Pegem Journal of Education and Instruction, Vol. 12, No. 4, 2022 (pp. 39-46)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Person Fit Statistics to Identify Irrational Response Patterns for Multiple-choice Tests in Learning Evaluation

Herwin Herwin[1*], Shakila Che Dahalan[2]

[1]Faculty of Education, Universitas Negeri Yogyakarta, Indonesia

[2]Faculty of Human Science, Universiti Pendidikan Sultan Idris, Malaysia

Abstract

This study aims to analyze and describe the response patterns of school exam participants based on the person fit method. This research is a quantitative study with a focus on research on social science elementary school examinations as many as 15 multiple choice items and 137 participant answer sheets. Data collection techniques were carried out with documentation by collecting the participants' exam results and then being scored based on the answer key. The data analysis technique used is statistical person fit analysis with the Rasch Model. Data were analyzed using the R Program with the package latent trait model (ltm). The results showed that as many as 93 or about 67.9% of examinees detected fit or were categorized as having rational response patterns and as many as 44 examinees or around 32.1% of examinees were detected as not fit or categorized as having irrational response patterns. Based on the results of the study, it was concluded that most elementary school exam participants had a response pattern that was fit (had a rational response pattern).

Keywords: Person Fit, Irrational Response Patterns, Multiple Choice Tests.

Introduction

Multiple choice tests are assessment techniques that are conducted where the respondent is given the opportunity to mark the answer choices from a series of choices. (Çiftçi, 2019). The characteristics of multiple-choice tests that provide answer choices often make the test participants do it by chance to get the correct answer. Primarily on multiple choice tests with scoring without correction (no penalty is imposed on the wrong answer) often impacts the desire of test participants to answer with guesses. This seems to be considered normal by the test participants so that during the assessment sometimes the test participants tend to fill the answer sheet with random responses even though they do not understand and master the question content in the test item. Instead of leaving blank answers, they tend to fill in even though it comes from guesswork.

Multiple choice items include statements called stem. This stem is the question must be answered or the problem must be solved. The other part is the response options that will be used to determine the correct choice. From this option there is one answer key and the other is the distractor. (Bailey & Curtis, 2015; Kılıçkaya, 2019). Motivation to complete all answer choices certainly affects the behavior of participants in giving answers. The phenomenon in the implementation of the test often raises questions for developers and test implementers that is sometimes found irrational participant response patterns. Sometimes participants can give correct responses to difficult questions but instead are not able to answer correctly on easy questions.

To improve the quality of education, one effort that can be done is to improve the quality of assessment (Herwin & Mardapi, 2017). A good measurement is a measurement that gets results that show the actual abilities of the participants. High or low ability of participants should come from the actual profile, not from the guesswork of the correct answer. Therefore, there is a need for control over cases of unreasonable response patterns in the implementation of assessments especially those using multiple choice test instruments.

Recognizing the challenges in applying multiple choice tests makes the implementation require statistical methods to maintain the quality of assessment and measurement results. Person fit statistics is a method used to detect deviant response patterns (Huang, 2012, p.28). This person fit is used as a basis for interpreting the response patterns shown by test takers (Mehrzi, 2011). Person fit in item response theory is a measurement made to detect anomalies in the test participant's response patterns (Meijer & Sijtsma, 2001; Pan & Yin, 2017).

In the test implementation, the score may not represent what is known by students and what can be done by students. Person fit provides important aspects of proof of validity as well as the results of analysis that are very useful to convey to education stakeholders (Walker & Engelhard, 2015). In the

Corresponding Author: herwin89@uny.ac.id

https://orcid.org: 0000-0002-8882-5087

How to cite this article: Herwin H, Dahalan SC (2022). Person Fit Statistics to Identify Irrational Response Patterns for. Multiple-choice Tests in Learning Evaluation. Pegem Journal of Education and Instruction, Vol. 12, No. 4, 2022, 39-46

Source of support: Nil

Conflict of interest: None.

