Romanian Journal of Applied Psychology
2025, Volume 27, Issue 1, 49-60

Official journal of West University of Timisoara &
The Euroregional Center for Applied Psychology
DOI: 10.2478/rjap-2025-0006

Consistency of aberrant response behavior: Are misfit persons consistent across two different questionnaires administered at the same time?

Muhammad Dwirifqi Kharisma Putra[1] & Faturochman[1]

[1]*Faculty of Psychology, Universitas Gadjah Mada, Yogyakarta, Indonesia*

Received 21.11.2024; Received revised 14.06.2025; Accepted 28.07.2025;
Available online 25.08.2025

Person-fit statistics are important statistics in the modern test theory, which are used to identify respondents with aberrant response behaviors, where, based on statistical evidence, classified as misfit persons. This study aims to determine whether misfit persons are consistent across two different questionnaires that administered simultaneously. The respondents in this study were 723 students in Indonesia. Respondents completed two instruments, namely the 24-item Metacognitive Skills Assessment in Research-Proposal Writing (MSARPW) and the 10-item Rosenberg Self-Esteem Scale (RSES), where the MSARPW was administered in the same Google Form as the RSES. Both instruments were analyzed using the Multidimensional Partial Credit Model (MPCM). The MPCM produced two person fit statistics, which were the focus of this research, namely person Outfit z and Zh statistics. The results of the regression analysis show that the regression coefficients between Zh MSARPW and Zh RSES are 0.375 ($p < .001$) and 0.424 ($p < .001$) for Person Outfit z MSARPW and RSES, respectively. To conclude, with a significant regression coefficient, this finding shows that misfit persons in responding to the MSARPW tend to be misfit persons in responding to the RSES, which indicates that the misfit persons are consistent between different measuring instruments administered at the same administration time. Limitations and implications for future research were also noted.

Keywords: aberrant response, person-fit, outfit statistics, Rasch measurement, Zh statistic

Address of correspondence: Correspondence regarding this article should be directed to Muhammad Dwirifqi Kharisma Putra, Universitas Gadjah Mada, Yogyakarta, Special Capital Region of Yogyakarta, Indonesia
E-mail: muhammad.dwirifqi@ugm.ac.id

Introduction

Modern test theory, such as the Rasch model, is a prominent psychometric framework for evaluating tests and questionnaires, with applications ranging from large-scale educational assessments to small-scale cognitive and noncognitive measurements. Although Rasch model has several practical advantages over classical test theory, the cost of using these models in practice is that inferences drawn from Rasch-based estimates are correct only to the extent that the empirical data meet the sometimes rather restrictive model assumptions, and thus the model fits the data (Crişan et al., 2017).

The application of Rasch models generally involves model fit tests. Standard 3.9 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) requires evidence of model fit when modern test theory, such as the Rasch model, is used to draw conclusions from a dataset (Haberman et al., 2013). To date, extensive research has been conducted on model fit testing in the Rasch model, which includes overall model fit, item fit, and person fit testing. However, although overall model fit and item fit testing have been extensively studied (see Maydeu-Olivares, 2013), person fit testing has been relatively understudied in a limited scope (Meijer et al., 2016).

From the Rasch model perspective, the concept of person fit refers to whether a person's response on test items aligns with the expected value of a model (Wright & Stone, 1999), where based on statistical evidence, the person can be categorized as a fit person which aligns with the expectation of the model or misfit persons which indicates measurement disturbances (Karabatsos, 2000; Zou & Bolt, 2023). Meanwhile, based on psychological or behavioral perspectives, the misfit persons resulted from aberrant response behavior (Meijer, 1996; Meijer et al., 2016; Sijtsma & Meijer, 2001), which is defined as the inconsistency of an individual's response pattern in taking a test (Karabatsos, 2003; Sijtsma & Meijer, 2001).

To understand the form of aberrant response behavior which in turn resulting several misfit persons in the Rasch model, for example, a total of 200 respondents took a 10-item multiple-choice exam with a dichotomous score format (1=true, 0=false) and a total score range of 0-10 (i.e., 0, 1, 2, …, 10). Further, three respondents with different response

patterns were selected. If the items are sorted by their difficulty (from easiest to most difficult), the first person's response pattern is '0000000001', the second person's is '0111111111', and the third person's is '1111100000'. The Rasch model classifies all three of these response patterns as misfits. Why is it interpreted that way?

In the Rasch measurement perspective, the first person does not fit the model because of an inconsistent response pattern called underfit, where this person gives wrong answers on the easy items but answers correctly on the most difficult items, which are far from this person's ability (this person ability is very low which only correct on 1 of 10 items). Such response patterns generally result from guessing or lucky guessing (Karabatsos, 2000; Meijer, 1996; Meijer & Sijtsma, 1995). Meanwhile, the second person had a very high ability (i.e., answering 9 out of 10 items correctly) but incorrectly answered the easiest item. This response pattern, which is also classified as an underfit, results from carelessness (Karabatsos, 2000). Furthermore, the third person has a response pattern that is too consistent or too ideal, which is called overfit. This response pattern is called the Guttman pattern, where the respondent succeeds in answering correctly the items that match their abilities but incorrectly answers the items that are above their abilities (Bond & Fox, 2007). To summarize, the illustration above provides a brief overview of the concept of person fit, where underfit and overfit persons are called misfit persons by statistical evidence based on a specified modern test theory model, such as the Rasch model or Item Response Theory (IRT) in general (Meijer, 1996; Smith, 1986).

Detecting misfit persons caused by aberrant response behavior is an important stage in the evaluation of the psychometric properties of an instrument because there are serious impacts when inconsistent response patterns are ignored (e.g., Karabatsos, 2003; Liu & Liu, 2019b; Panayides & Tymms, 2013; Wind & Schumacker, 2017). A previous study found that even if an instrument is demonstrated to have good psychometric properties, aberrant response patterns can disrupt the instrument's psychometric properties, leading to inaccurate estimation of respondent scores (Meijer, 1996). In modern test theory, the estimates of a person's trait level (person measure) are more reliable than the total score (raw score), but the person measures will be affected by person misfit and provide biased estimates such as underestimated or overestimated (Karabatsos, 2000; Panayides & Tymms, 2013), which means that there will be errors in the interpretation of the ability estimation results (Egberink et al., 2010).

A large body of literature has published numerous approaches for detecting misfit persons (e.g., Liu & Maydeu-Olivares, 2014). The importance of detecting misfit persons is closely linked to the recommendations on test quality assessment guidelines (see International Test Commission, 2014). The recommendation is consistent with the numerous significant negative consequences resulting from misfit persons, as identified in recent research (e.g., Turner & Engelhard, 2024; Zou & Bolt, 2023). To detect misfit persons, several statistics have been developed, known as person-fit statistics (e.g., Meijer, 1996).

Research on person-fit statistics has a long history, dating back to its development within the Rasch measurement perspective (Wright & Stone, 1999), the IRT perspective (Jones et al., 2023; Levine & Rubin, 1979; Meijer, 1996), and factor analysis (Ferrando et al., 2016). This topic is a growing theme in the field of measurement, resulting in a special issue in a journal discussing this concept (see Meijer, 1996). Based on our literature review, in the last 10 years, research on person-fit statistics is still growing, including the application of person-fit analysis to detect misfit persons in measurements in the fields of psychology and education (e.g., Liu & Liu, 2021b; Spoden et al., 2020), personality scales (e.g., Liu et al., 2019a), and psychiatry (e.g., Du et al., 2023). In general, by using person-fit statistics, misfit persons can be detected, identified, and flagged (Meijer & Sijtsma, 2001), their aberrant response can be corrected statistically (e.g., Andrich & Marais, 2014), or the misfit persons can be removed from analysis (e.g., Curtis, 2001; Linacre, 2005), resulting estimates of person trait level or ability that are “free” from aberrant responses (Andrich et al., 2016).

