Educenter : Jurnal Ilmiah Pendidikan
Vol 3 No 1 January 2024
ISSN: 2827-8542 (Print) ISSN: 2827-7988 (Electronic)
Open Access: https://jurnal.arkainstitute.co.id/index.php/educenter/index

# Identification of cheating methods in large-scale tests: Islamic education subject matter in the national examination

**Ahmad Niayatulloh[1*], Muhammad Haikal[2]**
[1,2]Universitas Darussalam Gontor, Indonesia
[1]ahmad.niayatulloh@unida.gontor.ac.id, [2]muhammad.haikal@unida.gontor.ac.id

**Article Info**

***Article history:***
Received October, 15th 2023
Revised December, 17th 2023
Accepted January, 25th 2024

***Keyword:***
Large-scale Test; National Exam; The Item Characteristics; The Cheating Identification

**ABSTRACT**

This study aims to determine item characteristics based on item response theory and identify cheating during the administration of the National Standardised School Examination (USBN) for Islamic religious education at the senior high school level in the 2015/2016 school year in Yogyakarta. The data source consists of 2929 answer sheets from USBN participants in the subject of Islamic Religious Education (PAI) under Package A of the KTSP curriculum. Using a quantitative approach with an ex post facto design, the analysis uses item response theory for item characteristics and person fit and test acceptance methods for identifying cheating. The results show that the PAI USBN test instrument for the 2015/2016 academic year has a moderate average level of difficulty, with a good information function. The Test Acceptance method, which identified 153 cases of knowledge sharing, 891 cases of ignorance sharing and 222 cases of response sharing, proved to be more effective in detecting cheating than the Person Fit method, which identified 2 individuals.



## INTRODUCTION

The position of Islamic religious education is of utmost importance in nurturing students to become believers, mindful of God, and possessors of noble character. The explicit role of religious education is clearly stated in Article 37 of Law Number 20 of 2003, emphasizing its coverage from primary to higher education. It aims to equip students with the knowledge and expertise to fulfill their religious duties and become scholars of religious sciences (Franck, 2021). Consequently, Islamic religious education plays a vital role in accelerating the attainment of national educational objectives. It is thus understandable that the values inherent in Islamic teachings should be integrated into all subjects, allowing for the internalization of these teachings in the learning process for students. Hence, the school curriculum rightly prioritizes placing religious education at the forefront.

The challenge faced in implementing religious education is ensuring that Islamic religious education not only imparts knowledge about religion but also directs students to become individuals who genuinely possess solid religious qualities. Therefore, educational material should not only focus on acquiring knowledge but also on shaping the attitudes and personalities of the students so that they become complete human beings (*Insan al-kamil*) with true faith, piety, and good moral character (*Akhlakul karimah*). There seems to be a phenomenon of shifting values resulting from moral degradation.

One of the values that is gradually eroding among students is the value of honesty (integrity). Integrity is paramount in the academic realm, serving as a character strength applicable to various facets of life, including education, research, and work (Bin-Nashwan et al., 2023). A significant factor undermining the integrity of education is the pervasive culture of cheating (Brimble, 2016). According to Kohlberg's theory of moral development, cheating behavior is closely tied to forming moral codes. Individuals cheat because they perceive it as forgivable and normalized, feeling pressured to achieve high grades for admission to higher-level schools (Anderman & Koenka, 2017).

Academic dishonesty poses a formidable challenge within the educational sphere, impeding the realization of educational objectives. Academic dishonesty, especially during examinations, significantly compromises the validity of student assessments (Winardi et al., 2017). The prevailing societal mindset, which places greater emphasis on students' achievements rather than the process of attaining them, exacerbates this issue. Students are compelled to prioritize high grades in tests over

genuine knowledge acquisition, leading to a considerable number of students resorting to cheating as a means to secure elevated grades and expedite exam completion.

Relevant research on the detection of cheating through answer copying has been conducted by Manoppo & Mardapi (2014). Their study aimed to identify instances of cheating behavior among students attending public high schools in the Maluku Province during the 2011/2012 academic year. The findings revealed a startling revelation. Out of the 1,620 sampled students, an overwhelming 1,556 individuals were found to have engaged in cheating, translating to a percentage of 96.04%. This figure underscores the issue's magnitude, highlighting its significance and necessitating immediate attention.

A study by Herwin & Heriyati (2016) focused on identifying cheating among exam participants using the person-fit method. Based on this method, the research objective was to analyze and describe the response patterns of school exam participants in Soppeng Regency, South Sulawesi. This quantitative study specifically examined 40 multiple-choice mathematics items in the elementary school exam, along with the answer sheets of 125 participants. The research findings revealed that most participants (approximately 64% or 80 individuals) demonstrated response patterns deemed fit, characterized by logical and consistent responses without indications of cheating.

In a separate study by (Kusaeri, 2017) titled "A Study of Cheating Behavior among Madrasah and Islamic School Students during National Examinations," the focus shifted to uncovering instances of cheating among Islamic high school students in East Java. Kusaeri aimed to assess cheating based on the student's performance in the National Examination (UN) and the Islamic Education National Examination (IIUN) in 2015. The cheating index was determined using pairwise and cumulative methods. The research findings highlighted a contrasting trend between Islamic and Christian/Catholic high school students. Around 5.26% of Islamic high school students in East Java met the passing criteria for the UN (above 55) with an IIUN score exceeding 70. In contrast, approximately 40% of Christian/Catholic high school students achieved UN scores above 55, with IIUN scores surpassing 70. These findings shed light on a higher prevalence of cheating among Islamic high school students in East Java compared to their Christian/Catholic counterparts.

Given the intriguing nature of the above phenomena, further analysis of cheating behaviors becomes imperative. Hence, the researchers became interested in identifying such behaviors during the National Examination for Islamic Religious Education (USBN PAI) among students in public high schools located in Yogyakarta. The study will utilize the person fit method and test acceptance as the main approaches to understanding the prevalence and characteristics of cheating within this context.

## RESEARCH METHODS

This study adopts a quantitative research design employing an ex-post facto approach. This approach aims to investigate the impact of treatment and explore the underlying factors that contribute to it. The data collection method utilized is documentation, specifically gathering student response answers from the USBN PAI (National Examination for Islamic Religious Education) test devices based on the KTSP curriculum, Package A, administered in public high schools in the Special Region of Yogyakarta. The answer sheets (LJK) data were obtained from the Department of Religion in Yogyakarta.

The population for this study comprises all answer sheets of participants in the USBN PAI test based on the KTSP curriculum, Package A, during the 2015-2016 academic year in the Special Region of Yogyakarta. The total number of students included in the study amounts to 2,929, encompassing students from various regional public high schools. Data analysis commences with describing the suitability (characteristics) of the USBN PAI test at the high school level, based on the KTSP curriculum, Package A. This is accomplished by applying a modern approach employing the Rasch model facilitated by the WINSTEPS software.