DOI: 10.47750/pegegog.12.04.05

Received : 17.12.2021

Accepted : 15.05.2022 **Published:** 01.10.2022

parametric context, person fit analysis rates higher where individual response patterns match the expected pattern of the model and its estimated level of trait. This assessment is very important because individual interpretations and valid conclusions based on an estimated trait level are only justified if the individual response patterns are consistent. In addition, in the analysis of external validity, the presence of the proportion of irrelevant respondents can influence relations with relevant criteria (Schmitt et al., 2015; Ferrando, 2012).

Performing measurements using multiple-choice tests can sometimes produce results that do not represent the ability of the test takers. Multiple choice characteristics that provide options for test takers allow test takers to give answers by guessing. Often found an unusual response pattern such as being able to answer correctly on difficult items but wrong on easy items. The pattern of guess responses by test takers will have an impact on increasing measurement errors in the assessment activities. This study is expected to contribute to improving the quality of the implementation of learning evaluation in schools. This is very important because a good evaluation is an evaluation that has justice for all parties without taking sides with one (Herwin & Phonn, 2019). In addition, the quality of the evaluation process and results can be a motivation for students to study further (Tjabolo & Otaya, 2019; Suleman et al., 2015).

The focus of this research is the application of person fit statistics on the conduct of assessment. This person fit is applied to identify irrational response patterns in the implementation of multiple-choice tests in elementary schools. This study aims to analyze and describe the response patterns of elementary school examinees based on the person fit method. Based on some of these things, the research questions are: (1) what are the characteristics of the multiple-choice test items used in the evaluation of learning? (2) what are the characteristics of the participants who have taken the test? (3) what is the response pattern of participants' answers based on the Person Fit Method?

Method

Research Design

In general, this research uses a descriptive quantitative approach. This research was conducted in state elementary schools in Bontomarannu District, Gowa Regency, South Sulawesi Province, Indonesia. This research is focused on the social science subject test set in the 2017/2018 academic year. The school exam questions are in the form of multiple-choice objective tests with as many as three options developed by the KKG Team appointed by the local government.

Population and Sample

To apply the person fit method, we need a response pattern of the test participants who come from the test participant's answer sheet. In this study, the examinees 'answer sheets were randomly sampled to obtain 137 test participants' answer sheets. The tests used are facilitated and prepared by the local government through the education office. This test is held twice a year to provide information to local governments regarding the learning progress of each school unit.

Data Collection

In this study, the researcher acts as an external party who analyzes the implementation of learning evaluation in general and analyzes the response patterns of participants in particular. The data collection technique in this research is the documentation technique. The instrument used in this study was a multiple-choice test that was used during elementary school exams. The test consists of 15 items with 3 options that discuss social science material. Other instruments are test takers' answer sheets which are collected through documentation activities.

Data Analysis

This research is focused on Social Sciences Questions of Elementary School Exams in Bontomarannu District, Gowa Regency, South Sulawesi Province. The exam questions analyzed in this study are the 2nd grade exam questions. The exam questions are in the form of multiple-choice objective tests consisting of 15 items. Through documentation techniques, 137 participants' answer sheets were netted for analysis based on the person fit method. After the data is collected, then the measurement model used in this study is the Rasch model so that all responses from the test sheet answer sheets are calibrated using the Rasch model. This model is focused on one parameter, namely the level of difficulty while the item discrimination index is assumed to be the same for all items (Baker, 2001; Baker & Kim, 2017; Herwin et al., 2019; Rizopoulos, 2006). The data analysis technique used is the person fit method. Person fit is calculated using the following formula.

$$Z = \frac{L(\theta)\text{ - }E[L(\theta)]}{SD([Ll(\theta)]}$$

with,

z : Person fit coefficient
L(θ) : Maximum Likelihood Estimation
E : Expected Value
SD : Standard Deviation

(Hulin, Drasgow, & Parsons, 1983; Reise, 1990; Armstrong & Stoumbos, 2007; Torre & Deng, 2008)

The research data obtained were then analyzed using the person fit method using the R Program with the Package Latent Trait Model. To determine whether the examinees are fit, criteria is needed. Fit criteria are used based on Reise's opinion (1990) which explains that the negative z value indicates that the examinee (person) concerned is detected as not fit and vice versa.

