JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v13i2.34601
http://journal.uinjkt.ac.id/index.php/jp3i

Evaluation of Different Person-Fit Measures in Cognitive Tests with Different Test Lengths

Sukaesi Marianti[1], Herdin Natalius Manao[2], Arij Faiha[3]

Department of Psychology, University of Brawijaya, Indonesia

s.marianti@ub.ac.id

Abstract

Test takers' characteristic is an exciting topic to discuss in psychometric research. In this study, person-fit is a part of the person characteristics applied in the context of cognitive tests. Given the importance of accurately estimating item and person parameters, person-fit is a statistical technique that can detect aberrant responses. Aberrance adversely affects the estimation process at the level of items and persons. The purpose of this study was to introduce and apply two popular person-fit statistics called l_z and l_z^*. These two statistics were applied in two studies, in study 1 using N = 317 and item = 16, and in study 2 using N = 331 and item = 49. The results showed that in studies 1 and 2, l_z^* detected more aberrant responses compared to l_z. Significant differences in estimated results from both techniques were also shown in Study 2. The outcomes of this study are valuable for researchers and practitioners in the field of psychometrics who rely on l_z and l_z^*, as a foundation for identifying aberrant responses.

Keywords: aberrant response pattern; l_z ; l_z^*; person-fit.

Abstrak

Karakteristik peserta tes adalah topik yang menarik untuk didiskusikan dalam penelitian psikometrik. Dalam penelitian ini person-fit adalah bagian dari karakteristik peserta tes yang dibahas dalam konteks tes kognitif. Mengingat pentingnya estimasi parameter item maupun parameter peserta tes untuk akurat, maka person-fit adalah suatu teknik statistik yang dapat digunakan untuk mendeteksi respons menyimpang, dimana respons menyimpang memberikan akibat yang buruk bagi proses estimasi pada level item maupun peserta tes. Tujuan dari penelitian ini adalah untuk memperkenalkan dan mengaplikasikan dua statistik person-fit yang popular, yaitu l_z dan l_z^*. Aplikasi dua statistik tersebut dilakukan di dalam dua studi, dimana dalam studi 1 menggunakan N=317 dan item=16, dan dalam studi 2 menggunakan N=331 dan item=49. Hasil penelitian menunjukkan bahwa di dalam studi 1 dan 2, l_z^*. lebih banyak mendeteksi respon menyimpang daripada l_z. Perbedaan hasil estimasi yang signifikan dari kedua teknik tersebut ditunjukkan di dalam studi 2. Hasil dari penelitian ini sangat berharga bagi para peneliti dan praktisi di bidang psikometrika yang mengandalkan l_z dan l_z^* sebagai dasar untuk mengidentifikasi respons yang menyimpang.

***Kata Kunci**: l_z ; l_z^*; person-fit; pola respon menyimpang.*

Introduction

In psychometrics, two aspects have been the focal points of research: item characteristics and test characteristics. These two aspects are interrelated and often discussed in conjunction. However, researchers have also directed their attention toward item characteristics and the characteristics of individuals or test participants. One pertinent topic related to test participants' attributes is person-fit.

Person-fit analysis has garnered significant attention among psychometric experts due to the necessity for accurate estimation outcomes that can serve as the foundation for decision-making. When aberrant responses are present within a dataset, they can impact estimation results, including estimations of item and test characteristics, subsequently influencing test takers' attributes. Simulation studies conducted by Mousavi and Cui (2020) revealed that aberrant responses introduce biases at both the item and test taker levels, causing bias in item parameter estimations such as difficulty and discrimination (b and a). This, in turn, affects test-taker classification.

Person-fit is a technique employed to measure the appropriateness level of a test participant's response pattern to a set of items within a given test. In practical terms, person-fit is utilized to identify test participants whose response patterns deviate from the norm, often referred to as aberrance (Karabatsos, 2003; Tendeiro & Meijer, 2014). Technically, person-fit is utilized to detect aberrance or deviations by identifying unexpected response patterns that do not align with the utilized model. This allows the distinction between the anticipated and observed responses to determine whether a test participant's response deviates or raises suspicion (Marianti et al., 2014; Van Der Linden & Guo, 2008). Aberrant responses can potentially serve as sources for identifying anomalous behavior exhibited by test participants during the test. The term "expected response" denotes theoretically generated responses that align with a well-fitting model. Conversely, "observed response" refers to the responses provided by test participants to the set of items.

Furthermore, Meijer and Sijtsma (2001) expounded on the aberrant response patterns commonly associated with person-fit. These types include (1) falsely enhancing ability estimates through aberrant response patterns, encompassing test takers who cheat and guess; and (2) falsely diminishing ability estimates through aberrant response patterns, involving test takers who become flustered during the test, those unable to complete all items, and those with language deficiencies.