The analysis, rooted in the modern approach, begins by verifying the assumptions of Item Response Theory (IRT). These assumptions include dimensionality, local independence, and parameter invariance (item and ability parameters) (Pardede et al., 2023). The analysis of item characteristics in this approach is determined as follows:

a) Each item and the distribution of participant responses should align with the model.
b) The estimated difficulty level of each item should range from -2 logits to 2 logits (Hambleton & Swaminathan in Istiyono, 2016).

`

c) The test provides valuable information if the Test Information Function (TIF) is greater than or equal to 10.

To assess the fit of items and participant response patterns based on the Rasch model, the outfit mean square (MNSQ) values can be examined. An item fits the model if its MNSQ value falls between 0.5 and 1.5 (Rahman et al., 2023).

Furthermore, the identification of cheating through the person fit method in this study utilized an index based on item response theory, precisely the logistic approach known as the one-parameter logistic model (Rasch Model). Person fit can be determined by analyzing the outfit statistic in the WINSTEPS output. The outfit statistic for a person indicates unexpected behavior exhibited by items with difficulty levels significantly different from the participant's ability (Ruijten et al., 2019). Outfit is based on the sum of squared standard residuals using conventional methods. Let X represent the observation, E denotes the expected value based on Rasch parameter estimation, and $\sigma^2$ represent the model's variance around its expectation (Qiu et al., 2021). The squared standard residual can then be calculated as:

**The squared standard residual**

$$Z^2 = \frac{X-E^2}{\sigma^2} \text{ to outfit } = \frac{\sum Z^2}{N}, (1)$$

Where:

$N$ : Number of observations.
$X$ : Observed value
$E$ : Expected value
$\sigma^2$ : Model variance
$Z$ : Standard residual

In this study, the researcher focused on three potential causes: cheating, careless mistakes, and lucky guessing. This narrowed focus was chosen to facilitate a more targeted identification of cheating using the person-fit method.

To examine these factors, a simulation was conducted using three distinct score patterns, each representing deviant behavior. These patterns were constructed based on 12 fictitious multiple-choice items, comprising four easy items (Items 1 to 4), four moderately complex items (Items 5 to 8), and four challenging items (Items 9 to 12). Using this simulated data, the researcher aimed to gain insights into the impact of different types of behavior on the person fit analysis.

**Table 1. Simulation of Cheating Identification**

| No | Item | | | | | | | | | | | | Behavior |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | |
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | Careless responding |
| 2 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | Lucky Guessing |
| 3 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | Cheating |

The criteria for the ability parameter of a participant, denoted as θ in Item Response Theory (IRT), characterize the individual's ability. The item parameters are expressed through a suitable logistic model. Hambelton noted that the ability parameter (θ) exists within the interval $-\infty \leq \theta \leq \infty$ and is scaled to approach a normal distribution with a mean of 0 and a standard deviation of 1. However, in practical terms, an individual's ability typically falls within $-3 \leq \theta \leq 3$. In this study, the researcher categorized the ability parameters into three groups to determine participant ability grouping (Hambleton & Swaminathan in Istiyono, 2016). The categorization is based on using the standard distribution rule, which theoretically spans a distance of 6. This approach was employed to derive empirical categorization results for participant abilities (Elo et al., 2014).

**Table 2. Categorization of Abilities into 3 Categories with 6 SD**

| No | Interval | Category |
|---|---|---|
| 1 | ( Mi + SD) < x ≤( Mi +3SD) | High |

`

| No | Interval | Category |
|---|---|---|
| 2 | > (Mi-1SD) < x ≤ (Mi+1SD) | Medium |
| 3 | (Mi-3SD) ≥ x ≤( Mi – 1SD) | Low |

Furthermore, the criteria for item difficulty in IRT range from $-\infty \leq b \leq \infty$ on the item response theory scale. However, in practice, items that are considered good have a difficulty level (b_i) ranging from $-2 \leq b \leq +2$. Items with difficulty levels (b_i) close to or above the +2.00 scale indicate that the items fall into the difficult category (Donati et al., 2021).

**Table 3. Design of Cheating Identification using Person Fit**

| Types of Anomalies | Ability | Item Criteria | Identification Criteria |
|---|---|---|---|
| Cheating | Low | Very difficult | Many correct answers on difficult items |
| Lucky guessing | Low | Very difficult | Correct on the most difficult item |
| Careless | High | Very easy | High frequency of incorrect responses (quantity > 1) |

The test acceptance testing is determined by utilizing empirical data obtained from participants who have taken the test. It involves comparing each student's answer responses with those of other students, a process known as pairwise comparison. These pairwise comparisons result in an empirical distribution, representing the combinations of all pairs. The pairs that exhibit high similarity, falling within the prominent cluster or group of responses in this empirical distribution, are considered acceptable. Conversely, pairs that deviate significantly (outliers) and display a notable dissimilarity in their answers indicate unacceptable similarity. Such outliers may arise due to misalignment (writing errors), guessing, or other variable behaviors.

The formula to calculate the total number of pairwise comparisons generated within a group of students can be expressed as follows (Kingsdorf & Krawec, 2014).

**The Total Number of Pairwise**

$$Pairwise = \frac{n \times (n-1)}{2} \qquad (2)$$

Where n = the number of students in each group.

In test acceptance, there are three forms of cheating detection: detecting shared knowledge (copying correct answers), shared ignorance (copying incorrect answers), and share response (copying both correct and incorrect answers simultaneously). To determine the percentage of share knowledge, share ignorance, and share response, the examination of identical correct answer strings and identical incorrect answer strings is conducted. The formulas for calculating these percentages can be found in (Conway et al., 2019).

**Share Ignorance**

$$Share\ knowledge = \frac{Twohigh}{Onehigh} \times 100\ \% \qquad (3)$$

Note :

*Twohigh* : The number of items with identical correct answers

*Onehigh* : The number of items answered correctly by either one or both pairs (pairwise)

`

$$Share\ Ignorance = \frac{Twolow}{Onehow} \times 100 \quad (4)$$

Note :
Twolow : The number of items with the same incorrect answer (same distractor)
Onelow : The number of items answered incorrectly by either one or both pairs (regardless of different distractors)

Furthermore, to identify cheating by copying both correct answers (share knowledge) and incorrect answers (share ignorance) simultaneously, the following formula can be used (Conway et al., 2019).

**Share Respons**

$$Share\ Respons = \frac{Twosame}{Twoobs} \times 100\ \% \quad (5)$$

Where :
Twosame : The count of items for which both correct and incorrect responses are the same
Twoobs : The count of items attempted by both pairs.

