2016 ■ Vol. 12 ■ no. 3

Person fit assessment using the PerFit package in R

Amin Mousavi[a,✉], Jorge N. Tendeiro[b] & Jalil Younesi[c]

[a]University of Saskatchewan, Canada
[b]University of Groningen, The Netherlands
[c]Allameh Tabataba'i University, Iran

Abstract ■ The validity of scores derived from an educational or psychological testing situation determines the accuracy and appropriateness of inferences made about an examinee based on his/her test score. Person fit assessment provides a framework for assessing the conformity of a test score to a given measurement model or to a group of examinees as an indicator of validity/invalidity of the test score. This paper presents a brief overview of person fit assessment, the effect of person misfit on ability estimation, the PerFit R package as a powerful tool for person fit assessment, and two practical examples on how to use PerFit for person fit assessment.

Keywords ■ Person fit, Item response theory, Aberrant responding, Test validity.

✉ amin.mousavi@usask.ca

AM: 0000-0002-6920-2319; *JNT*: na; *JY*: na

10.20982/tqmp.12.3.p232

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Introduction

In educational and psychological testing practice, the concept of validity is closely related to the accuracy and appropriateness of test scores. Invalid score inferences may be made if the measurement model fails to reflect accurately the real aspects of examinee responding processes. One of the situations that can lead to invalid score inferences is when the response pattern of an examinee does not fit the measurement model (e. g., an item response theory, IRT, model). Attempts to assess the fit of an examinee's response pattern to the measurement model have led researchers to studies of "person-fit" statistics (PFSs; see Tendeiro, Meijer, and Niessen; Tendeiro and Meijer, in press, 2014, for accessible overviews). Numerous studies have been conducted to develop PFSs aimed at evaluating the accuracy and appropriateness of scores obtained from a testing procedure (e. g., Armstrong and Shi; Donlon and Fischer; Drasgow, Levine, and Williams; Harnisch and Linn; Levine and Rubin; Meijer; Sijtsma; Sijtsma and Meijer; K. K. Tatsuoka and Tatsuoka; van der Flier, 2009, 1968, 1985, 1981, 1979, 1994, 1986, 1992, 1983, 1982). The logic of these statistics is typically to check whether an observed response pattern fits an expected response pattern derived from a testing model. For example, if an examinee produces correct answers to the more difficult items but fails to answer the easier ones successfully, the response pattern is considered to be "unexpected", "aberrant", or "misfitting".

Person-fit analyses are used in various empirical settings ranging from primary education (Meijer, Egberink, Emons, & Sijtsma, 2008) to high-stakes testing (Meijer & Tendeiro, 2014), and clinical testing (Conijn, 2013). The mathematical complexity of these statistics, combined with the lack of dedicated software, has prevented the widespread use of these techniques among practitioners. The PerFit R package (Team, 2015; Tendeiro et al., in press) will hopefully help to address this limitation.

Person-Fit Statistics

There are different types of PFSs but generally they are categorized as IRT-based (parametric) and group-based (nonparametric) indices (e.g. Sijtsma and Meijer; Meijer and Sijtsma; Karabatsos, 1992, 2001, 2003). In the group-based approach, PFSs are computed based on broad assumptions related to nonparametric IRT models (Sijtsma & Molenaar, 2002). Usually, group-based PFSs classify an observed response pattern as misfitting if too many easy items are answered incorrectly and/or too many hard items are answered correctly (Meijer & Sijtsma, 2001). Examples of group-based person-fit statistics include Harnisch and Linn's (1981) modified caution index C*, van der Flier's

Table 1 ■ Item parameters and response patterns (simulated data)

	Items									
Parameters	1	2	3	4	5	6	7	8	9	10
Discrimination	0.67	1.00	1.14	1.34	1.27	1.5	1.87	1.15	1.00	0.8
Difficulty	−2.00	−1.59	−0.85	−0.10	0.00	0.5	1.2	1.9	2.2	2.5
Guessing	0.01	0.20	0.15	0.15	0.10	0.25	0.20	0.11	0.05	0.01
Response patterns										
Examinee 1	1	1	1	1	1	0	0	0	0	0
Examinee 2	0	0	0	0	0	1	1	1	1	1

(1982) U3 index, K. K. Tatsuoka and Tatsuoka's (1983) norm conformity index NCI, and Sijtsma's (1986) HT coefficient.