Person-fit analysis identifies test takers with aberrant responses, allowing for subsequent interventions. Numerous person-fit statistics have been developed; Karabatsos (2003) compared 36 such statistics, categorizing them into two broad groups: parametric and non-parametric. Among several person-fit statistics, l_z and l_z^* have gained widespread popularity. The frequently employed l_z Statistic, standardized version of the lo statistic introduced by Levine and Rubin (1979), was developed by Drasgow et al. (1985). Meanwhile, l_z^* represents a modified version of l_z (Snijders, 2001).

The lz statistic is widely recognized as one of the most popular person-fit statistics in various fields, including educational measurement and item response theory (Combs, 2023; Sinharay, 2015; Zhu et al., 2022). The l_z Statistics follows an asymptotic standard normal distribution when theoretical values of true ability are employed. However, this asymptotic distribution deviates from normality when empirical ability values are utilized (Molenaar & Hoijtink, 1990).

Equation (1) serves to derive the value of l_o.

$$l_o = \Sigma_i^n\{u_i \; InP_i(\theta) + (1 - u_i) \; In[1 - P_i(\theta)]\} \qquad (1)$$

Where n stands for the number of items in the test, u_i represents the individual's response to item i, and $P_i(\theta)$ signifies the probability of the response given item i, considering the individual's ability θ. Subsequently, Equation (2) is employed to compute the value of l_z.

$$l_z = \frac{l_o - E(l_o)}{[Var(l_o)]^{1/2}} \qquad (2)$$

The value of E(lo) corresponds to the expected value of lo, while Var(lo) denotes the variance of lo. Negative values greater than l_z may indicate distinct response patterns among participants. Recognizing the limitations of l_z, Snijders (2001) introduced the enhanced l_z^* Statistics, which mitigates the shortcomings of l_z by incorporating modifications that allow for the use of empirical ability values. Despite its enhanced capabilities, the l_z^* Statistics finds less application among practitioners due to its relatively intricate computation (Magis et al., 2012). Equation (3) outlines the calculation procedure for the l_z^* Statistics, which offers a more robust alternative to address the challenges posed by the original l_z statistic.

$$l_z^* = \frac{l_0(\hat{\theta}) - E[l_0(\hat{\theta})] + c_n(\hat{\theta}) \, r_0 \, (\hat{\theta})}{\tilde{V}[l_0(\hat{\theta})]^{1/2}} \qquad (3)$$

where

$$\tilde{V}\left[l_0\left(\hat{\theta}\right)\right] = \sum_{i=1}^{n} \tilde{w}_i(\theta)^2 P_i(\theta) \; Q_i(\theta) \qquad (4)$$

Modifications to l_z^* are observed through the addition of $c_n\left(\hat{\theta}\right) r_0 \left(\hat{\theta}\right)$ in the numerator and the modification of the $w_i\left(\hat{\theta}\right)$ function to $\tilde{w}_i\left(\hat{\theta}\right)$ in the denominator. Additionally, the modified variance $\tilde{V}\left[l_0\left(\hat{\theta}\right)\right]$ in Equation (2) is smaller than the conventional variance V(l_o), resulting in greater variance for l_z^* when compared to l_z. These modifications are applied to bring the distribution of l_z^* closer to the standard normal distribution than l_z. In its application, person-fit employs the Item Response Theory (IRT). This statistical technique measures participants' test responses to the provided items (Desjardins & Bulut, 2018).

Item Response Theory was developed to address the limitations of Classical Test Theory (CTT). In IRT, each test item has its unique characteristics, and these individual item characteristics form the overall test characteristics. These characteristics encompass difficulty level, item discrimination, and pseudo-guessing probability. These three characteristics serve as logistic parameters that determine the IRT model to be used. IRT is acknowledged as a potent method that offers thorough insights into individuals, items, and tests, exceeding Classical Test Theory (CTT) in its scope and depth of information (Jabrayilov et al., 2016; Kohli et al., 2014). For this study, the employed IRT model is the 3-PL IRT, or the three-parameter logistic model.

This study aims to apply two person-fit statistics, l_z and l_z^* , while considering test length as a condition that can influence the performance of l_z and l_z^* in detecting aberrant response patterns. According to simulation studies (De La Torre & Deng, 2008; Reise & Due, 1991), test length impacts the detection capability of the person-fit statistic lzlz. Therefore, the application of l_z and l_z^* is divided into two studies: Study 1 examines the performance of l_z and l_z^* under conditions of tests with fewer items, while Study 2 assesses their performance in tests with a more significant number of items.

Based on the explanations provided in the introduction, the purpose of this study is to apply two person-fit statistics, l_z and l_z^*, under two different test conditions based on test length. The significance of the research findings lies in establishing a scientific foundation for determining whether it is appropriate to employ either one or both of these person-fit statistics to detect patterns of aberrant responses in both short and long test conditions. The implication of this study is to provide understanding and reference for researchers and practitioners who use person-fit. This understanding involves the use of person-fit statistics (l_z and l_z^*) the strengths and limitations of these two statistics in conditions related to test length.