Outlier detection can be performed by determining a threshold value that will be categorized as outlier data by converting data values into standardized scores, commonly known as z-scores. Z-score has a mean value equal to zero and a standard deviation equal to one. Therefore, the z-score is a standardized score representing the difference between an individual's score and the group mean divided by the standard deviation. In theory, to obtain the Z value, the formula is as follows:

**Z Score**

$$z - score = \frac{(X - \bar{X})}{SD} \quad (6)$$

Where:
$X$ = Observation value at index i
$\bar{X}$ = Mean of the observation values
$SD$ = Standard deviation of the observation values

When the values are expressed in standard format (Z-score), comparisons between different value magnitudes can easily be made. For small sample cases (less than 80), a Z-score with a value greater than 2.5 is considered an outlier. A Z-score is considered an outlier for larger samples if its value falls in the range of 3 to 4. For substantial sample sizes (above 80 observations), the evaluation guideline is that the threshold Z-score ranges from 3 to 4. Therefore, cases or observations with a Z-score greater than 3.00 are outliers. In this study, where the observed data exceeds 80 observations (pairs), the researcher employs the outlier classification if the Z-score is > 3 (Mowbray et al., 2019).

The procedure for identifying cheating using the test acceptance method can be presented in the form of Table 4 below:

**Tabel 4. Identification Design for Cheating in the Test Acceptance Method**

| Cheating Type | Formula | Score |
|---|---|---|
| *Share knowledge* | $Share\ knowledge = \frac{Twohigh}{Onehigh} \times 100\ \%$ | Z-score > 3 |

`

| Cheating Type | Formula | Score |
|---|---|---|
| *Share ignorance* | $Share\ Ignorance = \frac{Twolow}{Onehow} \times 100\ \%$ | Z-score > 3 |
| *Share respons* | $Share\ response\ = \frac{Twosame}{twoobs} \times 100\ \%$ | Z-score > 3 |

## RESULTS AND DISCUSSION

### Requirements for IRT Analysis

The research results indicate that the KMO-MSA value is 0.867, and Bartlett's test is significant at 0.000. The KMO-MSA value generated from the USBN PAI test instrument meets the requirements for factor analysis. This aligns with the conditions for factor analysis, where KMO-MSA > 0.5 and Sig. Bartlet's test < 0.05, as shown in the table. Factor analysis can proceed since the KMO-MSA and Bartlett's test requirements have been met. The KMO MSA and Bartlett's Test of Sphericity results indicate this.

**Table 5. Results of KMO MSA and Bartlett's Test of Sphericity**

| KMO and Bartlett's Test | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,867 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 8920,524 |
| | df | 1225 |
| | Sig. | ,000 |

According to (Basinska & Dåderman, 2023), the unidimensionality test is satisfied if the test measures a single dimension that assesses the same ability. An extraction process is performed to obtain items that measure the same dimension, resulting in several factors. Each formed factor has an eigenvalue, and factors with eigenvalues above 1.00 are retained. The eigenvalue values for the USBN PAI test can be seen in the figure.

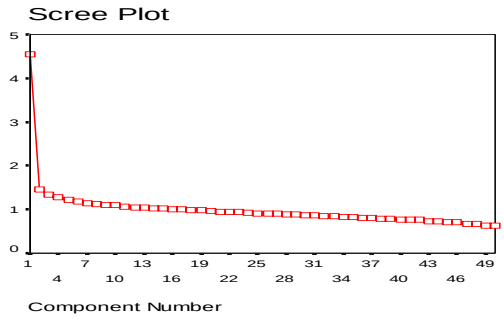


**Figure 1. Scree Plot of USBN PAI Scores**

Figure 1 showed that Factor 1 is far from Factor 2, indicating that Factor 1 is several times larger than Factor 2. This implies that the USBN PAI test is unidimensional. The assumption of local independence has been proven, as demonstrated by the unidimensionality of participant response data (Retnowati in Alfarisa & Purnama, 2019). This explanation can be interpreted as confirming the unidimensionality assumption, as presented in Table 9 and Figure 4. Therefore, the assumption of local independence has also been fulfilled.

The invariance parameter test aims to determine whether item characteristics remain unchanged even when answered by different groups of students. Similarly, the ability estimates will not change for the same group of students even if the item questions vary. If the correlation is positive and high, then the assumption of invariance of item parameters is met (Retnowati in Alfarisa & Purnama, 2019).

`

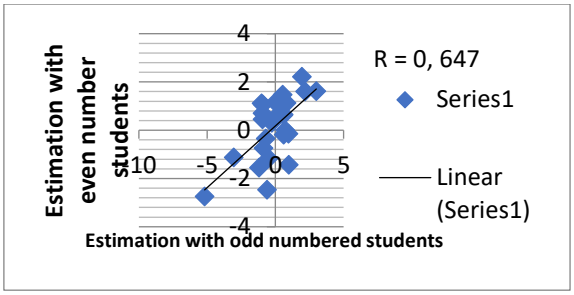


**Figure 2. Plot of Item Difficulty Parameter Invariance**

Figure 2 is a plot of estimates for the invariance of item parameters. Upon closer inspection of figure 2, it can be observed that the estimation values are relatively close to the straight line with a sufficiently high correlation value (0.647). Therefore, it can be concluded that the assumption of invariance of item parameters has been met.

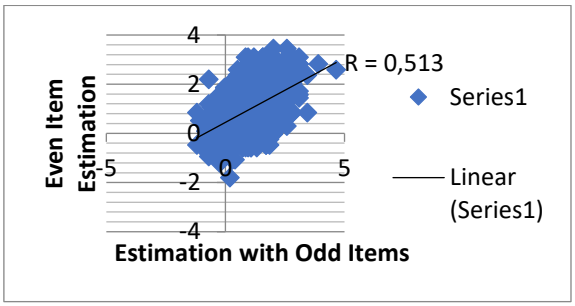


**Figure 3. Plot of Ability Parameter Invariance**

Figure 3 is a scatter plot of ability parameters based on the group of items answered by students. Based on this figure, the estimation values are relatively close to the straight line with a good (substantial) correlation value of 0.513. Therefore, the assumption of invariance of ability parameters has also been met.

The item difficulty and participant ability levels are considered suitable for the model if the OUTFIT MNSQ values fall within the range of 0.5 – 1.5 (Pitaloka et al., 2023). According to this criterion, all items in the USBN PAI test are suitable for the model, and approximately 7% of the 2929 test participants do not fit the Rasch model because they fall outside the specified Outfit MNSQ range. Figure 7 presents the level of individual fit to the model.