Further, in various research on the evaluation of the psychometric properties of instruments using the Rasch model, the results of person fit analysis are commonly reported (e.g., Du et al., 2023; Wanders et al., 2018), where several studies summarize more than 30 statistics available to test person fit of which the Rasch model is one of the pioneers (e.g., Karabatsos, 2003; Sijtsma & Meijer, 2001). In the Rasch model, the outfit z statistic, which can be used at the item, person, and even response category levels (i.e., Likert scales), is commonly employed (Karabatsos, 2000; 2003; Li & Olejnik, 1997). Apart from that, there is another person-fit statistic that has been widely studied to date, namely lz or its standardized version, namely Zh (Drasgow et al., 1983). A comparison of the accuracy of Outfit z and Zh has also been studied and compared (i.e., Felt et al., 2017), where Outfit z and Zh were found to produce the same conclusions about misfit persons due to aberrant response behavior (Pina & Montesinos, 2005).

However, many studies that compare several person-fit statistics generally focus on computational aspects, which can be challenging to implement for applied researchers in the field of psychology (Meijer et al., 2016). Previous studies provide a tutorial for researchers on the principles of person-fit statistics, aiming to bridge the gap between the application of person-fit testing to applied researchers in the field of psychology (i.e., Ferrando et al., 2016; Meijer et al., 2016). Further, there is a problem of disagreement regarding the decision on flagged misfit persons; there is a study that recommends removing misfit persons and re-analyze the data (Curtis, 2001; Linacre, 2005) to correct the response patterns such as treating the inconsistent response patterns as a missing data (Andrich & Marais, 2014), or study that only flagged the misfit person and not remove it and not re-analyzed the data (Meijer, 2003). There is a need to clarify which procedure needs to be used by the applied researcher. From a practical perspective, in psychological research, test administration is commonly performed by administering more than one instrument that measures various constructs simultaneously in a single administration (cross-sectional). It means that single test administration consists of multiple instruments, which has the impact that respondents must respond to a large number of items, often exceeding 100 items. A large number of items (combined from multiple instruments) was found to be one of the factors that initiated aberrant response behavior, resulting in misfit persons (e.g., Rolstad et al., 2011). Thus, it is necessary to examine whether individuals who are misfits on one test will also be misfits on other instruments (Wanders et al., 2018). However, the tutorial on person-fit analysis (e.g., Ferrando et al., 2016; Meijer et al., 2016) does not address the procedure for assessing whether misfit persons are consistent across multiple instruments.

Based on a literature review, several studies have focused on the consistencies of aberrant responses that

resulting misfit persons across multiple instruments (i.e., Panayides, 2009; Panayides & Tymms, 2012, 2013; Wanders et al., 2018). Some of those studies found that an aberrant response is not a stable characteristic among individuals across multiple measures (Panayides, 2009; Panayides & Tymms, 2012, 2013), which is contrary to other studies that found the aberrant response to be a stable characteristic of individuals (Wanders et al., 2018).

In more detail, in the context of cognitive tests, it was found that although misfits do occur, they do not predict misfits in other tests and are not dependent on the psychological or demographic characteristics of the test-takers (Panayides, 2009). Follow-up studies of the cognitive test also reached the same conclusion that aberrant response is a characteristic that varies between individuals (Panayides & Tymms, 2012, 2013). In contrast, using noncognitive tests with the Likert scale response format, it is found that aberrant responses are a stable characteristic among individuals across two different questionnaires (Wanders et al., 2018). Given the differences in findings between these studies, further research is needed specifically on this topic. Furthermore, based on the findings of the four studies mentioned above, another aspect that needs to be extended is the type of tests used in the three previous studies, which were cognitive tests (i.e., Panayides & Tymms, 2012, 2013). In contrast, measurements in psychology typically employ noncognitive tests, such as questionnaires. To the best of our knowledge, there are few studies (i.e., Conjin et al., 2014; Wanders et al., 2018) that focus on the consistency of person misfit across multiple questionnaires, even though the concept of person-fit analysis of noncognitive tests has been widely discussed and studied long time ago (e.g., Meijer et al., 2016; Reise & Flannery, 1996). Another study also stated that more real-data analyses are needed to demonstrate the usefulness of multiscale person-fit methods for noncognitive multiscale measures (Conjin et al., 2014). As a result, these are an important gap that the current study can address.

In the present study, the stage for testing whether misfit persons tend to be consistent across multiple instruments can be formalized by combining several procedures that were previously scattered in the literature as a separate "step", such as the tabulation of individual responses along with their person-fit statistics (e.g., Meijer, 1996; Meijer & Sijtsma, 1995) to confirm individual response patterns that are detected as misfits, classifying their type of aberrant response (e.g., Meijer et al., 2016), and followed by another stage, namely making decisions to exclude respondents (e.g., Curtis, 2001; 2004; Linacre, 2005) which detected as misfit persons in more than one instruments by more than one fit statistics. That approach contrasts with studies that exclude misfit persons based on person-fit statistics from a single instrument using a single statistic (e.g., Alnahdi & Yada, 2020; Zahra & Wirawan, 2025).

Further, we argue that the formalization of the step-by-step procedure proposed in this present study has the potential to be one of the alternative solutions to the debate among studies that choose to exclude misfit respondents from the analysis (e.g., Curtis, 2001; 2004; Linacre, 2010) or apply correction procedures to response patterns which are not removed the misfit persons from analysis (e.g., Andrich & Marais, 2014; Andrich et al., 2016). We hypothesized that if the proposed procedure were applied, the decision to exclude the misfit persons (e.g., Curtis, 2004) could be made for the "consistent misfit persons" with more robustness than the exclusion based on person-fit statistics from single instruments. Furthermore, the exclusion of misfit persons is also in line with the practice of outlier detection and procedures for removing outliers, which have been commonly used by applied psychology researchers in other statistical methods (see André, 2022). To this end, formalizing the separate steps into a single systematic procedure can serve as a bridge for implementing person-fit analysis, which can be utilized by applied researchers in the field of psychology, in line with the main motivation from previous studies (i.e., Meijer et al., 2016).

Research Questions

This study aims to test whether misfit persons' results from their aberrant response behavior are consistent between more than one measuring instrument administered in the same test administration. To answer this, two person-fit statistics, namely Outfit z and Zh, will be compared to confirm whether the same findings also apply to two different statistics based on the same model, the Rasch measurement model. More specifically, the following questions were addressed:

RQ1. Are misfit persons found to be consistent across two different measuring instruments based on statistical evidence?

RQ2. Do two different person-fit statistics, Outfit z, and Zh, provide the same information regarding the consistency or inconsistency of aberrant response behavior, which results in numerous misfit persons?

Methods

Participants

Data were gathered from 723 university students (409 males [56.5%], 314 females [43.4%]), who were recruited from 10 universities across Indonesia. The mean age of the sample was 26.841 years (SD = 7.384). All of the respondents were active students in their universities' faculties of education, with 476 (65.8%) being undergraduate students, 183 (25.3%) being graduate students, and 64 (8.8%) being doctoral students. The data were collected using an online approach using Google Form as a part of larger project on metacognitive model of self-esteem. Students were given information regarding the general aim of the study, and were assured that the data would be handled in a manner that protected their privacy. All students participated on a voluntary basis, and no incentive or compensation was provided to them for their participation. Also, informed consent was obtained from all individual participants included in the study.

