data from the National Longitudinal Study of Adolescent Health (NLSAH, 1994-1995), an education and health study of 7th-12th grade adolescents (Harris & Udry, 2010). The dataset is available at https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21600/datadocumentation. Yu and Cheng (2019) conducted exploratory factor analysis supporting a unidimensional structure. The data contain 19 items measuring emotional states using a 4-point Likert scale (coded 0-3). Since few examinees selected category 4 ("most of the time" or "always"), categories 3 and 4 were merged.

After excluding 97 examinees with missing data, 6,457 examinees remained for analysis. Step 1: We fitted GRM to the response data using PARSCALE 4.1 to obtain item parameters. With categories 3 and 4 merged, each item had 3 response categories, yielding 2 difficulty parameters ($_1$, $_2$) and 1 discrimination parameter ( ) per item. Parameter estimates appear in Table 8.

Step 2: To use and , we generated normal response data to obtain distributions for determining critical values. In MATLAB 2020a, we simulated 10,000 examinees with ~ N(0,1) (the 6,457 examinees' abilities approximated N(-0.0003, 0.9222)) and generated responses based on GRM probabilities to compute and . To approximate real conditions, we used estimated abilities for the null distribution. Based on simulation results, we selected the 5% Type I error cutoff to control false alarms while maximizing detection. The 95th percentile of 's distribution was 98.49, and the 5th percentile of 's distribution was -1.23 (Figure 5).

Step 3: We computed and for the 6,457 actual examinees and compared them to critical values. flagged 423 examinees (6.6%) as aberrant, while flagged 562 (8.7%), with 253 flagged by both. Since true aberrance status is unknown in empirical data, false alarm and detection rates cannot be calculated. However, simulation results suggest is generally more stringent than , so its higher flag rate was expected.

To further examine the statistics' performance, we analyzed response patterns of 15 flagged examinees across three categories (Table 9). Simultaneously flagged examinees showed extreme scores inconsistent with their estimated abilities (e.g., low-ability examinees scoring 2 on multiple items; high-ability examinees scoring 0 on easy items). Examinees flagged only by were predominantly low-ability with many 0s but occasional high scores—classic aberrant high-performance patterns, consistent with 's design. Those flagged only by showed more "uniform" patterns, indicating 's broader coverage across aberrant types compared to 's specificity for low-ability/high-performance patterns.

6. Discussion and Future Directions

This study proposed a residual-based person-fit statistic with advantages in detecting low-ability examinees exhibiting aberrant high performance. In the polytomous context, we compared and 's empirical distributions (positively skewed, negatively skewed), with skewness decreasing as test length increased. Simulation studies established empirical critical values at various Type I error rates to distinguish normal from aberrant examinees. By simulating different aberrant behaviors and severities, we evaluated detection rates and AUC. Although Study 2 presented results for three Type I error rates (0.01, 0.025, 0.05), practical applications require balancing error control with detection power; results suggest $\alpha = 0.05$ is optimal.

Key findings: At low-to-moderate aberrance (few items affected), outperforms , especially for lucky guessing—consistent with 's theoretical sensitivity to aberrant high performance. Under other conditions, performance is similar or slightly better. Note that 's actual Type I error rate tends to be slightly higher than 's, inflating its detection rate accordingly. Similarly, Type I error rates in simulation studies often show some inflation, so true detection rates may be slightly lower than reported. For AUC (independent of Type I error), 's advantage is more pronounced.

In practice, has high utility because low-ability examinees are more likely to actively engage in aberrant behaviors (cheating, guessing) to gain high-stakes benefits, often causing test security breaches. High-ability examinees' aberrant behaviors are typically passive and less damaging. Rupp (2013) systematically reviewed aberrant behavior research, noting that low-ability cheating and guessing receive more attention as widespread, critical problems. However, PFS effectiveness depends on item parameters; when parameters are unknown, estimation with aberrant data produces masking effects. In such cases, robust estimation methods (Cooperman et al., 2021) or data cleaning (Hong et al., 2020) should be considered. Given 's characteristics, combining it with other statistics may yield optimal results.

Our simulations used five-category scoring, while the empirical study used three categories after merging, with consistent results. Future research should validate across more scoring levels. Additionally, since uses empirical critical values, future work could develop a standardized version with normal or asymptotically normal distribution, eliminating the need for test-specific critical values. Improvements could also retain information for classifying aberrant types or extend to broader IRT models. Finally, this study did not examine item parameter estimation error effects, an important consideration for broader applications.

References

Buchanan, T., & Smith, J. L. (1999). Using the internet for psychological research: Personality testing on the world wide web. *British Journal of Psy-*

chology, 90(1), 125–144.