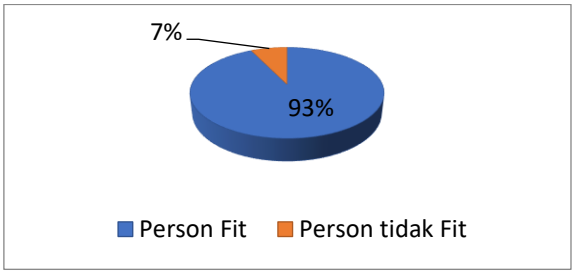


**Figure 4. Distribution of Individual Fit to the Model**

Based on the item difficulty criteria, the results show that out of 50 items, there are 7 items (14%) that fall into the category of not good. These items are 45, 36, 21, 48, 26, 31, and 25. The item with the highest difficulty level is item 45 with a difficulty level of +3.00 logits, while the easiest item is item 25 with a difficulty level of -5.18 logits.

`

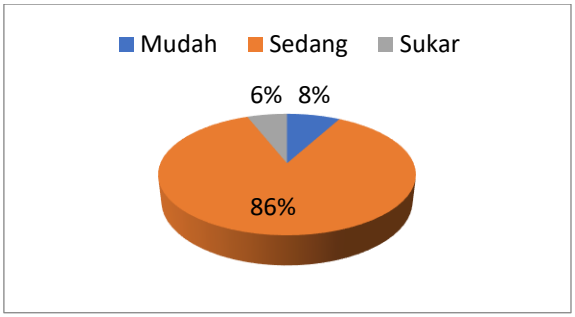


**Figure 5. Distribution of Item Difficulty Levels Based on IRT**

The calculation results show a maximum information function value of 9.35 logits at θ around +0.2. Meanwhile, the Standard Error of Measurement (SEM) for this test is 0.327. SEM is inversely proportional to the test information function. This means that the test will provide good information, with the most minor measurement error being 0.327 when taken by test participants with an ability level of around +0.2 logits. The graph depicts the relationship between item information function and SEM.

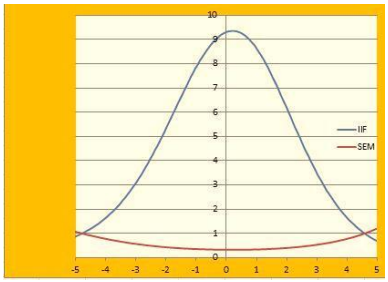


**Figure 6. Relationship between TIF and SEM**

Figure 6 presents the information function values of a USBN PAI test with 50 items. This test instrument has higher information values compared to its measurement error. The USBN PAI test is suitable for measuring students in the ability range of -4.78 to +4.52. If the questions are tested on participants with ability scales less than -4.78 and more than +4.52, the result will be a more significant measurement error compared to its information function value, identifikasi Cheating Metode Person Fit

The first parameter analyzed is the level of person-fit based on Winsteps output. By examining the outfit MNSQ values within the range of 0.5 – 1.5 (Pitaloka et al., 2023).

**Table 5. Results of Person Fit Analysis Using OUTFIT MNSQ**

| No | OUTFIT MNSQ Range | Status | Number | % |
|---|---|---|---|---|
| 1 | 0,5 < MNSQ < 1,5 | ***Fit*** | 2703 | 93 % |
| 2 | 0,5 ≤ MNSQ ≥ 1,5 | ***Misfit*** | 210 | 7 % |

In the analysis of the USBN PAI in Yogyakarta, it was found that 7% of respondents (210 individuals) had abnormal scores (misfit) out of 2929 respondents.

The second parameter analyzed is the difficulty level of the test items. The distribution of the difficulty level of the items and the ability of the respondents can be seen in Table 11 item measures. For each item, the estimated difficulty level ranges from -2 logits to 2 logits.

**Table 6. Results of Item Difficulty Level Estimation**

| No | Criteria | Item Numbers | Number | Percentage |
|---|---|---|---|---|
| 1 | Difficult ( $b > 2$) | 21, 36, 45 | 3 | 6 % |
| 2 | Moderate ( $-2 \leq b \leq 2$ ) | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 27, 28, 29, 30, 32, 33, 34, 35, | 43 | 86 % |

`

| No | Criteria | Item Numbers | Number | Percentage |
|---|---|---|---|---|
| | | 37, 38, 39, 40, 41, 42, 43, 44, 46, 47, 49, 50 | | |
| 3 | Easy ( $b < -2$) | 25, 26, 31, 48 | 4 | 8 % |

The table above showed that the USBN PAI items in Yogyakarta classified as difficult, with values greater than 2, are three items. Moderate difficulty items, with values ranging from -2 to 2, are 43. Meanwhile, easy items with values less than -2 are four items. The next step is to estimate the ability of the respondents.

The estimation of respondent abilities can be seen in the theta file at the person-measure value. In this study, 210 respondents are classified as misfit MNSQ among the participants of the USBN PAI in Yogyakarta. Subsequently, the researcher divided them into three categories based on the theta value. The categorization is determined using the standard distribution rule, theoretically spanning 6 SD. This is done to obtain empirical categories of participant abilities.

**Table 7. Results of Distributing Student Abilities Based on 3 Categories**

| No | Interval | Criteria | Number of Students | Percentage |
|---|---|---|---|---|
| 1 | ($\theta \leq 0.076$) | Low | 36 students | 17.15% |
| 2 | ($0.076 < \theta \leq 2.314$) | Moderate | 143 students | 68.09% |
| 3 | ($\theta > 2.314$) | High | 31 students | 14.76% |

Based on the table 7, the analysis results indicate that out of 210 students, 36 students (17.15%) have high abilities with $\theta > 2.308$ criteria, and 143 students (68.09%) have moderate abilities with criteria $0{,}071 < \theta \leq 2{,}308$ , dan 31 siswa (14,76 %) having low abilities with the criteria $b\theta \leq 0{,}071$.

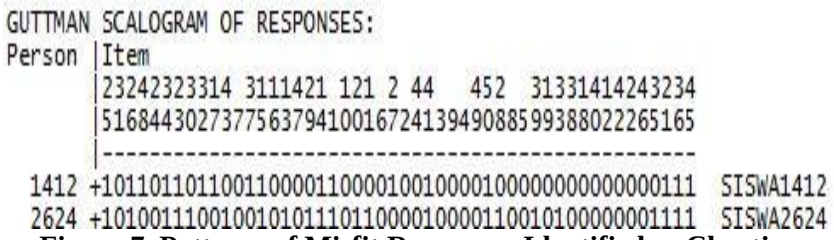


**Figure 7. Patterns of Misfit Responses Identified as Cheating**

Based on the figure 7, participants 1412 and 2624 have low abilities. They manage to overcome some of the easy items initially but make many mistakes afterward. They give up and then resort to cheating on the difficult items, successfully handling three items with high difficulty.