Instruments

In this research, there are two instruments used, namely the Metacognitive Skills Assessment in Research-Proposal Writing (MSARPW; Hayat et al., 2023) and the Indonesian version of the RSES. Both instrument were chosen as a part of the larger project and have nonsignificant scores correlation each other. The MSARPW contains a 24-item test with a 4-option Likert scale format. This instrument measures metacognition in the context of writing research proposals. There are two aspects measured through this instrument, namely regulation of cognition (12-items) and knowledge of cognition (12-items). This instrument has a multidimensional factor structure (two-correlated factor model). Examples of items from MSARPW are " *I know an effective strategy for aligning the research title, statement of the problem, research objectives and hypotheses to be tested*" and " *I have a well-planned strategy for selecting appropriate research methods for answering the research*

questions that have been formulated". Reliability of the scale by Cronbach's Alpha are 0.875 (regulation of cognition) and 0.893 (knowledge of cognition).

The second instrument used was the Rosenberg Self-esteem Scale (RSES; Rosenberg, 1965). This instrument contains a 10-item to measure global self-esteem. This instrument was adapted into an Indonesian language by Maroqi (2018). Response options in the RSES are a 4-point Likert scale from "0 = strongly disagree" to "3 = strongly agree". Even though it was initially developed with a unidimensional model, over time it was discovered that this instrument has a two-factor multidimensional structure where items 1 to 5 measure aspects of self-competence and items 6 to 10 measure aspects of self-liking (Ogihara & Kosumi, 2020). Examples of items from the RSES are "*I am able to do things as well as most other people*" and "*At times, I think I am no good at all*". Reliability of the scale by Cronbach's Alpha are 0.721 (self-competence) and 0.775 (self-liking).

Multidimensional Partial Credit Model (MPCM)

In this study, we used the Multidimensional Partial Credit Model (MPCM), a form of the MRCMLM (see Adams et al., 1997 for detailed explanations), to estimate item and person parameters of the MSARPW and the Indonesian RSES. The formula for MPCM is as follows:

$$P(u_{ij} = \mathrm{k}|\theta_{\mathrm{j}}) = \frac{e^{\sum_{\ell=1}^{k}(\theta_{j\ell} - b_{i\ell k})}}{1 + \sum_{r=1}^{K_i} e^{(\theta_{j\ell} - b_{i\ell r})}}$$

where $b_{i\ell k}$ is the difficulty parameter for item i on dimension l for score category k, and $\theta_{j\ell}$ is the trait level of person (Wang et al., 2004). There is an indeterminacy in the estimation of the $b_{i\ell k}$ parameters, then that parameters are set equal across the response categories, k=1,2,…, K_i (Adams et al., 1997; Wang et al., 2004). The item fit statistics which use in this study are Infit and Outfit mean square statistics (Infit and Outfit MNSQ) when the value is in the range of 0.7-1.3 indicating acceptable fit criteria and the values outside of these bounds may suggest a lack of fit between the item to the model (Karabatsos, 1998). As for this study, overall model fit or goodness-of-fit will be reported with the M_2 statistic, which if the p-value is not significant, indicates that the model is fit; RMSEA, which if the value is < 0.06, indicates that the model is fit, CFI and TLI, which if the value is > 0.90 shows that the model is fit and SRMR, if the value is <0.08, indicates that the model is fit (Tay et al., 2015).

Person-fit statistics

In this study, there are two person-fit statistics that used, namely Outfit z and Zh statistics. The reason why these two statistics are used in the present study is that Outfit z is a person fit statistic that is specifically developed in the Rasch measurement framework, while Zh is one of the most widely used model-based person-fit statistics (Conjin et al., 2014; Karabatsos, 2003; Meijer & Tendeiro, 2012). In addition, there have been previous studies that have conducted direct comparisons between the two statistics. Both statistics are available in standard software packages for Rasch analysis.

Person Outfit z, similar to the Outfit MNSQ and Outfit z statistics for items, are a parametric statistic calculated from the residual (the difference between the observed and expected response values) for each person taking each item. The Person Outfit MNSQ and z formula is as follows:

$$z_{ni} = \frac{\left(X_{ni} - E(X_{ni})\right)}{\sqrt{Var(X_{ni})}}$$

$$Outfit_n = \frac{\sum_{i=1}^{I} {z_{ni}}^2}{N}$$

As can be seen in the two formulas above, Outfit MNSQ is calculated from the standardized residual for respondent n and item i which is briefly the average of the sum of the squared residuals (i.e. unweighted). This value can be transformed into a standardized form with the Wilson-Hilferty transformation so that it is distributed in the form of a Student t-distributed distribution with infinite degrees of freedom (i.e. standard normal distribution), if the Rasch model holds (Artner, 2016). This means that the interpreted significance of Outfit z is 1.96, where respondents who have values above 1.96 and -1.96 are misfit respondents.

Meanwhile, the second person-fit statistic is the Zh statistic. Zh is the standardized person fit value of lz which is used in categorical data such as multiple choice items or Likert scales (Drasgow et al., 1985). The formula for *lz* statistics is (Felt et al., 2017):

$$lz = P(Y_i|\theta_i) = \sum_{i=1}^{n}\sum_{j=1}^{A+1} \delta_j(v_i) log\ P_g(\theta)$$

where Y_i represents the item responses, θ_i is contains the latent trait estimates, n represents the sample size, A is the number of j categories in item i. The random vector of item choices, denoted as $\delta_j(v_i)$, is used to ensure that only the probabilities of the chosen responses are summed. Thus, $\delta_j(v_i)$ is set equal to 1 when j = k, and it is set to 0 when j ≠ k. Finally, the standardized form of lz can be defined as Zh with the following transformation (Felt et al., 2017):

$$Zh = [lz(\theta) - E(lz(\theta))/SD(lz(\theta))]$$

where E(lz(θ)) represents the mean *lz*-value for the sample, and SD(lz(θ)) represents the standard deviation. This transformation simply standardizes the value to have a mean of 0 and variance of 1 by dividing the difference between lz and mean of lz with the standard deviation of the observed *lz*-value. The criterion that indicates that a respondent is not fit (misfit) is if the Zh value is greater than ±1.96 with a 0.5 significance level (α) (Conjin et al., 2014; Felt et al., 2017). Zh and Outfit z can be compared with each other considering that the units of measurement have been standardized.

Methodology Proposal: Examining person-fit consistencies

There are five stages that we proposed in testing person-fit consistencies (see Figure 1). In the first stage, the analysis begins with calibrating two instruments (MSARPW and Indonesian RSES) separately with MPCM. Second, based on the results of the MPCM analysis, Outfit z and Zh statistics are produced for each respondent. Each respondent has Outfit z and Zh values for each instrument, namely Outfit z and Zh-MSARPW and Outfit z and Zh-RSES.

There is an important point that researchers need to understand which is given that MSARPW and RSES have multidimensional models, the Outfit z and Zh statistics produced are only one for each instrument and not for each aspect. This also applies to Zh where Zh-MSARPW and Zh-RSES will be generated.

In the third stage, after person fit statistics were generated for each person in each instrument, we carried out response pattern analysis by tabulating each individual's response pattern to ensure that respondents who were marked as misfit by the two statistics (Outfit z and Zh) actually had the inconsistent response pattern. Response pattern analysis procedures in these steps have been explained comprehensively in both cognitive tests and non-cognitive tests in previous studies (i.e., Emons et al., 2005; Meijer, 1996; Meijer et al., 2016).

Next, in the fourth stage, to test whether person misfit is consistent between two different instruments, a simple linear regression was carried out between Outfit z-MSARPW against Outfit z-RSES as well as MSARPW. The MSARPW person fit statistics were used as the independent variable (X) because this instrument was presented first to respondents, while the RSES person fit statistics were used as the dependent variable (Y) because they were administered after than the MSARPW. A significant regression coefficient shows that misfit persons on the MSARPW tend to be misfit on the RSES. This means that these respondents are consistently misfit persons.