In the parametric IRT-based approach, PFSs assess the fit of a response pattern relative to a given IRT model such as the three parameter logistic model (3PLM; Embretson and Reise, 2000). Model-based PFSs use estimated item and ability parameters to compute expected response probabilities, which are then compared to the observed response patterns. If, according to the IRT model at hand, the probability of a correct response from an examinee is high, the hypothesis is posited that the examinee should answer that item correctly, and vice versa. A misfit is found when the hypothesis is not supported by the observed data. Examples of IRT-based person-fit statistics include Wright and Stone's (1979) U statistic, Wright and Masters's (1982) W statistic, Drasgow et al.'s (1985) lz statistic, and Snijders's (2001) lz* statistic. Interested readers can refer to Meijer and Sijtsma (2001) for an extensive review and discussion on several PFSs.

The effect person misfit on ability estimation

A misfitting response pattern can lead to over- or underestimating an examinee's trait level regardless of the kind of educational or psychological test. The effect of the misfit on the estimation of the trait or ability level can be illustrated by its effect on the likelihood function. In IRT, estimation of the trait level measured by a test can be achieved by maximizing a likelihood function for a given model and an observed response pattern.

In the following contrived example, two different examinees (assume that both have an ability level of zero) take a ten-item test in which items are sorted in ascending difficulty level. Item parameters for the 3PLM and response patterns are shown in Table 1. The 3PLM is one of the most popular IRT models in which item difficulty (denoted as b), item discrimination (denoted as a), and item pseudo-guessing level (denoted as c) are used to estimate the probability of a correct answer to an item given the ability level θ. The 3PLM is given by:

$$P_j(\theta_i) = c_j + \frac{(1-c_j)}{1+e^{-a_j(\theta_i - b_j)}}, \tag{1}$$

where θ_i is the latent trait for examinee i and $P_j(\theta_i)$ is the probability of a correct answer to the jth item by the ith examinee.

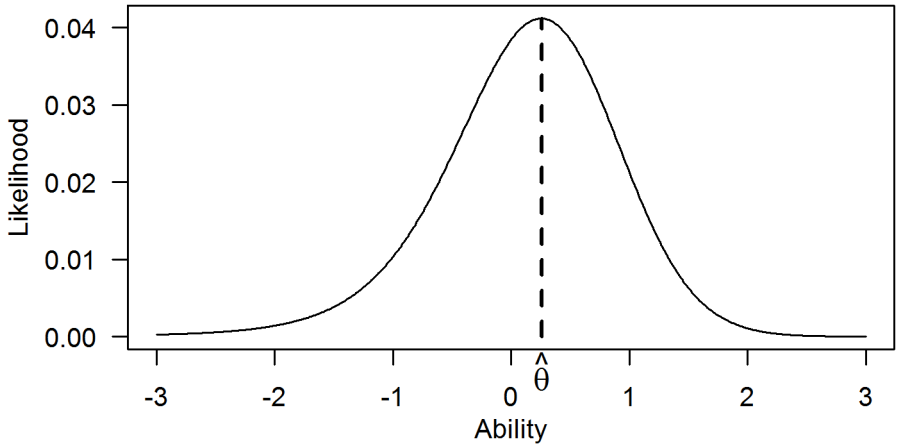
As can be seen in Table 1, Examinee 1 has answered the first five items correctly and the rest incorrectly, which means that s/he failed to answer items that have difficulty levels higher that his/her ability. It is possible to relate the item difficulty and ability parameters because both the ability and the difficulty parameters are on the same scale. Examinee 2, on the other hand, has answered the last five items correctly and the rest incorrectly, which means that s/he answered five difficult items with difficulty levels higher than his/her ability but failed to answer easy items with respect to his/her ability. The first examinee provided a fitting response pattern and the second examinee provided a misfitting response pattern. It should be noted that this is only one possible form of misfitting response pattern. For a review of different types/labels of misfitting response patterns defined in previous research, see Meijer (1996) and Rupp (2013).