Method

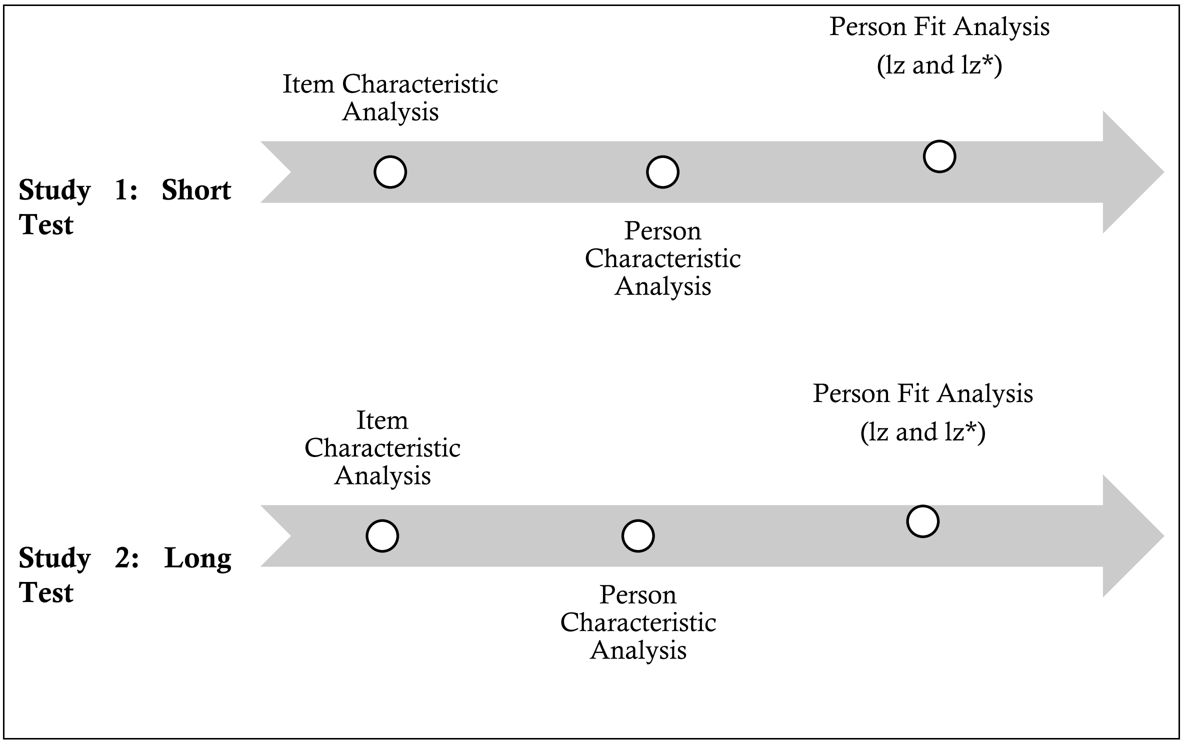
This quantitative psychometric study aimed at applying person-fit statistics under two test length conditions. The research was conducted in four stages: data collection, data partitioning based on conditions, application of person-fit statistics, and interpretation of detection figures. These stages are illustrated in Figure 1, which outlines the research design involving two test length conditions.

Participants

Real dichotomous data (0,1) were employed in this study, where 0 indicates incorrect participant responses and 1 indicates correct responses. This data was sourced from the open-source platform known as Harvard Dataverse (dataverse.harvard.edu). The study is divided into two parts: Study 1 and Study 2. Study 1 encompasses a group of 317 test participants who were administered a mathematics test of 16 items in 2018. Study 2, on the other hand, involved 331 participants who were given a reading subtest in 2018, comprising 54 items. The research sample for Study 1 consisted of 317 students (85 males and 232 females) aged 15, selected randomly from 295 schools in Indonesia. All the participants in Study 1 attempted the same set of test items. Similarly, Study 2 utilized the same set of items, with a sample size of 331 (89 males and 242 females) aged 15, drawn from 306 schools in Indonesia.

Instruments

The data employed in this study were derived from the Program for International Student Assessment (PISA) administration, an internationally conducted test with a three-year cycle. PISA is organized by the Organization for Economic Cooperation and Development (OECD). Its primary objective is to assess and compare the academic performance of school children worldwide, aiming to enhance educational methodologies and outcomes. PISA comprises four subtests: science, mathematics, reading, and financial literacy. In this study, Study 1 utilized the reading subtest data, while Study 2 employed mathematics subtest data, both administered in 2018.


Study 1: Short Test
Item Characteristic Analysis
Person Characteristic Analysis
Person Fit Analysis (lz and lz*)
Study 2: Long Test
Item Characteristic Analysis
Person Characteristic Analysis
Person Fit Analysis (lz and lz*)

Sources: Personal data (2021).