Findings

The findings of this study are described based on three main focuses which are the research questions. The focus in question is the characteristics of the multiple-choice test used in the evaluation of learning. The second focus is the characteristics of the participants and the last focus is the response pattern of the participants based on the Person Fit Method. The caliber results provide information on the parameters of the Social Science exam questions presented in Table 1 below.

Table 1 shows information on item characteristics after item calibration using the Rasch model. In the table, we can see that the characteristics of the question items are focused on one parameter, namely the item difficulty level parameter, while the discrimination index parameter is assumed to be equal to 0,692 for all items. To facilitate observing the level of difficulty parameters in all items, the following is presented in Figure 1 which is the distribution of the level of difficulty of items.

Table 1: Item Parameters for Social Sciences Elementary School Exams

Items	*Difficulty level*	*Item discrimination index*
1	-1,821	0,692
2	-1,759	0,692
3	-0,830	0,692
4	1,583	0,692
5	-0,880	0,692
6	1,040	0,692
7	-4,771	0,692
8	1,702	0,692
9	0,262	0,692
10	2,017	0,692
11	1,951	0,692
12	0,451	0,692
13	-0,633	0,692
14	-2,876	0,692
15	-3,177	0,692

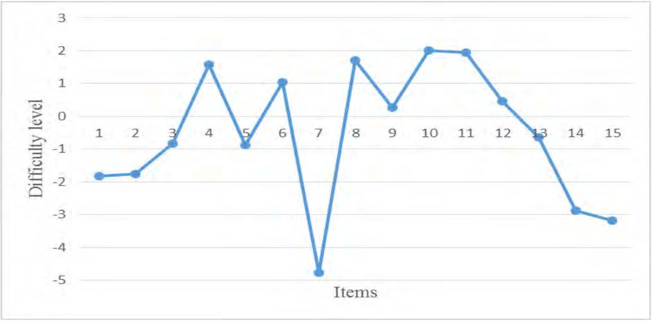

Fig. 1: Distribution of Test Item Difficulty

Based on the results of the calibration of the item items presented in Figure 1 shows that the most difficult item of the Social Sciences Examination for Elementary School Exams in Gowa Regency is item number 10 with an item difficulty coefficient of 2,017, while the easiest item is item number 7 with a coefficient item difficulty level of -4,771. For other forms of information presentation, the following Figure 2 presents an item characteristic curve for all items.

Based on the data presented in Figure 2, it can be explained that from the 15 items used have different characteristics based on the characteristics curve of the item. There are items that are very easy, there are items that are moderate, and there are also items that are very difficult for examinees. Thus, the fifteen items have different levels of difficulty. The following Figure 3 presents a comparison of the characteristics of items between easy items, medium items, and difficult items.

Figure 3 presents a comparison between easy, medium and difficult items. In the figure it can be seen that Item 7 is the easiest item because based on the curve even though participants have low ability, however they have a high