In the fifth stage, after the regression analysis is conducted, if a positive standardized regression coefficient (Beta) is found from instrument 1 (MSARPW) to instrument 2 (RSES), further analysis is carried out in the form of Outfit z and Zh differences testing for the two instruments. In simple terms, this procedure involves computing differences in person-fit statistics for each person between MSARPW and RSES. Then, from this difference, the p-value can be calculated, which shows whether there is a significant difference between the person fit statistics on MSARPW and their person fit statistics on RSES. Significant findings indicate that whether a person is fit or not differs between measuring instruments. Meanwhile, if it is significant, there is no significant difference between the person's individual fit statistics value on the two measuring instruments. In the final stage, adjustments are made to the p-value of this difference using the False Discovery Rate (FDR; Yekutieli & Benjamini, 1999) method to achieve a more reliable p-value.

In this study, all MPCM analysis, including generating Outfit z and Zh statistics, was carried out with the '*mirt*' package (Chalmers, 2012) using the marginal maximum likelihood estimation method. Additionally, the simple linear regression analysis was carried out with the '*lavaan*' package (Rosseel, 2012), and the post-hoc power analysis was performed using the '*semPower*' package (Moshagen & Bader, 2024). All of the packages were implemented using the RStudio program version 4.4.0 (RCore Team, 2015).

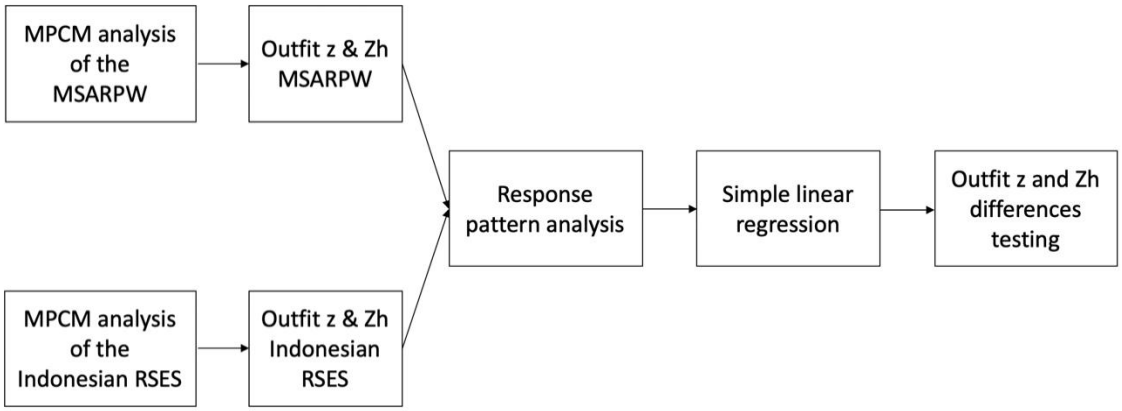

Figure 1. Steps of person-fit analysis for two separate measures

Results

Item calibration

In the first stage, separate calibration was carried out on two instruments, namely MSARPW and Indonesian RSES. This analysis was carried out to determine whether there was a model misspecification. Furthermore, after the data is found to be fit the model, Outfit z and Zh fit statistics are generated for each person on each measuring instrument. The MPCM goodness-of-fit statistics for both instruments can be seen in Table 1.

Based on the information in Table 1, it is found that MSARPW fits MPCM with a two-factor solution, as does Indonesian RSES, which also fits a two-factor model. After obtaining statistical evidence about the overall model fit, we examined the fit statistics of the individual items to the model. It was found that all MSARPW and RSES items fit the model, with MNSQ Infit and Outfit values in the range of 0.70-1.30. Thus, it can be concluded that the psychometric properties of the MSARPW and Indonesian RSES instruments meet the standards set by the Rasch model. Therefore, Outfit z and Zh can be generated for each respondent without any impact from model misspecification.

Person-fit analysis

The results of the MPCM analysis produce two person fit statistics for each instrument. Based on the cut-off Outfit z and Zh (± 1.96) we identified respondents which fit and misfit according to the model. The number of fit and misfit persons based on each statistic can be seen in Table 2.

As can be seen in Table 2, it is known that Outfit Z detected 48 misfit persons on the MSARPW measuring instrument while Zh detected 17 respondents where there is a difference in the number caused by the different mathematical basis of these two statistics. The same thing was also found in the Indonesian RSES instrument. In general, Zh detects a smaller number of misfit persons compared to Outfit Z. The distribution of Outfit z and Zh for all instruments can be seen in Figure 2.

Table 1. Goodness-of-fit statistics of MSARPW and Indonesian RSES

Fit Indices	MSARPW (two-factor model)	Indonesian RSES (two-factor model)
M_2	559.494	70.123
df	225	22
p-value of M_2	0.000	0.000
RMSEA	0.045	0.055
90% C.I. of RMSEA	0.041, 0.050	0.041, 0.070
SRMR	0.062	0.058
CFI	0.967	0.938
TLI	0.968	0.929

Table 2. Frequency of fit and misfit persons based on Outfit z and Zh

Instruments	Outfit Z		Zh	
	Fit	Misfit	Fit	Misfit
MSARPW	675	48	706	17
Indonesian RSES	656	67	695	28

As can be seen in Figure 2, although Zh and Outfit z have the same acceptable criteria (± 1.96), the interpretation of the positive and negative directions is different. In Outfit z, respondents who have Outfit z > -1.96 are respondents who are overfit (too ideal), whereas in Zh if the value is increasingly negative or > -1.96 then the respondent is an underfit respondent (too unpredictable). Based on these facts, researchers are expected to be careful in interpreting the direction of two different statistics even though the criteria are the same. The correlation matrix of the outfit z and Zh for both instruments can be seen in Table 3.