Figure 1. Research design with two conditions based on test length

Data Analysis

Data analysis in this study was conducted in two phases: firstly, by employing the Item Response Theory (IRT) model for item parameter estimation, and secondly, by performing a person-fit analysis. The data analysis process was facilitated using the R programming language. A more comprehensive ***explanation*** is provided as follows:

1. Item Parameters

In this study, the data analysis technique employed was the Three-Parameter Logistic Model (3-PLM) within the Item Response Theory (IRT) framework. The 3-PLM was selected for its flexibility and suitability for analyzing data with dichotomous responses. This model allows each item to possess its item discrimination and accommodates the probability of test participants successfully guessing an item's response. The operational mechanism of this model is articulated in Equation (5). The selection and application of this model were executed with the assistance of the 'ltm' R package (Rizopoulos, 2007).

2. Person-fit

The analytical approach adopted in this study involved employing the person-fit test to identify aberrant responses present within cognitive test data. This analysis process was facilitated using the R package named "PerFit," which encompasses a range of person-fit statistics (Tendeiro et al., 2016). Furthermore, two specific statistics were utilized: l_z and l_z^*. The computation for these two statistics is outlined in Equations (1) and (2) for l_z, and Equation (3) for l_z^*.

Results and Discussion

Results

This research was analyzed through two studies, namely Study 1 and Study 2, in which Study 1 utilized a short test and Study 2 employed a more extended test. The following explanation pertains to the data analysis outcomes of both studies. Before the person-fit analysis using the R package Perfit, item parameters were estimated using the 2PL IRT model, facilitated by the Irtoys package and ICL engine.

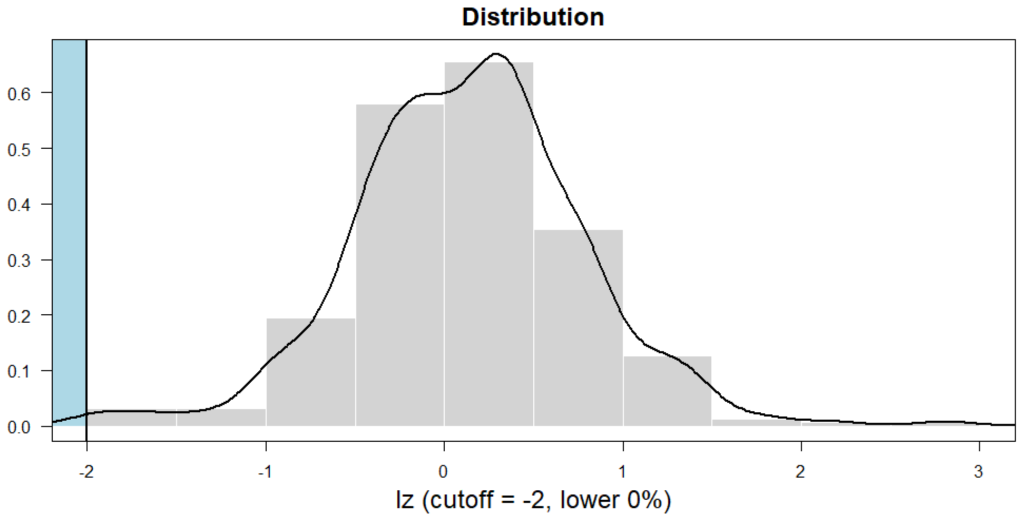
1. Study 1

The data analysis outcomes for the abbreviated 16-item test involving 317 student participants revealed the following. Employing a cutoff of -2, as established by (Meijer, 2009), no instances of aberrant responses were identified through the utilization of the l_z Statistics. In contrast, the l_z* Statistics detected 8 out of 317 participant response patterns (2.524%). Visual representations of the distribution statistics for both l_z and l_z* are provided in Figures 2 and 3, respectively. Detailed statistical descriptions concerning the person-fit values for both statistics are presented in Table 1.

Table 1. Descriptive statistics of person-fit analysis results

Statistic	Detection Rate (%)	Mean	SD	Min.	Max.
l_z	0	0.135	0.646	-1.9989	2.8043
l_z*	2.524	0.077	1.000	-2.814	2.086

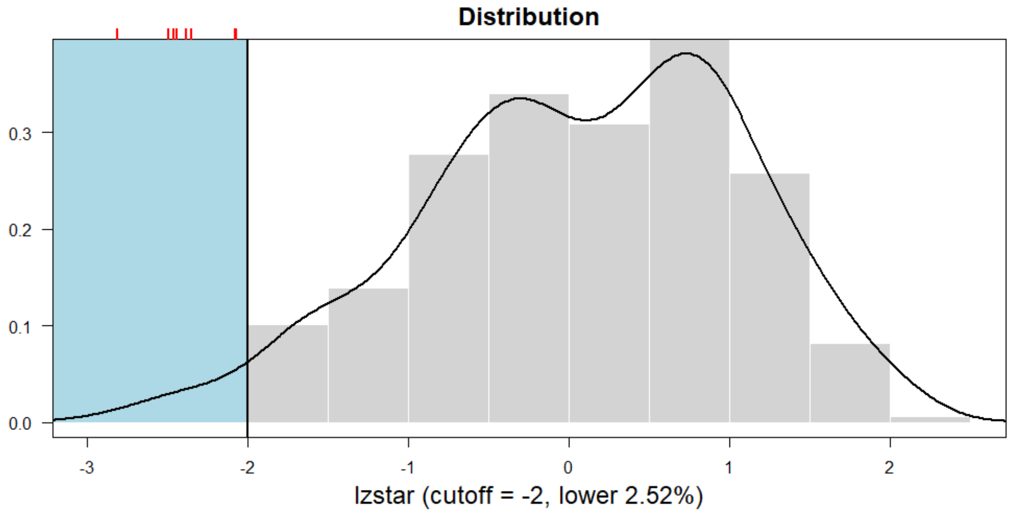
Sources: Analysis by the author in the current study (2021).