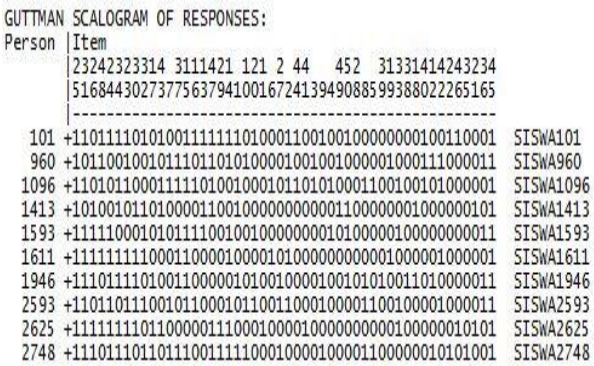


**Figure 8. Patterns of Misfit Responses Identified as Lucky Guessing**

Based on the Guttman matrix in figure 8, ten students are identified as engaging in lucky guessing. These students are identified with the IDs 101, 960, 1096, 1413, 1593, 1611, 1946, 2593, 2625, and 2748. All ten students have low abilities, i.e., $b \leq 0.076$. They are identified as engaging in

`

lucky guessing because, with their low abilities, they can answer item number 45, which is the most difficult item. However, for items with relatively lower difficulty, they provide many incorrect responses. Therefore, unexpectedly, they can give correct responses to some difficult items. No students are identified as careless.

**Table 8. Results of Cheating Identification Using Person Fit Method**

| Response Pattern | Ability | Item Difficulty | Identification Criteria | Count |
|---|---|---|---|---|
| Cheating | Low ($\theta \leq 0.076$) | Difficult ($b > 2$) | Consistently correct answers on difficult items (correct on three consecutive most difficult items) | 2 |
| Lucky Guessing | Low ($\theta \leq 0.076$) | Difficult ($b > 2$) | Correct on the most difficult item | 10 |
| Careless | High ($\theta > 2.314$) | Easy ($b < -2$) | Numerous incorrect responses (count > 1) | - |

Based on the data presented in the table above and using a scalogram or Guttman matrix where each item has a systematic sequence from the easiest to the most difficult item, it can be explained that there are 10 students identified with a Lucky Guessing response pattern. The identification of Lucky Guessing response patterns occurs in students who have low ability ($\theta \leq 0.076$) but unexpectedly provide correct responses to the most difficult items ($b > 2$). There are no students identified as Careless, and 2 students are identified as Cheating.

**Indentifikasi Cheating Metode Test Acceptence**

Test acceptance testing is determined by using empirical data from participants who take the test. First, each student's response is compared with other students' responses. That is called pairwise comparison. Pairs from pairwise comparisons will form an empirical distribution depicting all pairs' formation. The results of the number of pairwise comparisons at each school are presented in the following figure:

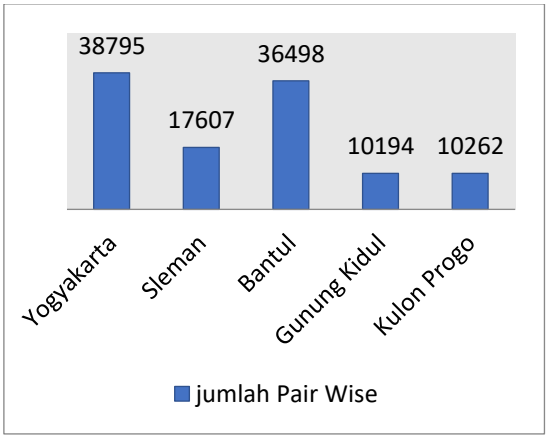


**Figure 9. The number of pair-wise comparisons for each school in each district/city.**

Based on the above figure, the results for each school pairwise are summed in their respective districts. The total pairwise for public Senior High Schools (SMAN) in Yogyakarta is 38,795 pairs of students; in SMAN Sleman district, there are 17,607 pairs of students, while in SMAN Bantul district with a total of 36,498 students, in SMAN Gunung Kidul district, there are 10,194 pairs of students, and the total pairwise for SMAN in Kulon Progo district is 10,262.

Each pairwise connection links the ability of exam participants with the average ability value in each pairwise as a combination of each pair. The "Outlier" points in this empirical distribution indicate a pair of exam participants with an exceptionally large (unusual) percentage of shared responses

`

for their ability levels. In this acceptance test, there are three forms of cheating detection, namely detecting shared knowledge (copying correct answers), sharing ignorance (copying incorrect answers), and sharing response (copying both correct and incorrect answers simultaneously).

Identification of cheating through shared knowledge is done by checking the similarity string of correct answers in each pairwise. The results of shared knowledge on the USBN PAI test at SMAN 1 Piyungan, Bantul district, can be acknowledged by looking at the output results table 35 in WinSteps, which are as follows:

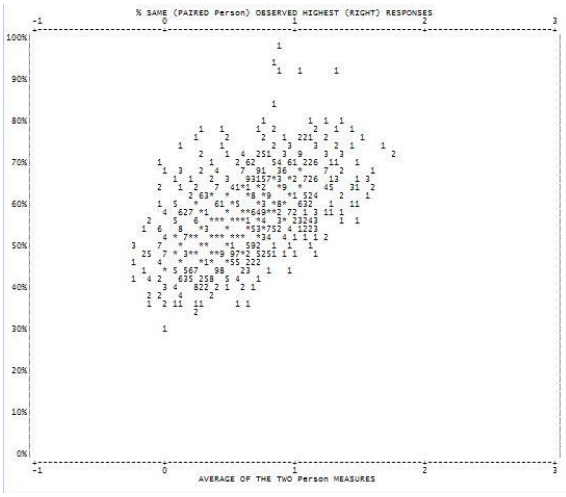


**Figure 10. Plot of share knowledge results for SMAN 1 Piyungan**

The above figure, the x-axis, represents the average ability between two students in a pair-wise comparison. In contrast, the y-axis represents the percentage of similarity or string of matching responses from the same correct answers multiplied by one hundred and divided by the total number of items answered correctly by one or both pairs. Values plotted between 1 and 9 represent the number of pairs falling on the x and y coordinates, while the asterisk (*) indicates the number of pairs exceeding 9.