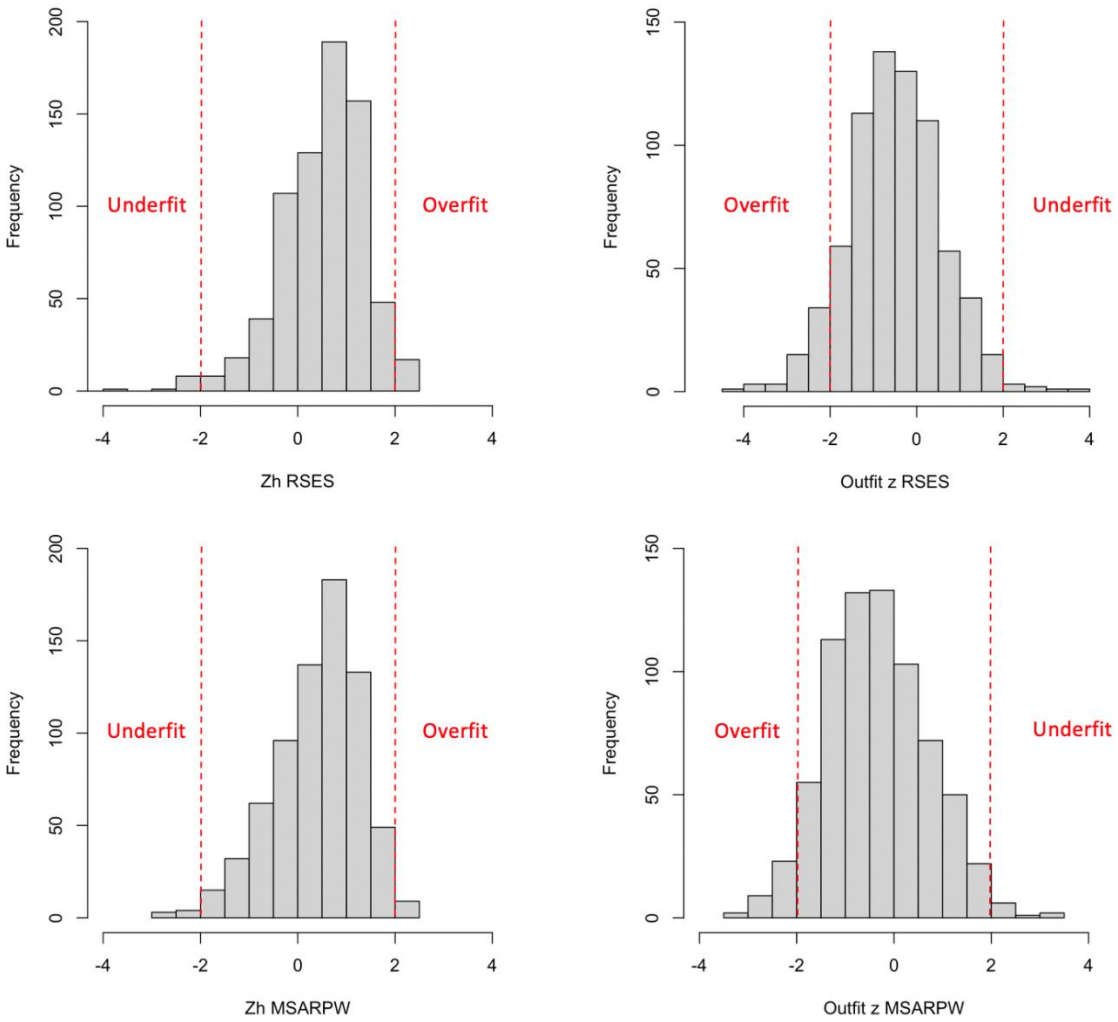

Figure 2. Distribution of Zh dan Outfit z statistics of MSARPW and RSES

Table 3. Correlation matrix of the Outfit z and Zh

Person fit statistics	Zh RSES	Zh MSARPW	Outfit z RSES	Outfit z MSARPW
Zh RSES	1			
Zh MSARPW	0.375	1		
Outfit z RSES	-0.838	-0.398	1	
Outfit z MSARPW	-0.400	-0.919	0.424	1

Note. *all correlations are significant*

As can be seen in Table 3, it is clear that Outfit Z and Zh were negatively correlated. The negative correlations mean that the way of interpreting the Outfit Z and Zh directions is opposite to each other, as shown in Figure 2. Consequently, researchers need to carefully interpret both statistics to avoid incorrect decisions (i.e., the opposite direction of Outfit z vs. Zh). Previous research has provided a comprehensive explanation of the procedures for interpreting Zh statistics, which psychology researchers can apply (Felt et al., 2017). Additionally, other research has demonstrated how to report person Outfit z statistics in health sciences instruments (Du et al., 2023). Both studies can serve as a reference for applied psychological researchers.

Response pattern analysis

The next information is an examination of the response pattern of the individual who is detected as a misfit person or which means that the individual gave an aberrant response. Table 4 contains examples of response patterns for three individuals who were detected as misfit based on Zh and Outfit z on the RSES instrument where we sorted the items from easiest to agree with (far left) to most difficult to agree with.

Table 4. Item score patterns and person-fit statistics for the Indonesian RSES (aspect 1)

Person ID	Theta	Items of aspect 1 (*ordered by difficulty*)					Outfit z	Zh
		Item 2	Item 5	Item 4	Item 1	Item 3		
212	-0.249	3	0	0	0	3	3.587	-3.680
707	-0.918	0	0	0	0	3	2.826	-2.407
722	0.956	0	3	3	3	3	2.162	-2.197
	M_{scores}	1.631	1.478	1.474	1.437	1.402		

In Table 4, there are three examples of response patterns that were detected as misfit persons by the Outfit Z and Zh statistics. All three persons were classified as misfits based on their aberrant response behavior. Both are underfitted, which means the response pattern is not consistent (too random). The first person with ID 212 (P212) had a theta level of -0.249, where the model expected P212 to respond 'agree' in four items from easiest to difficult (item 2, item 5, item 4, item 1). However, in the data, P212 answered 'strongly agree' to the easiest item (item 2) and the most difficult item (item 3) while responding 'strongly disagree' to the other items (item 5, item 4, item 1). This is what causes the P212 to be misfit person (Outfit z = 3.587 > 1.96, Zh = -3.680 > -1.96).

Furthermore, the second person with ID 707 (P707) had a theta level of -0.918, where P707 was expected by the model to respond 'agree' in four items from easiest to difficult (item 2, item 5, item 4, item 1), but P707 answered 'strongly disagree' on four items that were expected to be answered 'agree' (item 2, item 5, item 4, item 1) and instead answered 'strongly agree' to the most difficult item (item 3). This is what causes the P707 to be misfit person (Outfit z = 2.826 > 1.96, Zh = -2.407 > -1.96).

As for the last one, the third person with ID 722 (P722) has a theta level of 0.722, where the model expects P722 to agree with all the items from the easiest to the most difficult (item 2, item 5, item 4, item 1, item 3). However, in the data, P722 answered 'strongly agree' on four items and instead answered 'strongly disagree' on the easiest item (item 2); this is contrary to the expectations of the model (expected values). This is what causes the P722 to be misfit person (Outfit z = 2.162 > 1.96, Zh = -2.197 > -1.96). In the next phase of the analysis, the main hypothesis testing regarding whether a misfit person is consistent across two different instruments administered in the same test administration is reported, followed by Outfit z and Zh differences testing.

Tabel 5. Simple regression analysis results

Path	β	S.E.	z	p-value	R^2
Outfit Z MSARPW → Outfit Z RSES	0.424	0.029	14.567	0.000	0.180
Zh MSARPW → Zh RSES	0.375	0.031	12.177	0.000	0.141

Regression analysis

Simple linear regression analysis was conducted twice separately (see Table 5) due to violations of multicollinearity in the Outfit z and Zh if the regression analysis was performed concurrently. The results of the regression analysis show that the effect of Outfit Z MSARPW on Outfit Z RSES is significant in the positive

direction. This finding means that persons who are classified as misfits on the MSARPW tend to be a misfit person on the RSES (β = 0.424, p = .000). Furthermore, using different person fit statistics, namely Zh, it was found that the effect of Zh MSARPW on Zh RSES was significant in the positive direction. This means that persons who are classified as misfits on the MSARPW tend to be a misfit person on the RSES (β = 0.375, p = .000). Thus, based on both regression coefficients, it can be concluded that Outfit Z and Zh provide consistent findings, where these findings indicate that misfit persons are consistent between two different instruments across two different person-fit statistics.

Furthermore, from the two regression models above, it is evident that the R-squared (R^2) values for both models are significant. This finding indicates that the proportion of explained variance of the person fit statistics (Outfit z and Zh) in the RSES is explained by the variance of Outfit z or Zh on the MSARPW instrument. Additionally, we also conducted a post hoc power analysis of the two regression models in Table 5 and found that the power for both analyses was 1.000. This finding suggests that the sample size of 723 in this study had sufficient power to detect the two regression effects.

Outfit z and Zh differences testing

In the last stage, the Outfit z differences and Zh differences testing was carried out to find out whether there were significant differences between each person's Outfit z and each person's Zh on two different measuring instruments. The results of the analysis in the form of the distribution of p-value Outfit and Zh differences can be seen in Figure 3.