The likelihood function for the ith examinee can be computed using the following formula,

$$L_i = \prod_{j=1}^{J} P_j(\theta_i)^{X_{ij}} \left(1 - P_j(\theta_i)\right)^{1-X_{ij}} \tag{2}$$

where X_{ij} is the binary $(0,1)$ response to item j ($j = 1, 2, \ldots, J$) by examinee i, with score 0 (respectively 1) indicating an incorrect (respectively correct) answer. The maximum likelihood estimate (MLE) of θ, $\hat{\theta}$ occurs at the maximum of the likelihood function, where the first derivative of the likelihood function equals zero. This MLE approach can be easily extended to polytomously scored items. Ploytomous items are essentially an extension to dichotomous or binary items in which each item has more than two response categories. One popular example of polytomously scored items are Liket-style items where respondents can choose one option out of usually 3, 5 or 7

Figure 1 ■ Likelihood function of the fitting response pattern of the first examinee



possible options.

The likelihood function for Examinee 1 (i. e. fitting response pattern) is presented in Figure 1. As showed in this figure, a fitting response pattern results in a likelihood function with a maximum on the estimated trait level, sharply dropping-off at other values of ability. The likelihood function for Examinee 2 (i. e. misfitting response pattern) is presented in Figure 2. As seen in this figure, this specific misfitting response pattern results in a likelihood function without an exact maximum (i. e., in the ability range of -3 to 3), as it is mainly flat at its peak. This makes it difficult, if not impossible, to accurately determine a relative maximum. Hence, the likelihood function does not accurately reflect the true ability due to the nature of the misfitting response pattern.

Although both examinees achieved a number-correct score of five, due to the sharp difference between the response patterns, their ability estimates are different. The estimated ability for Examinee 1 is 0.26, which is close to true ability level of 0. However, for Examinee 2, there is no obvious estimate for his/her ability but the likelihood function has higher values at large negative ability values. So, any decision for Examinee 2 based on his/her ability estimate is invalid as his/her ability estimate is not accurate and valid. This example expressed the need for assessing fit of a response pattern to a given model. Additionally, this example points at a potential cause for moderate detection power of parametric PFSs compared to non-parametric PFSs. Since parameteric PFSs (e.g., l_z^*) use estimated ability values in their computation, biased ability estimate may lead to less accurate person fit analysis for misfitting response patterns.

An overview of the PerFit package

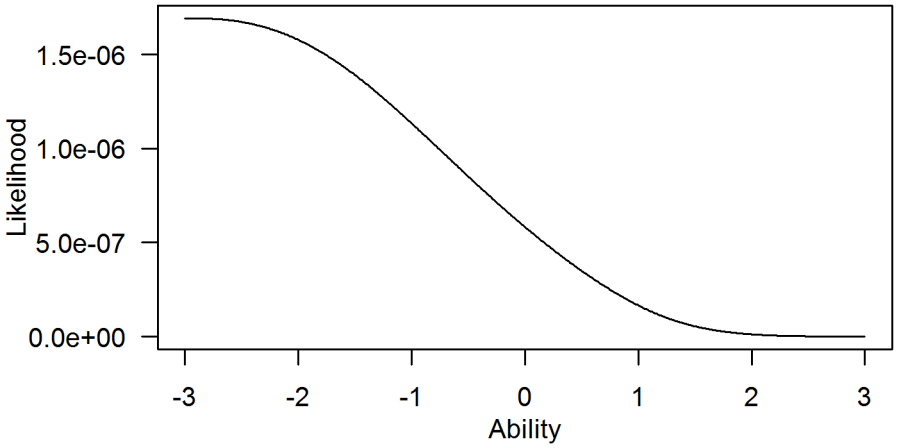
The PerFit (Team, 2015; Tendeiro et al., in press) package contains several person-fit functions. The goal is to detect response vectors that seem to be strange in terms of the sample of respondents or in terms to the IRT model. The current version (i.e., 1.4) of PerFit package includes the person-fit statistics listed in Table 2.