Chen, Q., Ding, S., Zhu, L., & Xu, Z. (2010). Three-parameter graded response model and its parameter estimation. *Journal of Jiangxi Normal University (Natural Science), 34*(2), 117–122. [陈青, 丁树良, 朱隆尹, 许志勇. (2010). 三参数等级反应模型及其参数估计. 江西师范大学学报 (自然科学版), 34(2), 117–122.]

Cheng, X., Ding, S., Zhu, L., & Wu, H. (2012). The stratified item selection strategy with maximal information under graded response model. *Journal of Jiangxi Normal University (Natural Science), 36*(5), 446–451. [程小扬, 丁树良, 朱隆尹, 巫华芳. (2012). 等级评分模型下的最大信息量分层选题策略. 江西师范大学学报 (自然科学版), 36(5), 446–451.]

Cooperman, A. W., Weiss, D. J., & Wang, C. (2021). Robustness of adaptive measurement of change to item parameter estimation error. *Educational and Psychological Measurement*, Advance online publication.

Curran, P. G., Kotrba, L., & Denison, D. (2010, April). Careless responding in surveys: Applying traditional techniques to organizational settings. Paper presented at the 25th annual conference of the Society for Industrial/Organizational Psychology, Atlanta, GA.

Dodd, B. G., Ayala, R. J. De., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5–22.

Doval, E., & Delicado, P. (2020). Identifying and classifying aberrant response patterns through functional data analysis. *Journal of Educational and Behavioral Statistics, 45*(6), 719–749.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement, 28*(1), 105–113.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*(1), 59–80.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*(2), 159–177.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224–247.

Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association, 88*(422), 515–519.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika, 72*(2), 159–180.

Gulliksen, H. (1950). *Theory of mental tests.* John Wiley & Sons Inc.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer et al., *Measurement and Prediction* (pp. 60–90). Princeton: Princeton University Press.

Harris, K. M., & Udry, J. R. (2010). *National Longitudinal Study of Adolescent Health (Add Health), 1994–2008: Core files [restricted use]* (Technical report). Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Hong, S., Steedle, J., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement, 80*(2), 312–345.

Hotaka, Maeda. (2017). Robust latent ability estimation based on item response information and model fit (Dissertation). Milwaukee.

Huang, J. L., Bowling, N. A., Liu, M. Q., & Li, Y. H. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*, 299–311.

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

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

Li, J., & Ding, S. (2018). The several stratified methods of CAT in the presence of calibration error on GRM. *Journal of Jiangxi Normal University (Natural Science), 42*(4), 374–378. [李佳, 丁树良. (2018). 基于 GRM 模型的 CAT 分层方法在校准误差中的应用研究. 江西师范大学学报 (自然科学版), 42(4), 374–378.]

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(4), 29–37.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 101–121). New York, NY: Springer.

Meade, A. W. (2016). Understanding and detecting careless responding in survey research. Retrieved from https://cba.unl.edu/outreach/carma/documents/CARMA-Meade-Presentation.pdf

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455.

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135.

Meijer, R., & Sijtsma, K. (2001). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8*(3), 261–272.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*(2), 121–129.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*(3), 200–219.

Rogers, H. J., & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47–57.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(4), 1–97.

Schnipke, D. L. (1996). How contaminated by guessing are item-parameter estimates and what can be done about it? Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–238.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika, 81*(4), 1118–1141.

Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika, 81*(4), 992–1013.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342.

van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*(3), 273–282.

van Krimpen-Stoop, 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.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40*(4), 307–330.

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

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

Xiong, J., Luo, H., Wang, X., & Ding, S. (2018). The online calibration based on graded response model. *Journal of Jiangxi Normal University (Natural Science), 42*(1), 62–66. [熊建华, 罗慧, 王晓庆, 丁树良. (2018). 基于 GRM 的在线校准研究. 江西师范大学学报 (自然科学版), 42(1), 62–66.]

Xiong, J., Ding, S., Luo, F., & Luo, Z. (2020). Online calibration of polytomous items under the graded response model. *Frontiers in Psychology, 10*, 3085.

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods, 24*(5), 658–674.

Yuan, K. H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology, 38*(1), 329–388.

Appendix A: Comparison of and Under Five-Category Scoring

Table A-1 False alarm and detection rates for and (test length = 40 items)

Aberrant Type	Critical Value	Low (0.1)	Moderate (0.25)	High (0.5)
Cheating				
Lucky Guessing				
Random				
Careless				
Creative				
Mixed				

Table A-2 False alarm and detection rates for and (test length = 80 items)

Aberrant Type	Critical Value	Low (0.1)	Moderate (0.25)	High (0.5)
Cheating				
Lucky Guessing				
Random				
Careless				

Aberrant Type	Critical Value	Low (0.1)	Moderate (0.25)	High (0.5)
Creative				
Mixed				

Note: Figure translations are in progress. See original paper for figures.