Identification of outliers in the results of shared knowledge in the above figure is done using z-scores. The threshold for the z-score value to be considered an outlier is if the z-score value is greater than 3.00. To view the z-score values, SPSS software can be used. The results of the z-score values for pair-wise sharing knowledge at SMAN 1 Piyungan are as follows:

**Table 9. Outliers Result of Share Knowledge at SMAN 1 Piyungan**

| Percentage | Student 1 | Student 2 | Z-Score | Indication |
|---|---|---|---|---|
| 97% | 2214 | 2215 | 4.70934 | Outliers |
| 94% | 2215 | 2216 | 4.37156 | Outliers |
| 92% | 2208 | 2211 | 4.14044 | Outliers |
| 92% | 2174 | 2178 | 4.09007 | Outliers |
| 91% | 2214 | 2216 | 4.06272 | Outliers |
| 83% | 2200 | 2203 | 3.13301 | Outliers |
| 81% | 2178 | 2179 | 2.81399 | Not |
| 80% | 2170 | 2190 | 2.80621 | Not |
| 80% | 2211 | 2214 | 2.75018 | Not |
| 79% | 2210 | 2211 | 2.69129 | Not |
| 79% | 2213 | 2216 | 2.61098 | Not |

Table 9 shows the results of the z-score values for pairwise sharing knowledge in SMAN 1 Piyungan. The total number of students in SMAN 1 Piyungan is 59, resulting in 1711 pairwise comparisons (the specific number of pairwise comparisons for each school can be seen in the attachment). Out of the 1711 pairwise comparisons, six pairs were detected as outliers with z-score values exceeding 3 in the share knowledge plot. This is because these six pairs fall outside the acceptable limits of matching correct responses, linked to the average ability level of each pairwise

`

comparison, or can be considered abnormal pairs. Therefore, the six pairs identified as outliers are unacceptable altogether.

Identification of cheating through shared ignorance involves examining the similarity of incorrect answers in each pairwise comparison. The results of share ignorance in the USBN PAI test at SMAN 1 Srandakan, Bantul Regency, can be observed in the output table 35 of Winsteps as follows:

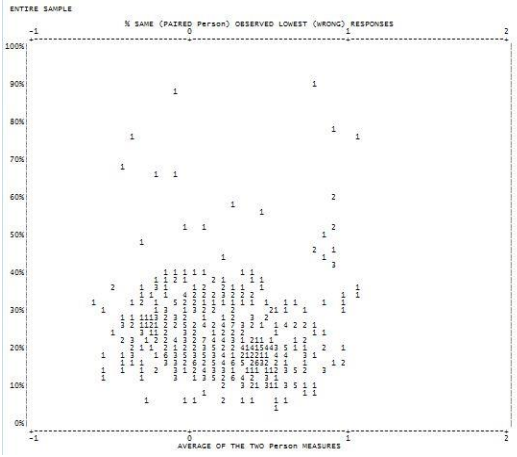


**Figure 11. Plot of Share Ignorance Results at SMAN 1 Srandakan**

The results of the z-score values for pairwise pairs in share ignorance at SMAN 1 Srandakan are as follows:

**Tabel 10. Hasil Outliers Share Ignorance SMAN 1 Srandakan**

| Percentage | Student 1 | Student 2 | Z-Score | Indication |
|---|---|---|---|---|
| 89% | 2502 | 2505 | 5.93046 | Outliers |
| 88% | 2489 | 2492 | 5.83981 | Outliers |
| 79% | 2497 | 2498 | 4.98769 | Outliers |
| 77% | 2491 | 2492 | 4.78343 | Outliers |
| 76% | 2499 | 2501 | 4.76586 | Outliers |
| 68% | 2511 | 2512 | 3.98410 | Outliers |
| 67% | 2489 | 2491 | 3.88779 | Outliers |
| 66% | 2512 | 2515 | 3.78485 | Outliers |
| 60% | 2499 | 2502 | 3.29071 | Outliers |
| 60% | 2499 | 2505 | 3.29071 | Outliers |
| 58% | 2493 | 2495 | 3.08402 | Outliers |
| 56% | 2494 | 2495 | 2.93245 | Not |
| 52% | 2501 | 2502 | 2.60832 | Not |

Table 10 above showed the results of z-score values for pairwise matches in share ignorance at SMAN 1 Srandakan. The total number of students at SMAN 1 Srandakan is 34, resulting in 561 pairwise matches. Among these 561 pairwise matches, 11 pairs are detected as outliers, with z-score values exceeding 3 in the share ignorance plot. This indicates that 11 pairs of students are suspected of cheating by copying incorrect answers.

Cheating identification through shared response is conducted by examining the similarity of correct and incorrect answers in each pairwise match. The results of the share response in the USBN PAI test at SMAN 1 Pleret, Bantul Regency, can be seen in the output table 35 Winsteps as follows:

`

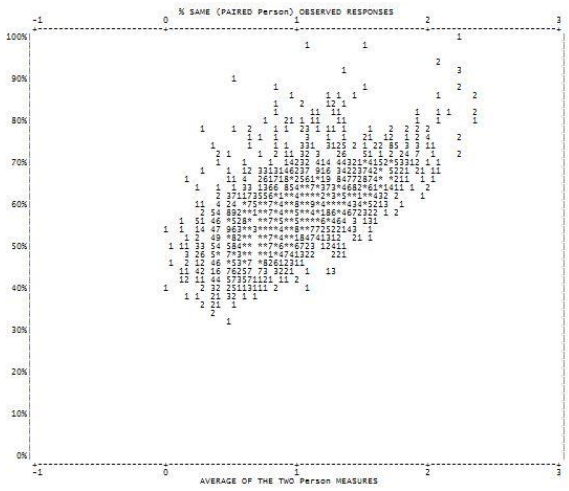


**Figure 12. Plot of Share Response Results at SMAN 1 Pleret**

The results of z-score values for pairwise matches in shared response at SMAN 1 Pleret are as follows:

**Table 12. Outliers Result of Share Response at SMAN 1 Pleret**

| Percentage | Student 1 | Student 2 | Z-Score | Indication |
|---|---|---|---|---|
| 100% | 2225 | 2228 | 4.88127 | Outliers |
| 98% | 2286 | 2287 | 4.64905 | Outliers |
| 98% | 2295 | 2297 | 4.64905 | Outliers |
| 94% | 2225 | 2232 | 4.18459 | Outliers |
| 94% | A2228 | 2232 | 4.18459 | Outliers |
| 92% | 2225 | 2230 | 3.95236 | Outliers |
| 92% | 2228 | 2230 | 3.95236 | Outliers |
| 92% | 2230 | 2231 | 3.95236 | Outliers |
| 92% | 2254 | 2255 | 3.95236 | Outliers |
| 90% | 2235 | 2239 | 3.72013 | Outliers |
| 88% | 2225 | 2231 | 3.48791 | Outliers |
| 88% | 2228 | 2231 | 3.48791 | Outliers |
| 88% | 2294 | 2298 | 3.48791 | Outliers |
| 88% | 2297 | 2300 | 3.48791 | Outliers |
| 86% | 2224 | 2225 | 3.25568 | Outliers |
| 86% | 2224 | 2228 | 3.25568 | Outliers |
| 86% | 2230 | 2232 | 3.25568 | Outliers |
| 86% | 2283 | 2286 | 3.25568 | Outliers |
| 86% | 2295 | 2300 | 3.25568 | Outliers |
| 86% | 2297 | 2298 | 3.25568 | Outliers |
| 86% | S2298 | 2300 | 3.25568 | Outliers |
| 84% | 2279 | 2282 | 3.02345 | Outliers |
| 84% | 2281 | 2282 | 3.02345 | Outliers |
| 84% | 2283 | 2287 | 3.02345 | Outliers |
| 84% | 2295 | 2298 | 3.02345 | Outliers |
| 84% | 2296 | 2300 | 3.02345 | Outliers |
| 84% | 2297 | 2299 | 3.02345 | Outliers |
| 84% | 2299 | 2300 | 3.02345 | Outliers |