Sources: Analysis by the author in the current study (2021).

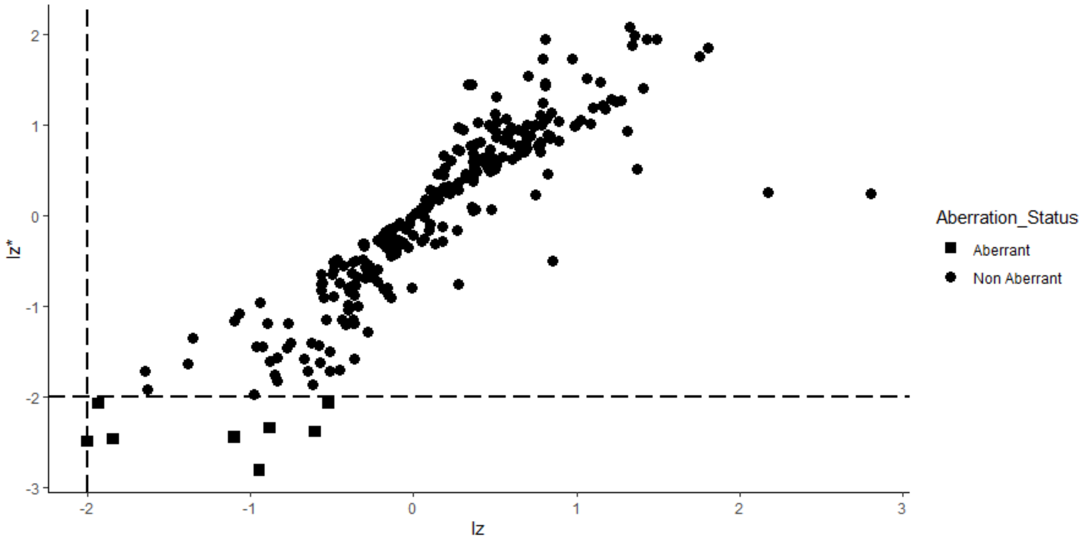
Figure 2. Distribution of person-fit statistic based on l_z

Comparison between l_z and l_z^* can be observed in the overlay plot presented in Figure 4. Four regions are identified, with the upper-left region representing test participants detected as aberrant based on X but not detected by Y. The second region corresponds to the lower-left area, indicating test participants detected as aberrant by both statistics. This region exhibits a notably more significant concentration compared to the first region. The third region, situated in the upper-right area, depicts test participants not detected as aberrant by either statistic; most test participants are observed to cluster in this region. The final region, located in the lower-right quadrant, represents participants not detected as aberrant by X but detected as aberrant by Y.



Sources: Analysis by the author in the current study (2021).

Figure 3. Distribution of person-fit statistic based on l_z^*



Sources: Analysis by the author in the current study (2021).

Figure 4. Comparison of the distribution of person-fit statistics based on l_z *and* l_z^*

To enhance the precision of the comparison between l_z and l_z^*, a paired sample t-test analysis is conducted. The t-test results yield a t-value of 1.889, with α=0.060, indicating that there is no significant difference in the abilities of l_z and l_z^* in estimating person-fit statistics, which are subsequently employed for detecting aberrant responses (aberrant responses).