There are several overview papers that help comparing among the various competing PFSs (e.g., De la Torre & Deng, 2008; Karabatsos, 2003; Mousavi & Cui, 2013; Nering & Meijer, 1998; Tendeiro & Meijer, 2014). This package also contains plotting utilities for the distribution of calculated PFSs with the cutoff value superimposed, and for non-parametric person response functions (PRFs; Sijtsma and Meijer, 2001), which may also be requested in order to help interpreting individual answering behaviors. This feature is only available for dichotomous items. There are also other useful functions incorporated in the package such as cutoff for estimating `cutoff` value for a given PFS and `flagged.resp` for identifying misfitting response patterns based on the estimated cutoff value that can be used for person fit assessment.

A step-by-step example

Even thought there are not established procedures for conducting person fit assessments, a typical procedure involves: a) Choosing the PFS(s), b) depending on the choice of PFS, determining the method for estimating a cutoff value, c) calculating the PFSs for the data, and d) identifying the misfitting response patterns. As suggested by Emons, Sijtsma, and Meijer, 2005, plotting PRFs for flagged response patterns can be used for diagnosis purposes. The

Figure 2 ■ Likelihood function of the misfitting response pattern of the second examinee. Note the vertical axis range compared to Figure 1.



PRF for a fitting response pattern is expected to be non-increasing, so a PRF plot can used to inspect local increases in a misfitting response pattern.

For the purpose of illustration of the above mentioned steps, a dataset available in the `PerFit` package called "InadequacyData", comprising 28 dichotomously scored items for 806 respondents, is used. Similar to other R software packages, we can load both the PerFit package and the dataset using the following commands:

```
library(PerFit)
data(InadequacyData)
```

We use the H^T coefficient for this example but this procedure can be replaced by any of the PFSs listed in Table 2. All the PFS functions in the PerFit package can handle missing values in the dataset by either pairwise deletion or single imputation. For more information on available imputation methods or other details concerning a specific PFS it is possible to read the help for each PFS function within R (e.g., `?Ht`). In order to calculate the PFS scores based on the HT coefficient, we can use the following command:

```
Ht.out <- Ht(InadequacyData)
```

This command saves all the outputs of `Ht` function in the "Ht.out" object. Then we can estimate the cutoff value by using `cutoff` function as:

```
Ht.cut<-cutoff(Ht.out)
```

The `cutoff` function employs a bootstrap resampling procedure (default is 1000 samples) for approximating the sampling distribution of the PFS based on generated fitting response patterns. The fitting response patterns can be generated with respect to a parametric IRT model (i.e., ModelFit="Parametric" option) or nonparametric model (i.e., ModelFit = "NonParametric" option as default). By default, the cutoff value is estimated at the significance level of 0.05; this level can be changed by means of the Blvl option. Due to the nature of the bootstrap procedure, every time the cutoff function is run a (slightly) different cutoff value will be estimated. So, for the sake of consistency, it is better to save the output of the `cutoff` function in an R object (e.g., `Ht.cut`).

The next step is to use the estimated cutoff value as a decision-making rule in order to identify misfitting response patterns. This can be done using the `flagged.resp` function.

```
flagged<-flagged.resp(Ht.out,cutoff.obj=
    Ht.cut,scores=F)$PFSscores
```

The `flagged.resp` function takes the output from the PFS function and the estimated cutoff value. If the cutoff value is not provided then it will be estimated internally. We used "scores=F" to prevent showing observed response patterns for identified misfitting cases in conjunction with "$PFSscores" for a cleaner output. In our analysis 52 response patterns were identified as misfitting which is equal to 6.45% of the sample. We can use the PRFplot function for inspecting the person response function associated to the misfitting response patterns, for example case number 30, using this command:

```
PRFplot(InadequacyData,respID=30)
```