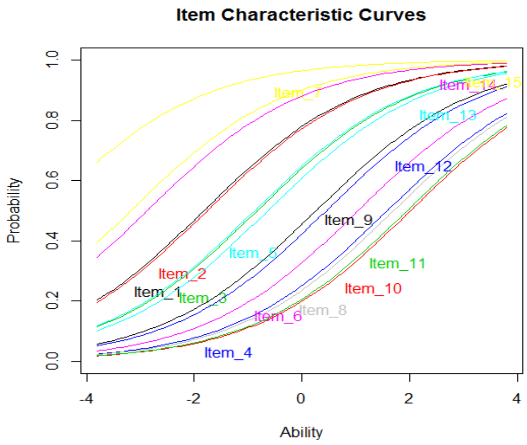

Fig. 2: Characteristic Curves for All Items

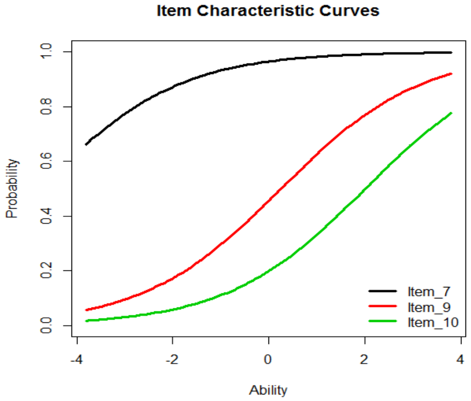

Fig. 3: Comparison of Characteristic Curves Between Easy, Medium and Difficult Items

probability of answering Item 7 correctly. The opposite occurs in Item 10. In item 10 the probability of answering items correctly is only owned by participants with high ability. Based on this presentation, ideally, low-ability participants have a small chance to answer difficult question items correctly and conversely participants with high abilities should be able to answer easy items correctly.

After obtaining the item parameters through item calibration, the next step is to analyze the test participant's response patterns based on the person fit method. This was done to obtain the value of z (coefficient of person fit) on each pattern of response test participants. The coefficient z is used to conduct testing and draw conclusions related to the status of the examinees. The person fit analysis is performed using software assistance, namely the R Program with the Latent Trait Model Package. The results of the analysis of person fit are presented in Table 2.

Based on the presentation of the data in Table 2, it shows information that out of 137 students (examinees) analyzed, as many as 93 or about 67.9% of the examinees were detected as having a fit response pattern and 44 of the 32.2% examinees were detected as having a response pattern. which is not fit. These results provide information that the majority of participants in the Elementary School of Social Sciences Examination in Bontomarannu District, Gowa Regency were not identified as having an irrational response pattern. This can be proven after the participant response patterns are analyzed by the person fit method. In addition, it seems that there are still other participants (32.1%) who are identified as having an irrational response pattern. This was shown after being analyzed by the person fit method showing a negative z coefficient (not fit). The results of a more complete person fit statistics analysis from Table 2 can be seen in the following link: https://drive.google.com/file/d/1ECoAYstHxIQS00pNovH26zhUDmCjYpj6/view.

Table 2: Person Fit Examinees Analysis Results

Coefficient	*Detected*	*Total Examinees*	*%*
Positive	Fit	93	67.9
Negative	Not Fit	44	32.1

Furthermore, to explain the application of person fit on examinees, the following empirical data is presented in the example of two examinees who have the same total score but have different person fit coefficients. The two intended examinees are examinees with codes 003 and 108. Participants with codes 003 and 108 basically have the same total score of 10 results from 10 items that were answered correctly and 5 items that were answered incorrectly. The following Table 3 and Table 4 are presented, which are the response patterns of examinee with code 003 (examinee who has a rational response pattern) and examinee with code 108 (examinee who has irrational response patterns)

Based on the information presented in Table 3 and Table 4, it can be explained that both participants have the same total score of 10. If using the classic method that has been used so far, the conclusions for both participants are the same, ie both have scores 10 and this score is quite high and is in the top group of examinees. Another thing will happen if the person fit method is applied to both response patterns. After being analyzed by the person fit method, the z coefficient for an examinee with code 003 is 1,741. The coefficient shows that an examinee with code 003 is detected fit or in the concept of person fit an examinee is categorized as having a rational response pattern. Another decision was obtained from an examinee's response pattern with code 108. Based on the results of the analysis with the person fit method, a coefficient of -2.193 was obtained. The coefficient indicates that an examinee with code 108 detected has a response pattern that is not fit or in the concept of person fit an examinee is categorized as having an irrational response pattern.