As can be seen in Figure 3, the number of respondents who have significant differences in Outfit z and Zh ($p < 0.05$) is very low (red bar). This means that respondents tend to have consistent person fit on the two instruments, where this finding shows that respondents which classified as misfit on the MSARPW are also classified as misfits on the RSES. Conversely, this finding also applies to fit person, where person that fit in responding the MSARPW are also fit in RSES, where only a small percentage are vice versa (fit in MSARPW but misfit in RSES). This finding emphasizes the consistency regarding person misfit which has been proven to be consistent between two different measuring instruments administered at the same administration time.

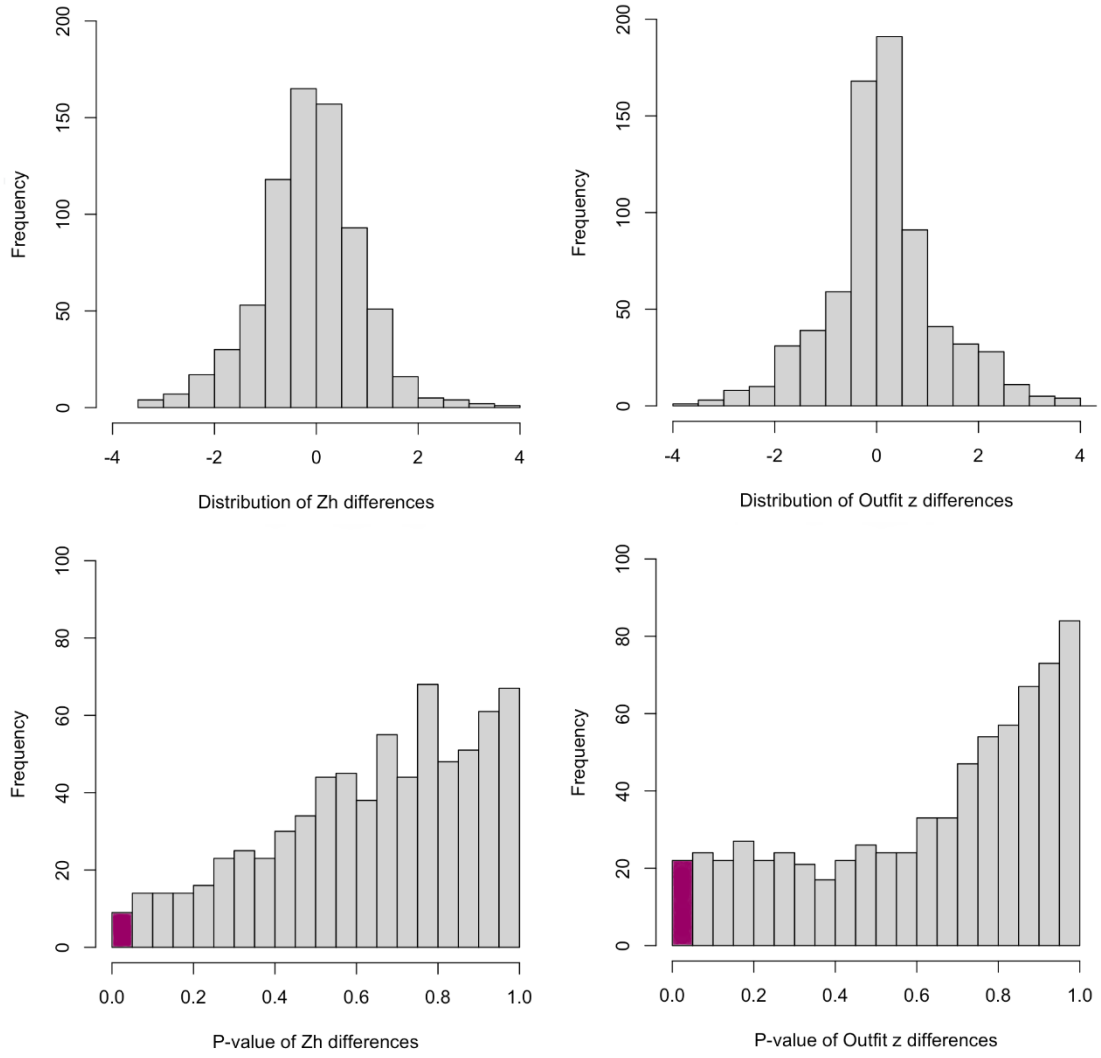

Figure 3. Distribution of p-value from differences testing

Discussion

This study aims to test whether misfit persons are consistent across two different measuring instruments and to determine if two different statistics (i.e., Outfit z and Zh) yield consistent findings or the same conclusion. To this end, we formalized the five-stage procedure used to test intended consistency. Based on the five-stage procedure, we found that respondents classified as misfit persons on one instrument also tended to be misfit persons on other instruments administered simultaneously. This finding aligns with the study conducted by Wanders et al. (2018), which reported similar findings and discussed them in depth. Additionally, the test used in that study was also non-cognitive, highlighting the similarity between the two studies. However, that study used linear correlation for testing consistency and also only employed one person-fit statistic, namely Zh. In contrast, the present study used linear regression with adjusted p-values and two person-fit statistics (Outfit z and Zh).

However, our findings are not in line with previous studies, which found that an aberrant response is not a stable characteristic among individuals (Panayides, 2009; Panayides & Tymms, 2012, 2013). Although these studies also used Outfit z, the method used to test consistency was the chi-square test, which differs from the one used in the present study. In addition, the three studies employed cognitive tests or maximum performance tests that differed from those used in the present study. Therefore, further study is needed to determine whether the procedure we formalized also yields consistent findings on cognitive tests.

In addition, when comparing Outfit Z and Zh, we found that both fit statistics yield the same conclusion about the consistency of misfit persons across two different questionnaires. The comparison of Outfit z and Zh statistics can be perform directly because they both follow a normal distribution (Karabatsos, 2003; Li & Olejnik, 1997). Our finding complements the findings of previous studies that conducted direct comparisons between Outfit z and lz or Zh (Goldammer et al., 2024; Jones et al., 2023), although these studies have not specifically tested the consistency of misfit persons across two different questionnaires. This means that our study can serve as an initial reference for testing two different person-fit statistics across two different questionnaires. Additionally, it is worth noting that other studies have evaluated more than 10 fit statistics simultaneously and are more comprehensive than our study (e.g., Goldammer et al., 2024; Karabatsos, 2003). This is another limitation of the present study.

In addition, it is essential to acknowledge that the present study has various constraints and limitations; therefore, we do not want our findings to be overgeneralized without further in-depth studies to address these constraints. The first constraint is that only one model is used, namely MPCM, and the model is found to be a good fit in both instruments. Based on the literature review, many previous studies have tested the condition of the poor-fitting model to better understand its relationship with misfit persons (i.e., Hong et al., 2020; Meijer & Tendeiro, 2012). Conversely, our study is limited only to the well-fitting model conditions. Regarding the use of MPCM only, we followed the model selection based on previous studies that had only used PCM in their assessment of person fit consistency (i.e., Panayides & Tymms, 2013). Thus, future studies can compare various models other than PCM to test whether the procedure that we formalized applies to other models (e.g., IRT).

Additionally, some challenges should be noted. As stated in previous studies, interpreting misfit persons is the biggest challenge (Meijer & Tendeiro, 2012). we performed individual response pattern analysis and found information that the underfit pattern occurred because respondents answered “strongly agree” to items that were difficult to agree with and were far above their trait level (i.e., self-esteem level). This finding aligns with previous studies, which have found that this pattern constitutes a form of aberrant response (Karabatsos, 2000; Meijer et al., 2016). In the present study, many forms aberrant response patterns which resulting misfit persons could be explained through response pattern analysis in both instruments. However, our study did not test specific types of aberrant responses, even though previous studies have identified various forms, such as guessing, sleeping, constant response set, unmotivated, and sabotaging (see Karabatsos, 2000; Liu et al., 2019; Meijer, 1996). Future studies can extend our findings by reviewing item content and individual responses in more detailed to draw conclusions about various types of aberrant responses.