2. *Study 2*

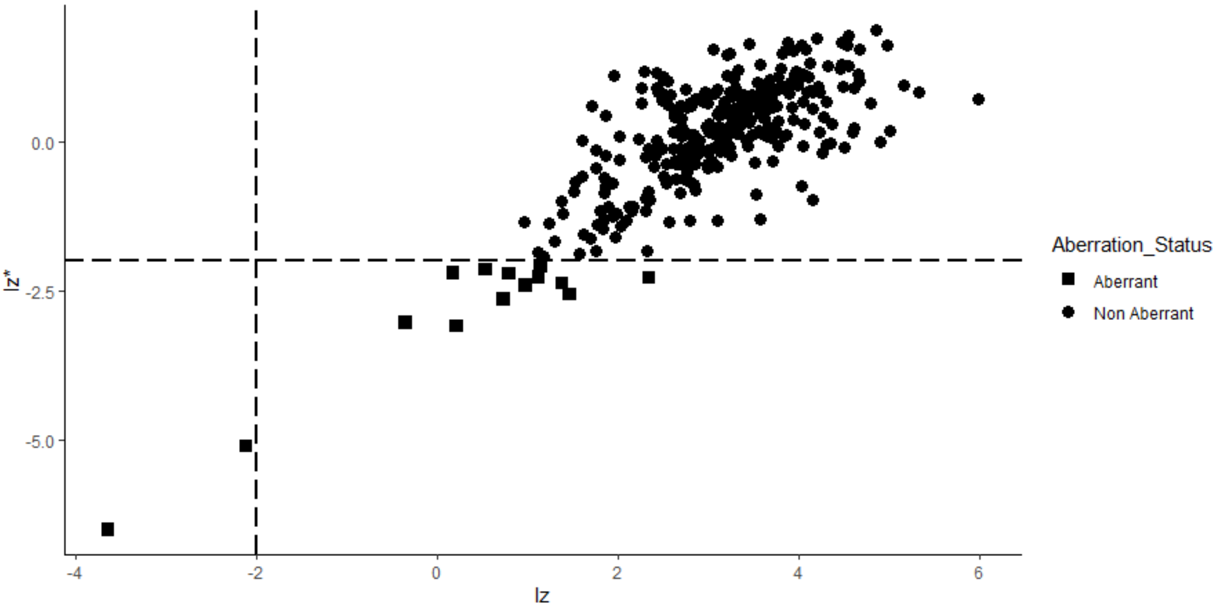
Study 2 was conducted based on the reading subtest comprising 54 items, with 331 participants (89 male and 242 female students). Five items were eliminated after data screening due to a lack of participant responses, indicated by all items containing “Not Available” (NA) values. Consequently, a total of 49 suitable items remained for analysis. The results of the person-fit analysis using the l_z and l_z^* statistics are presented in Table 2 and are also supported by Figures 6 and 7, which show that among the 331 participants' response patterns, 2 (0.604%) were identified as exhibiting aberrant response patterns based on the l_z statistic. Furthermore, the l_z^* statistic detected 14 response patterns (4.230%) that indicated suspicious responses.

Table 2. Descriptive statistics of person-fit analysis results

Statistic	Detection Rate (%)	Mean	SD	Min.	Max.
l_z	0.604	3.041	1.041	-3.636	5.985
l_z*	4.230	0.079	1.039	-6.507	1.868

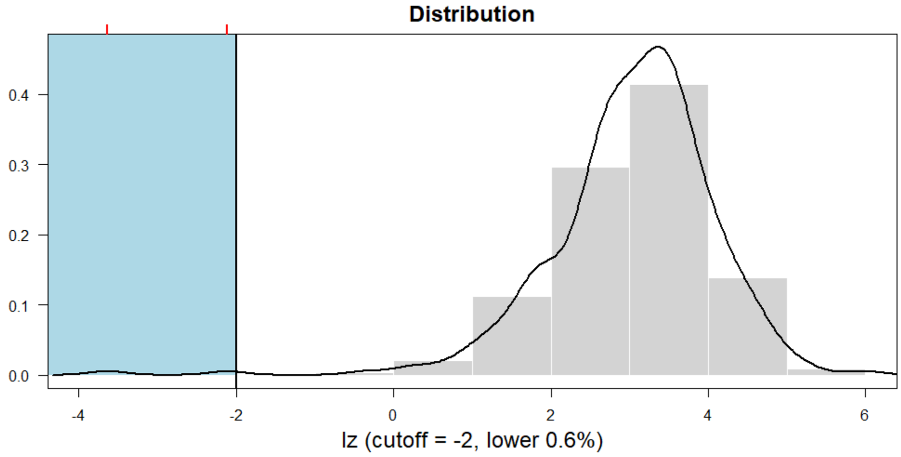
Sources: Analysis by the author in the current study (2021).

The distribution plots of the l_z and l_z^* statistics are depicted in Figure 5. From Figure 5, it can be observed that several aberrant response patterns went undetected by l_z but were identified by l_z^*. This observation aligns with the t-test results obtained from the paired sample t-test analysis, revealing a t-value of 78.675 and α value of 2.20E-16. These findings signify differences in the abilities of l_z and l_z^* to estimate person-fit statistics as well as to detect aberrant responses.