To more easily understand the concept of person fit, we can see the case of examinees with codes 003 and 108. Examinee 003 seems to be wrong on items 8, 10, and 11. This is considered logical considering the status of the three items is the most difficult item on the exam questions. (see Figure 1) after calibrating the grains. It means that examinee 003 is

Table 3: *Response Pattern of Examinee with Code 003 (Total Score = 10 and Coefficient z = 1.741)*

Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Response pattern	1	1	1	0	1	0	1	0	1	0	0	1	1	1	1

Information:
1: The score for the item answered is correct
0: The score for the item answered is incorrect

Table 4: Response Pattern of Examinee with Code 108 (Total Score = 10 and Coefficient z = -2.193)

Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Response pattern	0	1	1	1	1	0	1	1	0	1	0	1	1	1	0

Information:
1: The score for the item answered is correct
0: The score for the item answered is incorrect

fit with the model that is able to answer the easy and answer incorrectly on difficult items so that it has a rational response pattern. Another thing occurred in the 108 participants who were detected as not fit. Empirical data shows that 108 participants were able to correctly answer points 8 and 10 which were difficult items, but instead were unable to correctly answer items 1 and 15 which were basically easy items in the exam questions. This is what caused the 108 participants to be detected as a participant who has an irrational or unsuitable response pattern.

Discussion

The practice of measurement and testing in the field of education by using multiple choice test instruments certainly makes it easy for implementation on a large scale. However, it is important to realize that the characteristics of multiple-choice tests that have a choice of possible answers are chosen by test takers. This choice impacts several possibilities that can be done by the participants starting from answering with analysis, making a fortune or guessing, or just choosing to improvise to meet the answer sheet. This requires careful control so that the decisions taken reflect the true condition of the test takers. This is something that must be understood by the teacher as a pedagogic competence that is their responsibility (Tjabolo & Herwin, 2020; Saptono et al., 2021; Pujiastuti et al., 2021).

The results of the analysis by the person fit method were obtained by some participants who were detected as not fit. If these empirical findings are related to some previous findings that it is most likely that the pattern of response is not fit caused by respondents not interested (not motivated) in the assessment in the exam so they show the behavior of answering many items at random to produce an inconsistent response pattern (Ferrando, 2015). One of the causes of response patterns or item score patterns that are not fit is caused by the behavior of test participants who do not have the attraction or motivation in conducting the test so that they answer or respond to questions randomly.

One important problem in psychological measurement and education assessment is the problem of the validity of individual scores. Deviant response behavior is an important factor influencing the validity of scores in test execution. Deviant item response patterns can have an impact on the predictive power of a test that is incompatible with the individual's true abilities. This is the basis of the importance of person-fit statistics that are useful for detecting deviant behavior in individual response patterns (Avşar, 2019). Person-fit is one of the bases for proving the validity of test scores (Tendeiro et al., 2016; Walker & Engelhard, 2015).

The response patterns that are detected as not fit tend to be caused by cheating behavior of the examinees. Such behavior can be in the form of attempts to violate exam rules such as looking at references, copying other people's work, and even the motivation of examinees to seek the benefits (lack of motivation) (Meijer & Stoop, 2001). This is categorized as a response that deviates from the nature of the implementation of the test (Woods, Oltmanns, & Turkheirmer, 2008). For examinees with a deviant response pattern, it will have an impact on the estimation of inaccurate abilities and also have an effect on the validity of the measurement, which is misleading, thus the decision based on the score will not be fair or inappropriate. Test results like this have serious consequences and tend to be detrimental for decision making (Torre & Deng, 2008).

The accuracy of a policy decision is highly dependent on the accuracy of the assessment measure. Assessment plays an important role in maintaining accountability for promotion of students from class one to the next class. However, the threat of carrying out the assessment, namely the scores generated by students, may not reflect actual abilities. For example, if a student answers incorrectly on an easy item while being able to answer correctly on a difficult item. This situation can have an impact on misinterpretation and test score actions which can violate the perspective of accuracy and fairness (Mousavi, 2019).