The table above shows the results of the z-score values for pairwise response sharing at SMAN 1 Pleret. The number of students at SMAN 1 Pleret is 78, generating 3003 pairwise responses (the specific count for each school is available in the attachment). Of these 3003 pairwise responses, 28 pairs are identified as outliers with z-score values exceeding 3 in the share response plot. This indicates that 28 pairs of students are simultaneously suspected of cheating by copying correct and incorrect answers.

By estimating the difficulty level of items and the ability level of students who exhibit misfits, it is possible to determine the difficulty level of test items and the ability level of misfit students based on their respective categories. This information can be further used to identify misfit students detected

as cheaters. Based on the analysis of cheating identification, the diagram illustrating the results of cheating identification can be depicted as follows:

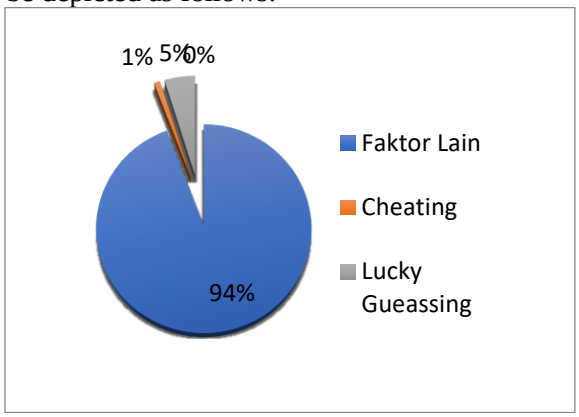


**Figure 13. Diagram Results of Cheating Identification Based on the Person Fit Method**

Based on the empirical findings from the analysis of 210 students identified as misfits, cheating identification using the person fit method in the USBN PAI participant identified two students, which accounts for 0.95% of the students. Additionally, ten students, or 4.76% of the total, were identified as lucky guessers, and no participants were identified as careless. The remaining 94% is attributed to other factors.

In this acceptance test, there are three forms of cheating detection: detecting shared knowledge (copying correct answers), shared ignorance (copying incorrect answers), and shared response (copying both correct and incorrect answers simultaneously). Thus, the acceptance test has three distinct analysis results. Overall, the diagram of cheating identification using the acceptance test method is as follows:

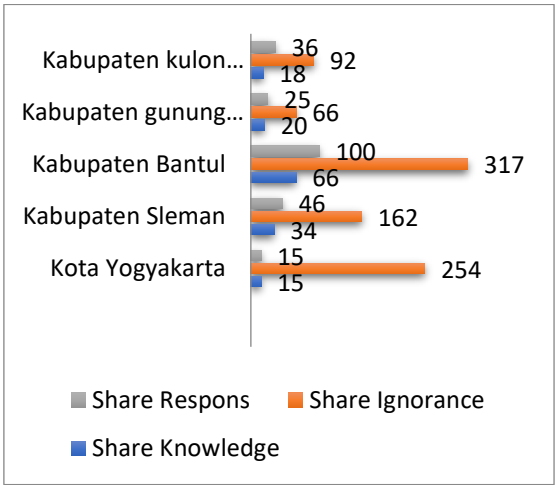


**Figure 14. The Number of Pairs Identified for Cheating using the Acceptance Method in USBN PAI in the Special Region of Yogyakarta**

The data above shows that the identification of cheating using the acceptance method in all three forms mainly occurs in the district areas. Bantul district is most identified in both Share Knowledge and Share Ignorance, with 100 pairs of Share Knowledge and 317 pairs of Share Ignorance. Furthermore, the Gunung Kidul district is most identified in Share Response with 66 pairs.

Based on tables 16, 17, and 18 (outliers results), the outcomes of outlier pairs with a z-score > 3.00 are closely located in the identification numbers of each pair. This proves that the pairwise identified as cheating in the acceptance method are situated close to each other in terms of seating. This implies that one of the pairs (the one that is not normal) is assumed to be the source (the one being copied) and the other copies the answers.

All pairwise in the USBN PAI test in DIY identified as cheating in the acceptance test are pairs that have been diagnosed with excessive similarity in the same answers (correct/incorrect), presented

`

at the level of each pair's ability by looking at the average value of each pair. The pairwise identified as cheating unexpectedly falls on the outlier point in the empirical distribution, thus going beyond most of the distribution. The outlier points indicate that these pairs have an unusually large (unusual) similarity response for their ability level, so the pairs are out of the acceptability limits for accepting similar responses connected with the average ability level of each pair or can be said as pairs that are abnormal and unacceptable in their similarity.

This is in line with what (Maxim et al., 2014) stated that copying is only one cause, namely a very similar response string, the examination of the empirical distribution type resulting in the identification of pairs of exam participants with very unusual string responses (abnormal), so there is no reason to accept this string. This becomes a strong indicator that responses with unusual similarities are not independent.

This is consistent with what Cizek stated that answer copying statistics can be grouped into two types (Yormaz & Sünbül, 2017). In the second method, the probability of observed patterns is compared with the distribution of values originating from independent pairs of exam participants who took the same test. Examples of such statistics are the K index (Holland, 1996).

## CONCLUSION

Based on the analysis and discussion of the National Examination for Islamic Religious Education (PAI) Package A for Senior High Schools in the Academic Year 2015/2016 in the Special Region of Yogyakarta, several conclusions can be drawn. Firstly, employing a modern approach, the characteristics of the USBN PAI test instrument for the academic year 2015/2016 exhibit an average difficulty level. The test instrument demonstrates a commendable information function, with maximum information obtained at 9.35 logits around $\theta + 0.2$, and a standard error of measurement (SEM) of 0.327. Secondly, the analysis results indicate that two individuals were identified as cheating in the USBN PAI test using the person fit method. Concurrently, employing the acceptance test method in the USBN PAI test revealed 153 pairs identified in shared knowledge, 891 pairs in shared ignorance, and 222 pairs in shared response.