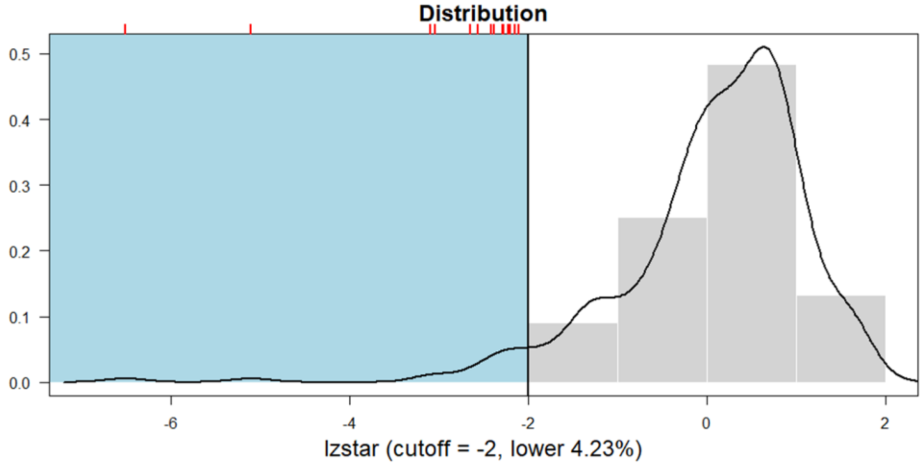
Sources: Analysis by the author in the current study (2021).

Figure 5. Comparison of the distribution of person-fit statistics based on l_z and l_z^*



Sources: Analysis by the author in the current study (2021).

Figure 6. Distribution of person-fit statistic based on l_z



Sources: Analysis by the author in the current study (2021).

Figure 7. Distribution of person-fit statistic based on l_z^*

Discussion

The purpose of this research is to apply two person-fit statistics, l_z and l_z^*, under two different test conditions based on test length. These conditions are divided into Study 1 and Study 2. Study 1 involves N=317 participants and 16 items, while Study 2 involves N=331 participants and 49 items. Study 1 has fewer items compared to Study 2, leading to the hypothesis that the performance of the two statistics, l_z and l_z^*, will differ between the two studies. Data analysis results from Study 1 indicate that l_z did not detect any participants with aberrant response patterns, whereas l_z^* detected only 8 participants (2.524%) with aberrant response patterns. Such discrepancies are common, where l_z^* often exhibits higher detection rates compared to l_z.

However, a t-test for paired samples shows no significant difference between the estimation outcomes of the person-fit statistics l_z and l_z^*. This suggests that the estimated person-fit values using both techniques do not substantially differ when dealing with shorter test conditions. Moreover, Figure 3 vividly illustrates the distribution of person-fit statistics from both techniques. Based on this figure, the majority of test participant responses are non-aberrant, indicating that both person-fit techniques do not detect them. Nevertheless, 8 individuals are identified as aberrant by l_z^* but remain undetected by l_z.

Study 2 presents different outcomes compared to Study 1, wherein the detection rates of aberrant responses are higher for both techniques than in Study 1. Statistic l_z detected 2 aberrant responses (0.604%), while l_z^* detected 14 aberrant responses (4.230%). This scenario showcases an increase in detection rates by l_z and l_z^* in Study 2. Previous research by Meijer and Sijtsma (2001) suggests that the detection of aberrant responses increases based on factors such as the type of aberrant response pattern, theta values, and test length.

Figure 5 displays a more dispersed distribution of person-fit statistics than Figure 4. While there are resemblances in displaying non-aberrant responses, discrepancies emerge in detecting aberrant responses between l_z and l_z^* in Figure 5. The figure indicates that X responses are identified as aberrant by l_z but not by l_z^*. Similarly, X responses are detected as aberrant by l_z^* but not by l_z.

The findings in Figure 5 are supported by paired sample t-test comparisons, where the analysis results suggest a significant difference in the person-fit statistics estimation outcomes l_z and l_z^*. This points to distinct estimations by l_z and l_z^* in Study 2, involving a longer test length. Research by Meijer et al.(1994)suggests that aberrant response patterns are more easily detected in longer tests with higher discriminative power.

Considering the data analysis outcomes of Study 1 and Study 2, it is evident that l_z and l_z^* exhibit significant differences in estimating person-fit statistics for longer tests. However, for shorter tests, both person-fit techniques show no differences in estimation outcomes. Therefore, it is not advisable to employ l_z and l_z^* for shorter tests. In longer tests, researchers can employ l_z and l_z^* for person-fit estimation, with l_z^* leading to a higher number of detected aberrant responses compared to l_z.

Numerous studies have demonstrated the superiority of both l_z and l_z^* compared to other person-fit techniques (Gorney & Wollack, 2023; Kim & Moses, 2018; Lee et al., 2014). Nevertheless, these two techniques possess strengths in specific conditions, such as test length. The robustness of these techniques cannot be generalized beyond these specific conditions.

Conclusion

Based on the findings of both conducted studies, it can be inferred that significant differences in statistical estimation outcomes of l_z and l_z^* values can be identified, particularly in Study 2, where the employed test was of considerable length. In Study 1, however, both statistical techniques exhibited no significant variance in estimation outcomes. Another noteworthy discovery is that l_z^* demonstrates a higher capability in detecting aberrant responses across all studies. Nevertheless, amidst these differences, there is concurrence between the two methods in detecting aberrant responses, where both techniques identify the same test participants.