From a statistical perspective, to sharpen the findings regarding consistency, a difference testing analysis on Outfit z and Zh statistics reveals that the number of respondents with significant Outfit z and Zh differences is very low. The p-value resulting from this study has been adjusted as suggested by previous research (e.g., Yekutieli & Benjamini, 1999). Thus, these findings provide us with more in-depth statistical evidence about the consistency of misfit and fit persons.

However, with few previous studies discussing person misfit consistency, we suspect that person misfit consistency may be caused by several causes that have not been tested in this study but are widely discussed in previous studies, such as test-taking motivation (Lundgren & Eklof, 2023), the number of items in the instrument (Burchell & Marsh, 1992), response burden, namely the effort required by the respondent to answer a questionnaire (Briz-Redon, 2021; Rolstad et al., 2011), as well as the number of instruments administered in a single administration. To determine this, future research can collect data such as response times (RT) to clarify the amount of time respondents spend on each item (e.g., van der Linden & van Krimpen-Stoop, 2003). With RT data, it will be evident that the aberrant response pattern was produced in a short time, allowing it to be concluded that the response given by the person was random or aberrant. This is a promising topic for further exploration of our findings.

Ultimately, we hope that the five-step procedure formalized in this study can overcome the limited implementation of person-fit analysis by applied psychology researchers, as stated in previous studies. Previous studies have explained that the person-fit analysis procedure for detecting inconsistent responses is too difficult for nonspecialist researchers to understand (Meijer et al., 2016). Through our study, we developed a procedure that can serve as a reference for the application of person-fit analysis, particularly for nonspecialist researchers. Apart from that, we feel that by using Outfit z and Zh, there is a practical advantage that respondents that classified as misfits if the z statistics values are > ±1.96 with a 0.5 significance level (a) where these criteria tend to be similar to the item fit criteria (or general significance testing of z-statistics) which are more widely

known to nonspecialist researchers.

Finally, from a practical perspective, our procedure can be used as the basis for excluding "consistent misfit persons" (persons who misfit on both instruments), which is more robust than excluding misfit persons on only one instrument with only one statistic (i.e., Alnahdi & Yada, 2020; Zahra & Wirawan, 2025). What we mean by "robust" is that "consistent misfit persons" are identified using a five-step process that provides more detailed statistical evidence based on multiple person-fit statistics and multiple instruments.

Conclusion

In conclusion, based on statistical evidence, we found that individuals who are misfits on one instrument also tend to be misfits on other instruments that are administered simultaneously, indicating the consistency of misfit persons. Both person-fit statistics, Outfit z and Zh, were also found to draw the same conclusions, indicating the consistency of different person-fit statistics. Furthermore, we formalized a systematic procedure that includes a five-step data analysis process for assessing the consistency of misfit persons; although this is not a novel concept, this finding is significant considering that in practice, applied psychology researchers commonly collect data on various constructs involving multiple instruments in the same test administration, where the five-step procedure can be used in that condition to assess the consistency of person misfits. It is hoped that this five-step procedure can serve as a bridge for applied psychological researchers to extend our findings by using the five-step procedures and studying the consistency of aberrant response behavior in the future.

Declaration of Interest Statement

The authors declare that there are no conflicts of interest with respect to the authorship of this paper.

Funding

There is no funding to report in association with this paper.

Ethical Approval

The ethical approval was granted by the Institute for Research and Community Services (LPPM; Lembaga Penelitian dan Pengabdian Pada Masyarakat), Universitas Negeri Jakarta, Jakarta, Indonesia (No. B/680/UN39.14/PJ/X/2022).

Data Availability Statement

Full dataset including raw data, software code, and output was available online in the anonymous materials located on the open science framework (OSF) with the link of: https://osf.io/eqh9j/?view_only=1871348c03564f20a696f5e4338823b9

References

Adams, R.J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23. https://doi.org/10.1177/0146621697211001

Alnahdi, G. H., & Yada, A. (2020). Rasch analysis of the Japanese version of Teacher Efficacy for Inclusive Practices Scale: Scale unidimensionality. *Frontiers in Psychology, 11:* 1725. https://doi.org/10.3389/fpsyg.2020.01725

American Educational Research Association (AERA), American Psychological Association (APA), National Council for Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

André, Q. (2022). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General, 151*(1), 213–223. https://doi.org/10.1037/xge0001069

Andrich, D., & Marais, I. (2014). Person proficiency estimates in the dichotomous rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement, 38*(6), 432-449. https://doi.org/10.1177/0146621614529646

Andrich, D., Marais, I., & Humphry, S. (2016). Controlling guessing bias in the dichotomous Rasch model applied to a large-scale, vertically scaled testing program. *Educational and Psychological Measurement, 76*(3), 412-435. https://doi.org/10.1177/0013164415594202

Artner, R. (2016). A simulation study of person-fit in the Rasch model. *Psychological Test and Assessment Modeling*, *58*(3), 531–563.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates.

Briz-Redon, A. (2021). Respondent burden effects on item non-response and careless response rates: An analysis of two types of surveys. *Mathematics, 9*(17), 2035. https://doi.org/10.3390/math9172035

Burchell, B., & Marsh, C. (1992). The effect of questionnaire length on survey response. *Quality and Quantity, 26*(3), 233-244. https://doi.org/10.1007/BF00172427

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Conjin, J. M., Emons, W. H. M., & Sijtsma, K. (2014). Statistic lz-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement, 38*(2), 122-136. https://doi.org/10.1177/0146621613497568

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement, 41*(6), 439–455. https://doi.org/10.1177/0146621617695522

Curtis, D. D. (2001). Misfits: People and their problems. What might it all mean? *International Education Journal, 2*(4), 91-99.

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*(2), 125-144.

Drasgow, F., Levine, M. V., & William, 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

Du, J., Wang, Y., Wu, A., Jiang, Y., Duan, Y., Geng, W., Wan, L., Li, J., Hu, J., Jiang, J., Shi, L., & Wei, J. (2024). The validity and IRT psychometric analysis of Chinese version of Difficult Doctor-Patient Relationship Questionnaire (DDPRQ-10). *BMC Psychiatry, 23*: 900. https://doi.org/10.1186/s12888-023-05385-5

Egberink, I. J. L., Meijer, R. R., Veldkamp, B. P., Schakel, L., & Smid, N. G. (2010). Detection of aberrant item score patterns in computerized adaptive testing: An empirical example using the CUSUM. *Personality and Individual Differences, 48*(8), 921-925. https://doi.org/10.1016/j.paid.2010.02.023

Emons, M. H. W., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person fit analysis using person-response functions. *Psychological Methods*, *10*(1), 101-119. https://doi.org/10.1037/1082-989X.10.1.101

Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology, 8:* 863. https://doi.org/10.3389/fpsyg.2017.00863

Ferrando, P. J. (2015). Assessing person fit in typical-response measures. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 128–155). Routledge/Taylor & Francis Group.