In education measurement, detection of response patterns is very important to do. This is important because the difference between ability and test results can result in incorrect corrective steps or it can also have an inaccurate selection decision. Such response patterns can be identified through person-fit analysis (Santos et al., 2019). Another thing that can be obtained is information collected about individuals such as background information or the results of their work can be used to reveal the sources of actual errors (Cui & Roberts, 2013; Torre & Deng, 2008). Furthermore, the application of person-fit statistics is very good for finding item score patterns that can systematically reveal specific behaviors such as guessing, cheating, planning, sleeping, or other mismatch mechanisms (Dardick & Weiss, 2017; Miejer et al., 2016).

In the practice of measuring and testing the use of information the total score of the test takers is not enough to be relied upon as the only basis for decision-making information. Further investigation is needed on the score patterns shown by test takers (Wind & Walker, 2019). Person fit is designed to detect anomaly response patterns or aberrant response patterns (Fox & Marianti, 2017; Meijer & Sijtsma, 2001; Pan & Yin, 2017). Person-fit analysis is useful and promising in interpreting individual score results. This is because through person-fit analysis we can detect potential threats to the validity of the test taker's score inference. This person-fit is also very good for contextualizing individual performance in carrying out tests and can highlight a subset of certain items that test takers do unexpectedly (Walker, Jennings, & Engelhard, 2018). However, it should be considered that the use of statistical person-fit mechanics to detect fraud based on response time can be risky, because deviant behavior may arise due to poor time management (Linden & Guo, 2008; Sinharay, 2018).

The application of the person fit method is very beneficial for the implementation of a test. This person fit method can be applied to control the behavior of participants in an exam. This control is important, because in essence the implementation of the test aims to measure how much the test participant's understanding of the competency being tested is not just how capable the participant is giving the correct answer to the items given. The person-fit method is also very useful in the validity of measurements in the implementation of the test. This is considered important because basically the right measurement will produce an objective decision. Conducting a person-fit evaluation for cognitive diagnostic assessments is very important because failure to identify the wrong response can lead to misinterpretation to decision errors (Cui & Li, 2015). In addition, Goodness of Fit can also be used to measure the accuracy of empirical data that is relevant to rational expectations (Herwin & Nurhayati, 2021; Herwin et al., 2022).

The results of this study indicate that most of the test participants already have a fit response pattern or a rational pattern of 67.9%. This means that basically the majority of test takers are identified to provide answers with rational analysis, i.e., they are able to answer questions more easily and fail at more difficult questions. Although there are still a small number of test takers who are detected as not fit, this certainly is an evaluation material for the organizers for further analysis. The results of this study and the support of some previous findings indicate that person-fit statistics are very useful in the practice of measurement and testing. This can minimize measurement errors and decision errors. Primarily on the deployment of a multiple-choice test instrument that allows many types of behavior to emerge. With this person-fit analysis it is hoped that we can distinguish between participants who master competencies and those who lack material.

Conclusion

Based on the results of research and discussion, it was concluded that the characteristics of the multiple-choice test items used in the evaluation of learning showed varying results with a score of 2.01 as the hardest item and a score of -4.77 for the easiest item. The same thing also happened to the characteristics of participants who showed varying levels of achievement, namely there were those who showed fit coefficients and some who did not. The majority of participants in the Primary School Examination in Bontomarannu Subdistrict, Gowa Regency were detected as rational response patterns (logical, consistent, and not cheated response patterns). This can be seen after applying the person fit method around 67.9% of the examinees detected fit.

Suggestion

Based on the conclusions of the study, it is recommended that in the holding of the exam it is recommended to apply the person fit method because the method is very useful for controlling the behavior of the examinees in taking the exam. Considering how great the benefits of this person fit method, it is recommended that the person fit method be applied on an ongoing basis in the implementation of exams, both for school exams and for other forms of exams such as the selection exam. To the Gowa Regency Government in order to facilitate teachers especially those who are fully involved in conducting the exam by holding item calibration training based on Item Response Theory (IRT) in general and person fit caliber training specifically so that the implementation of school exams can be better in the future.