The result is shown on the left panel of Figure 3. It is also possible to plot the PRF for fitting response patterns as a

Table 2 ■ PFSs available in the PerFit package

Person-fit statistic (R function)	Reference	Item type	Type of PFS
r.pbis	(Donlon & Fischer, 1968)	Dichotomous	NonParametric
C.Sato	(Sato, 1975)	Dichotomous	NonParametric
G, Gnormed	(van der Flier, 1977; Meijer, 1994)	Dichotomous	NonParametric
A.KB, D.KB, E.KB	(Kane & Brennan, 1980)	Dichotomous	NonParametric
U3, ZU3	(van der Flier, 1980, 1982)	Dichotomous	NonParametric
Cstar	(Harnisch & Linn, 1981)	Dichotomous	NonParametric
NCI	(K. K. Tatsuoka & Tatsuoka, 1982, 1983)	Dichotomous	NonParametric
lz	(Drasgow, Levine, & Williams, 1985)	Dichotomous	Parametric
lzpoly	(Drasgow, Levine, & Williams, 1985)	Polytomous	Parametric
Ht	(Sijtsma, 1986)	Dichotomous	NonParametric
Gpoly	(Molenaar, 1991)	Polytomous	NonParametric
Gnormed.poly	(Molenaar, 1991; Emons, 2008)	Polytomous	NonParametric
lzstar	(Snijders, 2001)	Dichotomous	Parametric
U3poly	(Emons, 2008)	Polytomous	NonParametric

term of comparison. The right panel of Figure 3 illustrates the PRF for respondent 29, which was not flagged by H^T.

In the above plots, the items are sorted in ascending order of difficulty on the horizontal axis. As it can be seen in Figure 3, the PRF for a fitting response pattern (right panel) shows an expected trend where the probability of a correct answer decreases when item difficulty increases. But the PRF for the misfitting response pattern (left panel) depicts an increase in the probability of a correct answer for more difficult items, while there is a close-to-zero probability of a correct answer for easy items. This could be a case of item disclosure where the examinee knew the correct answers for difficult items (Emons et al., 2005).

Finally, we can use the `plot` function in order to generate a graphical representation of the observed distribution of PFS scores using the following command:

```
plot(Ht.out,cutoff.obj=Ht.cut)
```

We used "cutoff.obj" for the sake of consistency because if the cutoff value is not provided in the `plot` function then a cutoff value will be estimated internally. Figure 4 illustrates the observed distribution of calculated PFS scores with the cutoff value superimposed (i.e., vertical line). The colored area on the left side of Figure 4 indicates range of values that determine potentially misfitting response patterns. Additionally, the confidence interval for the estimated cutoff value is shown by a green marker on the x-axis and respondents flagged as misfitting are displayed on the top-left using red ticks. The H^T values smaller than the cutoff value (i.e., shaded area) indicate misfitting response patterns.

An empirical example

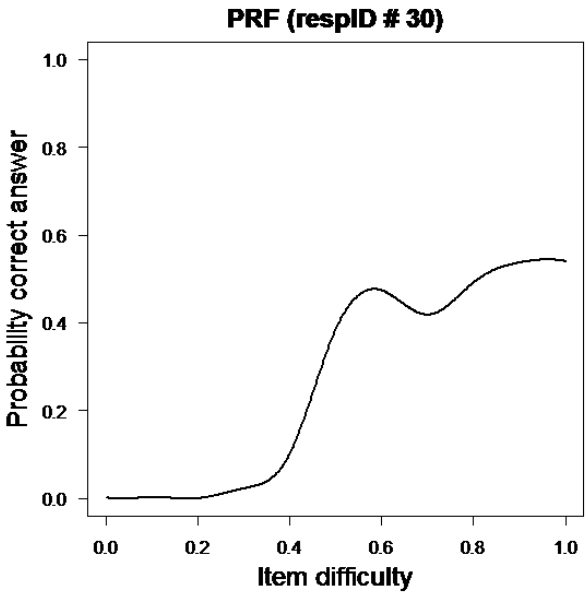
Although person fit assessment is mainly applied to achievement testing data, there are several examples of utilizing person fit indices in psychological studies (e.g., Emons, Meijer, & Denollet, 2007; Müller, Hasselbach, Loerbroks, & Amelang, 2015; Widhiarso & Sumintono, 2016). This section presents a simple example of such an application. In a psychological study on high school students in IRAN, the NEO-FFI personality test (Costa & McCrae, 1992) and Cattell's culture fair intelligence test (Cattell, Krug, & Barton, 1973) were administered to 430 individuals. The intelligence test data were further examined for person misfit using the HT coefficient. Results indicated that 40 (i.e. about 9.3%) response patterns were classified as misfitting. Figure 5 shows the distribution of H^T values. The shaded area (on the left side of the graph) represents the H^T values representing misfit.