Ferrando, P. J., Vigil-Colet, A., & Lorenzo-Seva, U. (2016). Practical person-fit assessment with the linear FA model: New developments and a comparative study. *Frontiers in Psychology*, 7: 1973. https://doi.org/10.3389/fpsyg.2016.01973

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika, 78*(3), 417–440. https://doi.org/10.1007/s11336-012-9305-1

Hayat, B., Rahayu, W., Putra, M. D. K., Sarifah, I., Puri, V. G. S., & Isa, K. (2023). Metacognitive Skills Assessment in Research-Proposal Writing (MSARPW) in the Indonesian university context: Scale development and validation using multidimensional item response models. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia, 12*(1), 31-47. https://doi.org/10.15408/jp3i.v12i1.31679

Hong, S. E., Monroe, S., & Falk, C. F. (2020). Performance of person-fit statistics under model misspecification. *Journal of Educational Measurement,* 57(3), 423-442. https://doi.org/10.1111/jedm.12207

International Test Commission (ITC). (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing,* 14(3), 195-217. https://doi.org/10.1080/15305058.2014.918040

Jones, E. A., Wind, S. A., Tsai, C-L., & Ge, Y. (2023). Comparing person-fit and traditional indices across careless response patterns in surveys. *Applied Psychological Measurement, 47*(5-6), 365-385. https://doi.org/10.1177/01466216231194358

Karabatsos, G. (1998). Analyzing nonadditive conjoint structures: Compounding events by Rasch model probabilities. *Journal of Outcome Measurement, 2*(3), 191-221.

Karabatsos, G. (2000). A critique of rasch residual fit statistics. *Journal of Applied Measurement*, 1(2), 152-176.

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

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/10.2307/1164595

Linacre, J. M. (2005). When to stop removing items and persons in Rasch analysis? *Rasch Measurement Transactions*, *23*(4), 1241.

Li, M.-n. F., & Olejnik, S. (1997). The power of Rasch person–fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*(3), 215–231. https://doi.org/10.1177/01466216970213002

Liu, Y., & Maydeu Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*, *49(*4), 354-371. https://doi.org/10.1080/00273171.2014.910744

Liu, Y., & Liu, H. (2021). Detecting noneffortful responses based on a residual method using an iterative purification process. *Journal of Educational and Behavioral Statistics, 46*(6), 717-752. https://doi.org/10.3102/1076998621994366

Liu, T., Lan, T., & Xin, T. (2019a). Detecting random responses in a personality scale using IRT-based person-fit indices. *European Journal of Psychological Assessment*, 35(1), 126-136. https://doi.org/10.1027/1015-5759/a000369

Liu, T., Sun, Y., Li, Z., & Xin, T. (2019b). The impact of aberrant response on reliability and validity. *Measurement: Interdisciplinary Research and Perspectives,* 17(3), 133-142. https://doi.org/10.1080/15366367.2019.1584848

Lundgren, E., & Eklof, H. (2023). Questionnaire-taking motivation: Using response times to assess motivation to optimize on the PISA 2018 student questionnaire. *International Journal of Testing,* 23(4), 231-256. https://doi.org/10.1080/15305058.2023.2214647

Maroqi, N. (2018). Uji validitas konstruk pada instrumen Rosenberg Self-Esteem Scale dengan metode confirmatory factor analysis (CFA). *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia, 7*(2), 92-96. https://doi.org/10.15408/jp3i.v7i2.12101

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

Maydeu-Olivares, A. (2013). What should we assess the goodness of fit of IRT models? *Measurement*, *11*(3), 127-137. https://doi.org/10.1080/15366367.2013.841511

Meijer, R., R. (1996). Person fit research: An introduction. *Applied Measurement In Education*, 9(1), 3-8. https://doi.org/10.1207/s15324818ame0901_2

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*(1), 72–87. https://doi.org/10.1037/1082-989X.8.1.72

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8*(3), 261–272. https://doi.org/10.1207/s15324818ame0803_5

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., & Tendeiro, J. N. (2012). The use of the lz and lz* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics, 37*(6), 758-766. https://doi.org/10.3102/1076998612466144

Meijer, R. R., Niessen, M. S. A., & Tendeiro, N. 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), 56-62. https://doi.org/10.1177/1073191115577800

Moshagen, M., & Bader, M. (2024). semPower: General power analysis for structural equation models. *Behavior Research Methods, 56*(4), 2901-2922. https://doi.org/10.3758/s13428-023-02254-7

Ogihara, Y., & Kusumi, T. (2020). The developmental trajectory of self-esteem across the life span in Japan: Age differences in scores on the Rosenberg Self-Esteem Scale from adolescence to old age. *Frontiers in Public Health, 8*: 132. https://doi.org/10.3389/fpubh.2020.00132

Olson J. F., & Fremer J. (2013). *TILSA Test security guidebook: Preventing, detecting, and investigating test security irregularities*. Council of Chief State School Officers.

Panayides, P., & Tymms, P. (2012). Is aberrant response behavior a stable characteristic of students in classroom math tests? *Rasch Measurement Transactions, 26*(3), 1382-1383.

Panayides, P., & Tymms, P. (2013). Investigating whether aberrant response behaviour in classroom maths tests is a stable characteristic of students. *Assessment in Education: Principles, Policy & Practice*, *20*(3), 349-368. https://doi.org/10.1080/0969594x.2012.723610

Pina, J. A. L., & Montesinos, M. D. H. (2005). Fitting Rasch model using appropriateness measure statistics. *The Spanish Journal of Psychology, 8*(1), 100-110. https://doi.org/10.1017/S113874160000500X

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Reise, S. P., & Flannery, W. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9*(1), 9–26. https://doi.org/10.1207/s15324818ame0901_3

Rolstad, S., Adler, J., & Ryden, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health, 14*(8), 1101-1108. https://doi.org/10.1016/j.jval.2011.06.003

Rosenberg, M. (1965). *Rosenberg Self-Esteem Scale (RSES)* [Database record]. APA PsycTests. https://doi.org/10.1037/t01038-00 0

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*(2), 191–207. https://doi.org/10.1007/BF02294835

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, *46*(2), 359–372. https://doi.org/10.1177/001316448604600210

Spoden, C., Fleischer, J., & Frey, A. (2020). Person misfit, test anxiety, and test-taking motivation in a large-scale mathematics proficiency test for self-evaluation. *Studies in Educational Evaluation, 67:* 100910. https://doi.org/10.1016/j.stueduc.2020.100910

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3-46. https://doi.org/10.1177/1094428114553062

Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2024a). Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. *Disability and Rehabilitation, 46*(3), 591–603. https://doi.org/10.1080/09638288.2023.2169771

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024b). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation, 46*(3), 604–617. https://doi.org/10.1080/09638288.2023.2169772

Turner, K. T., & Engelhard, G., Jr. (2024). Using functional clustering to diagnose person misfit. *Journal of Experimental Education, 92*(2), 377–397. https://doi.org/10.1080/00220973.2022.2161088

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*(2), 251–265. https://doi.org/10.1007/BF02294800

Wanders, R. B. K., Meijer, R. R., Ruhé, H. G., Sytema, S., Wardenaar, K. J., & de Jonge, P. (2018). Person-fit feedback on inconsistent symptom reports in clinical depression care. *Psychological Medicine, 48*(11), 1844-1852. https://doi.org/10.1017/S003329171700335X

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*(1), 116–136. https://doi.org/10.1037/1082-989X.9.1.116

Wind, A. S., & Schumacker, E. R. (2017). Detecting measurement disturbances in rater mediated assessments. *Educational Measurement: Issues and Practice*, 36(4), 44-51. https://doi.org/10.1111/emip.12164

Wright, B. D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wide Range, Inc.

Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference, 82*(1-2), 171-196. https://doi.org/10.1016/S0378-3758(99)00041-5

Zahra, N. S., & Wirawan, H. (2024). Empowering digital transformation: Developing and validating a Digital Leadership Scale through Rasch model analysis. *Measurement: Interdisciplinary Research and Perspectives*. Advanced online publication. https://doi.org/10.1080/15366367.2024.2334591

Zou, D., & Bolt, D. M. (2023). Person misfit and person reliability in Rating Scale Measures: The role of response styles. *Measurement: Interdisciplinary Research and Perspectives, 21*(3), 167-180. https://doi.org/10.1080/15366367.2022.2114243