Limitation

This research was carried out at the level of learning evaluation activities where the number of participants was still medium. Future research is expected to use a wider and more number of participants.

References

Armstrong, R. D., & Stoumbos, Z. G. (2007). On the performance of the lz person fit statistic. *Practical Assessment Research & Evaluation*, *12*(16), 1–16. https://doi.org/10.7275/xz5d-7j62

Avşar, A. Ş. (2019). Comparison of Person-Fit Statistics for Polytomous Items in Different Test Conditions. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *10*(4), 348–364. https://doi.org/10.21031-epod.525647-830847

Bailey, K. M., & Curtis, A. (2015). *Learning about language assessment: Dilemmas, decisions, and directions (2nd ed.).* Boston, MA: National Geographic Learning.

Baker, F. B. (2001). *The basics of item response theory.* Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Springer.

Çiftçi, S. (2019). Metaphors on open-ended question and multiple-choice tests produced by pre-service classroom teachers. *International Electronic Journal of Elementary Education*, *11*(4), 361–369. https://doi.org/10.26822/iejee.2019450794

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

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

Dardick, W. R., & Weiss, B. A. (2017). Entropy-based measures for person fit in item response theory. *Applied Psychological Measurement*, *41*(1), 1–18. https://doi.org/10.1177/0146621617698945

Ferrando, P. J. (2015). *Assessing person fit in tipical response measures. Handbook of item response theory modeling: Aplications to tipical performance assessment*. New York: Routledge.

Ferrando, Pere J. (2012). Assessing inconsistent responding in E and N measures : An application of person-fit analysis in personality. *Personality and Individual Differences*, *52*(6), 718–722. https://doi.org/10.1016/j.paid.2011.12.036

Fox, J.-P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*(2), 243–262. https://doi.org/10.1111/jedm.12143

Herwin, H., Fathurrohman, F., Wuryandani, W., Dahalan, S. C., Suparlan, S., Firmansyah, F., Kurniawati, K. (2022). Evaluation of structural and measurement models of student satisfaction in online learning. *International Journal of Evaluation and Research in Education, 11*(1), 152-160. http://doi.org/10.11591/ijere.v11i1.22115

Herwin, H., & Mardapi, D. (2017). An emotion assessment model for elementary school students. *Jurnal Penelitian Dan Evaluasi Pendidikan, 21*(1), 80-92. https://doi.org/10.21831/pep.v21i1.14504

Herwin, H., & Nurhayati, R. (2021). Measuring students' curiosity character using confirmatory factor analysis. *European Journal of Educational Research, 10*(2), 773-783. https://doi.org/10.12973/eu-jer.10.2.773

Herwin, H., Phonn, S. (2019). The application of the generalized lord's chi-square method in identifying biased items. *Jurnal Penelitian dan Evaluasi Pendidikan*, *23*(1), 57–67. https://doi.org/10.21831/pep.v23i1.20665

Herwin, H, Tenriawaru, A., & Fane, A. (2019). Math elementary school exam analysis based on the Rasch model. *Jurnal Prima Edukasia*, *7*(2), 106–113. https://doi.org/10.21831/jpe.v7i2.24450

Huang, T. (2012). Aberrance detection powers of the BW and person-fit indices. *Educational Technology & Society*, *15*(1), 28–37.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory*. Illinois: Dorsey Professional Series.

Kılıçkaya, F. (2019). Assessing L2 vocabulary through multiple-choice, matching, gap-fill, and word formation items. *Lublin Studies in Modern Languages and Literature*, *43*(3), 155–166. https://doi.org/10.17951/lsmll.2019.43.3.155-166

Linden, W. J. van der, & 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

Mehrzi, R. A. (2011). Comparison among new residual-based person-fit indices and wright's indices for dichotomous three-parameter IRT model with standardized tests. *Journal of Educational and Psychological Studies*, *4*(2), 14–26. https://doi.org/10.24200/jeps.vol4iss2pp14-26

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Meijer, R. R., & Stoop, V. K. (2001). *Person fit across subgroups: An acheivement testing example. Essay on item response theory*. New York: Springers.