A correlation analysis showed significant (i.e., p-value ≤ 0.05) relationships between the H^T values and Neuroticism (r= -0.32) and Openness ($r = 0.48$) sub-scales on NEO-FFI. Additionally, there was a significant relationship between the H^T values and intelligence test scores ($r = 0.54$). These are interesting findings as they can shed light on potential sources or correlates of person misfit and can provided further insights on why aberrant responses occur.

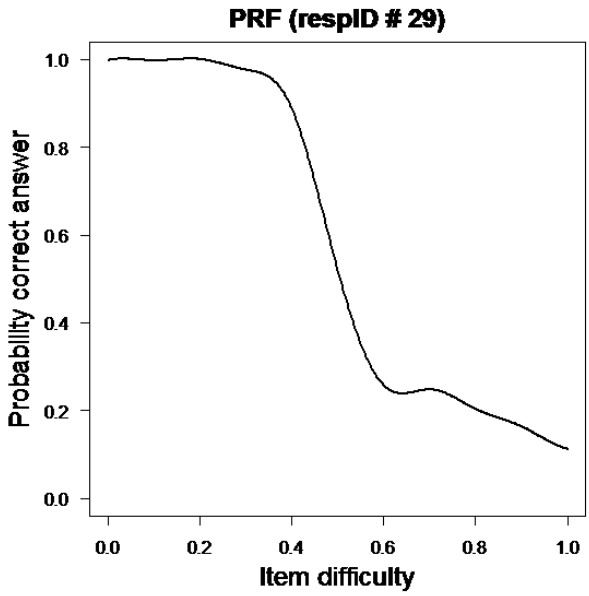
Based on these findings, persons with lower degrees of Neuroticism and higher degrees of Openness are less likely to produce misfitting response patterns. Furthermore, persons with higher IQ are also less likely to respond aberrantly. Figure 6 depicts the person response functions for selected fitting and misfitting response patterns.

Figure 3 ■ Person response function for case number 30 (misfitting; panel a) and case number 29 (fitting; panel b)

(a)



(b)



The graph for case number 2 shows a reasonable curve because an increase of the item difficulty is associated with a decrease of the probability of correct answer. On the other hand, the graph for case number 394 shows higher probability of correct answer for easy and hard items in contrast to lower probability of correct answer for items with moderate difficulty.

The profile of NEO-FFI sub-scales in Figure 7 complies with the results of the correlation analysis. The case number 394 had higher score on Neuroticism compared to the case number 2 and lower score on Openness. Moreover, the IQ score for person with misfitting response pattern (i.e., case number 394) was 84 and for the fitting response pattern (i.e., case number 2) was 109. Using these information, researchers can make decisions on how to deal with aberrant response patterns with more confidence rather than simply relying on a single score (i.e., the person-fit score).