Furthermore, this study is recommended to implement the person fit method in the examination, as this method proves beneficial in controlling the behavior of exam participants during the test. Given the substantial advantages of both the person fit method and acceptance test, it is advisable to consistently apply these methods in examinations, encompassing school exams and other selection exams. Secondly, cheating in exams poses a significant threat to validity. Therefore, examiners and invigilators should proactively adopt measures to minimize the possibility of cheating by implementing preventive procedures. Lastly, schools and teachers play a crucial role in fostering a sense of honesty among students facing the National Examination (USBN). Emphasizing that cheating practices are not a viable means to enhance USBN scores, schools should prioritize the implementation of a robust learning system for academic success.

## REFERENCES

Alfarisa, F., & Purnama, D. N. (2019). Analisis butir soal ulangan akhir semester mata pelajaran ekonomi SMA menggunakan Rasch model. *Jurnal Pendidikan Ekonomi Undiksha*, *11*(2), 366–374. https://doi.org/10.23887/jjpe.v11i2.20878

Anderman, E. M., & Koenka, A. C. (2017). The Relation Between Academic Motivation and Cheating. *Theory Into Practice*, *56*(2), 95–102. https://doi.org/10.1080/00405841.2017.1308172

Basinska, B. A., & Dåderman, A. M. (2023). Psychometric properties of the Bern illegitimate tasks scale using classical test and item response theories. *Scientific Reports*, *13*(1), 7211. https://doi.org/10.1038/s41598-023-34006-0

Bin-Nashwan, S. A., Sadallah, M., & Bouteraa, M. (2023). Use of ChatGPT in academia: Academic integrity hangs in the balance. *Technology in Society*, *75*, 102370. https://doi.org/10.1016/j.techsoc.2023.102370

`

Brimble, M. (2016). Why Students Cheat: An Exploration of the Motivators of Student Academic Dishonesty in Higher Education. In T. Bretag (Ed.), *Handbook of Academic Integrity* (pp. 365–382). Springer Singapore. https://doi.org/10.1007/978-981-287-098-8_58

Conway, B., Gary Martin, W., Strutchens, M., Kraska, M., & Huang, H. (2019). The Statistical Reasoning Learning Environment: A Comparison of Students' Statistical Reasoning Ability. *Journal of Statistics Education*, *27*(3), 171–187. https://doi.org/10.1080/10691898.2019.1647008

Donati, M. A., Borace, E., Franchi, E., & Primi, C. (2021). Using the Short Form of the MSBS to Assess State Boredom Among Adolescents: Psychometric Evidence by Applying Item Response Theory. *Assessment*, *28*(3), 928–941. https://doi.org/10.1177/1073191119864655

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative Content Analysis: A Focus on Trustworthiness. *SAGE Open*, *4*(1), 215824401452263. https://doi.org/10.1177/2158244014522633

Franck, O. (2021). Gateways to accessing powerful RE knowledge: A critical constructive analysis. *Journal of Religious Education*, *69*(1), 161–174. https://doi.org/10.1007/s40839-021-00133-x

Herwin, H., & Heriyati, H. (2016). *Identifikasi kecurangan peserta ujian melalui metode person fit*. 91–96.

Istiyono, E. (2016). Developingassessment Instrumentbased Quizstarin Theory Of Kinetic Gas To Measure Cognitive Abilities Senior High School Students. *Jurnal Pendidikan Fisika*, *5*(7), 437–445.

Kingsdorf, S., & Krawec, J. (2014). Error Analysis of Mathematical Word Problem Solving Across Students with and without Learning Disabilities. *Learning Disabilities Research & Practice*, *29*(2), 66–74. https://doi.org/10.1111/ldrp.12029

Kusaeri, K. (2017). Studi Perilaku Cheating Siswa Madrasah Dan Sekolah Islam Ketika Ujian Nasional. *Edukasia : Jurnal Penelitian Pendidikan Islam*, *11*(2), 331. https://doi.org/10.21043/edukasia.v11i2.1727

Manoppo, Y., & Mardapi, D. (2014). Analisis Metode Cheating Pada Tes Berskala Besar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 115–128. https://doi.org/10.21831/pep.v18i1.2128

Maxim, L. D., Niebo, R., & Utell, M. J. (2014). Screening tests: A review with examples. *Inhalation Toxicology*, *26*(13), 811–828. https://doi.org/10.3109/08958378.2014.955932

Mowbray, F. I., Fox-Wasylyshyn, S. M., & El-Masri, M. M. (2019). Univariate Outliers: A Conceptual Overview for the Nurse Researcher. *Canadian Journal of Nursing Research*, *51*(1), 31–37. https://doi.org/10.1177/0844562118786647

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *Research and Evaluation in Education*, *9*(1), 86–117. https://doi.org/10.21831/reid.v9i1.63230

Pitaloka, D. A. E., Kusuma, I. Y., Pratiwi, H., & Pradipta, I. S. (2023). Development and validation of assessment instrument for the perception and attitude toward tuberculosis among the general population in Indonesia: A Rasch analysis of psychometric properties. *Frontiers in Public Health*, *11*, 1143120. https://doi.org/10.3389/fpubh.2023.1143120

Qiu, X.-L., Chiu, M. M., Wang, W.-C., & Chen, P.-H. (2021). A new item response theory model for rater centrality using a hierarchical rater model approach. *Behavior Research Methods*, *54*(4), 1854–1868. https://doi.org/10.3758/s13428-021-01699-y

Rahman, Y. A., Rentina, L. H., & Dhini, U. R. (2023). Person Fit Analysis For Assessing Academic Writing Performance Using Rasch Model. *Jurnal Pendidikan Glasser*, *7*(2), 301. https://doi.org/10.32529/glasser.v7i2.2571

`

Ruijten, P. A. M., Haans, A., Ham, J., & Midden, C. J. H. (2019). Perceived Human-Likeness of Social Robots: Testing the Rasch Model as a Method for Measuring Anthropomorphism. *International Journal of Social Robotics*, *11*(3), 477–494. https://doi.org/10.1007/s12369-019-00516-z

Winardi, R. D., Mustikarini, A., & Anggraeni, M. A. (2017). Academic Dishonesty Among Accounting Students: Some Indonesian Evidence. *Jurnal Akuntansi Dan Keuangan Indonesia*, *14*(2), 142–164. https://doi.org/10.21002/jaki.2017.08

Yormaz, S., & Sünbül, Ö. (2017). Determination of Type I Error Rates and Power of Answer Copying Indices under Various Conditions. *Educational Sciences: Theory and Practice*, *17*(1), 5–26.