Miejer, R., Niessen, A., & Tendeiro, J. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, *23*(1), 52–62. https://doi.org/10.1177/1073191115577800

Mousavi, A. (2019). An examination of different methods of setting cutoff values in person Fit research. *International Journal of Testing*, *19*, 1–22. https://doi.org/10.1080/15305058.2018.1464010

Pan, T., & Yin, Y. (2017). Using the bayes factors to evaluate person fit in the item response theory. *Applied Measurement in Education*, *30*(3), 213–227. https://doi.org/10.1080/08957347.2017.1316275

Pujiastuti, P., Herwin, H., & Firdaus, F. M. (2021). Thematic learning during the pandemic: CIPP evaluation study. *Cypriot Journal of Educational Sciences, 16*(6), 2970–3980. https://doi.org/10.18844/cjes.v16i6.6481

Reise, S. P. (1990). A comparison of item and person fit methods of assessing model data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127–137. https://doi.org/10.1177/014662169001400202

Rizopoulos, D. (2006). ltm : An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5). https://doi.org/10.18637/jss.v017.i05

Santos, K. C. P., Torre, J. de la, & Davier, M. von. (2019). Adjusting person fit index for skewness in cognitive diagnosis modeling. *Journal of Classification*, 1–22. https://doi.org/10.1007/s00357-019-09325-5

Saptono, B., Herwin, H., & Firmansyah, F. (2021). Web-based evaluation for teacher professional program: Design and development studies. *World Journal on Educational Technology: Current Issues, 13*(4), 672–683. https://doi.org/10.18844/wjet.v13i4.6253

Schmitt, N., Sacco, J. M., Mcfarland, L. A., & Jennings, D. (2015). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*(1), 41–53. https://doi.org/10.1177/01466219922031176

Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times. *Journal of Educational Measurement*, *55*(4), 457–476. https://doi.org/10.1111/jedm.12188

Suleman, Q., Gul, R., Ambrin, S., & Kamran, F. (2015). Factors contributing to examination malpractices at secondary school level in Kohat division, Pakistan. *Journal of Education and Learning, 9*(2), 165-182. https://doi.org/10.11591/edulearn.v9i2.1732

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), 1–27. https://doi.org/10.18637/jss.v074.i05

Tjabolo, S. A., & Herwin, H. (2020). The influence of teacher certification on the performance of elementary school teachers in Gorontalo Province, Indonesia. *International Journal of Instruction, 13*(4), 347–360. https://doi.org/10.29333/iji.2020.13422a

Tjabolo. S. A., Otaya, L. G. (2019). Quality of school exam tests based on Item Response Theory. *Universal Journal of Educational Research, 7*(10), 2156-2164. https://doi.org/10.13189/ujer.2019.071013.

Torre, J. D. L., & 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

Walker, A. A., & Engelhard, G. J. (2015). Exploring person fit with an approach based on multilevel logistic regression. *Applied*

Measurement in Education, *28*(4), 274–291. https://doi.org/10.1080/08957347.2015.1062767

Walker, A. A., Jennings, J. K., & Engelhard, G. (2018). Using person response functions to investigate areas of person misfit related to item characteristics. *Educational Assessment*, *23*(1), 47–68. https://doi.org/10.1080/10627197.2017.1415143

Wind, S. A., & Walker, A. A. (2019). Exploring the correspondence between traditional score resolution methods and person fit indices in rater-mediated writing assessments. *Assessing Writing*, *39*, 25–38. https://doi.org/10.1016/j.asw.2018.12.002

Woods, C. M., Oltmanns, T. F., & Turkheirmer, E. (2008). Detection of aberrant responding on personality scale in a military sample: An aplication of evaluation person fit with two level logistic regression. *Psychological Assessment*, *20*(2), 159–168. https://doi.org/10.1037/1040-3590.20.2.159