The results of this study have implications for research and applications in the field of psychometrics, using l_z and l_z^* as statistics to provide a foundation for making decisions that determine whether participants' test responses are aberrant. However, considering the findings from Study 1 and Study 2, future research endeavors should consider conducting simulation studies to design test conditions and participant attributes meticulously. The design of diverse conditions will offer more precise insights into the strengths and limitations of the person-fit technique l_z and l_z^*. A greater number of designed conditions will expand the possibilities for generalizing simulation results to real-world data.

Acknowledgment

I extend my sincere gratitude to the Faculty of Social and Political Sciences, Universitas Brawijaya for providing the research grant that made this study possible.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this research. All data and results presented in this study have been analyzed and reported impartially, without any external influence or bias. The research has been conducted solely for academic and scientific purposes.

Authors Contribution

Author 1: Designed the study, wrote the manuscript, and conducted data analysis.

Author 2: Collected data and contributed to data analysis.

Author 3: Collected data, searched for references, and assisted in editing the manuscript.

References

Combs, A. (2023). A New Bayesian Person-Fit Analysis Method Using Pivotal Discrepancy Measures. *Journal of Educational Measurement*, *60*(1), 52–75. https://doi.org/10.1111/JEDM.12342

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. https://doi.org/10.1111/J.1745-3984.2008.00058.X

Desjardins, & Bulut. (2018). *Handbook of Educational Measurement and Psychometrics Using R*. https://www.crcpress.

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. https://doi.org/10.1111/J.2044-8317.1985.TB00817.X

Gorney, K., & Wollack, J. A. (2023). Using Item Scores and Distractors in Person-Fit Assessment. *Journal of Educational Measurement*, *60*(1), 3–27. https://doi.org/10.1111/JEDM.12345

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied psychological measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Karabatsos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement in Education*, *16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Kim, S., & Moses, T. (2018). The Impact of Aberrant Responses and Detection in Forced-Choice Noncognitive Assessment. *ETS Research Report Series*, *2018*(1), 1–15. https://doi.org/10.1002/ETS2.12222

Kohli, N., Koran, J., & Henn, L. (2014). Relationships Among Classical Test Theory and Item Response Theory Frameworks via Factor Analytic Models. *Educational and Psychological Measurement*, *75*(3), 389–405. https://doi.org/10.1177/0013164414559071

Lee, P., Stark, S., & Chernyshenko, O. S. (2014). Detecting Aberrant Responding on Unidimensional Pairwise Preference Tests. *Applied psychological measurement*, 391–403. https://doi.org/10.1177/0146621614526636

Levine, M. V., & Rubin, D. B. (1979). Measuring the Appropriateness of Multiple-Choice Test Scores. *Journal of educational statistics*, *4*(4), 269–290. https://doi.org/10.3102/10769986004004269

Magis, D., Raîche, G., & Béland, S. (2012). A Didactic Presentation of Snijders's lz* Index of Person Fit With Emphasis on Response Model Selection and Ability Estimation. *Journal of Educational and Behavioral statistics*, *37*(1), 57–81. https://doi.org/10.3102/1076998610396894

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for Aberrant Behavior in Response Time Modeling. *Journal of Educational and Behavioral Statistics*, *39*(6). https://doi.org/10.3102/1076998614559412

Meijer, R. R. (2009). Person-Fit Research: An Introduction. *Applied Measurement in Education*, *9*(1), 3–8. https://doi.org/10.1207/S15324818AME0901_2

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of Test and Person Characteristics on Nonparametric Appropriateness Measurement. *Applied Psychological Measurement*, *18*(2), 111–120. https://doi.org/10.1177/014662169401800202

Meijer, R. R., & Sijtsma, K. (2001). Methodology Review: Evaluating Person Fit. *Applied psychological measurement*, *25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75–106. https://doi.org/10.1007/BF02294745/METRICS

Mousavi, A., & Cui, Y. (2020). The Effect of Person Misfit on Item Parameter Estimation and Classification Accuracy: A Simulation Study. *Education Sciences, 10*(11), 324. https://doi.org/10.3390/EDUCSCI10110324

Reise, S. P., & Due, A. M. (1991). The Influence of Test Characteristics on the Detection of Aberrant Response Patterns. *Applied Psychological Measurement*, *15*(3), 217–226. https://doi.org/10.1177/014662169101500301

Rizopoulos, D. (2007). ltm: An R Package for Latent Variable Modeling and Item Response Analysis. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/JSS.V017.I05

Sinharay, S. (2015). Assessing person fit using l_z^* and the posterior predictive model checking method for dichotomous item response theory models. *International Journal of Quantitative Research in Education*, *2*(3/4), 265. https://doi.org/10.1504/IJQRE.2015.071730

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342. https://doi.org/10.1007/BF02294437/METRICS

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of Invalid Test Scores: The Usefulness of Simple Nonparametric Statistics. *Journal of Educational Measurement*, *51*(3), 239–259. https://doi.org/10.1111/JEDM.12046

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5). https://doi.org/10.18637/JSS.V074.I05

Van Der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/S11336-007-9046-8/METRICS

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