Conclusion

In this paper, we briefly reviewed the main concepts of person fit assessment, discussed the effect of misfitting responding on the estimated ability parameter, and presented two examples on how to do such analysis in R software (Team, 2015) using the PerFit package. The person fit assessment is an active field of research which gains more attentions across the diverse spectrum of test developers, test users, and psychometric researchers. There are several research streams in this field such as evaluating the performance of existing PFSs under different conditions, identifying sources of misfitting responding (e.g., Meijer & Tendeiro, 2014; Cui & Mousavi, 2015) and applications of PFSs in testing situations (e.g., Tendeiro, Meijer, Schakel, & Maij-de Meij, 2013).

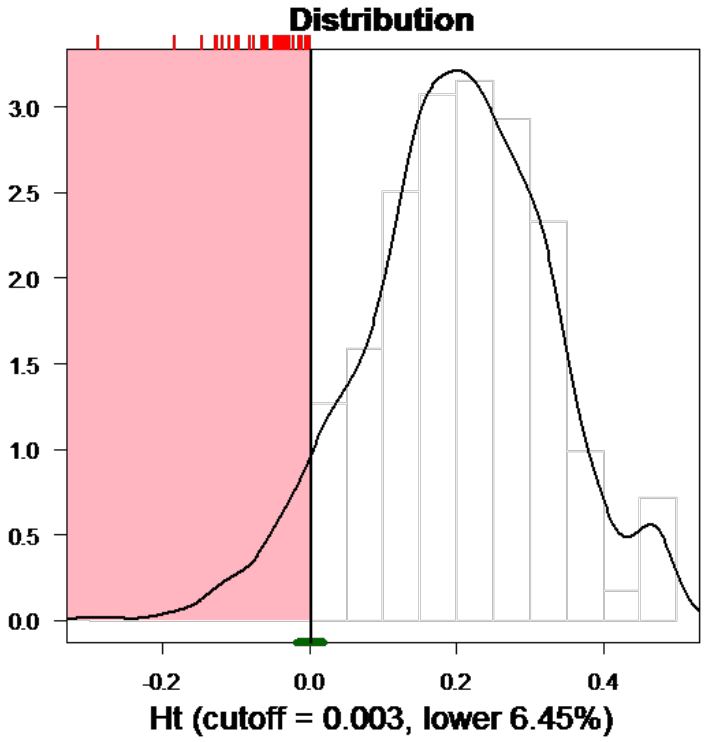
Recently, some researchers tried to develop guidelines on choosing appropriate PFSs and administrating person fit analysis (e.g., Rupp, 2013; Tendeiro & Meijer, 2014). The suggestions involve the steps we discussed in this paper in addition to follow-up quantitative and qualitative inspections for exploring potential sources of misfitting response patterns. Moreover, investigation of sources of person misfit can provide valuable information for checking the validity of scores for psychological instruments that are susceptible to fake responding.

We believe that one of the reasons that person fit assessment is still an academic research area and has not been used widely in practical testing situations was the lack of appropriate software for doing such analysis and unavailability of person fit indices in commercial and popular test calibration software. The PerFit package and this paper are steps towards making the person fit assessments more accessible.

References

Armstrong, R. D. & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psy-*

Figure 4 ■ Distribution of HT values in the sample with cutoff value.



chological Measurement, *33*, 391–410. doi:10 . 1177 / 0146621609331961

Cattell, R. B., Krug, S. E., & Barton, K. (1973). *Technical supplement for the culture fair intelligence tests, scales 2 and 3*. Champaign, IL: IPAT.

Conijn, J. M. (2013). *Detecting and explaining person misfit in non-cognitive measurement*. Netherlands: Unpublished doctoral dissertation). Tilburg University.

Costa, P. T. & McCrae, R. R. (1992). *Revised neo personality inventory (neopir) and neo fivefactor inventory (neoffi) professional manual*. FL, Psychological Assessment Resources: Odessa.

Cui, Y. & Mousavi, A. (2015). Explore the usefulness of person-fit analysis on large-scale assessment. *International Journal of Testing*, *15*, 23–49. doi:10 . 1080 / 15305058.2014.977444

De la Torre, J. & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*, 159–177. doi:10 . 1111 / j . 1745 - 3984 . 2008 . 00058.x

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

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x

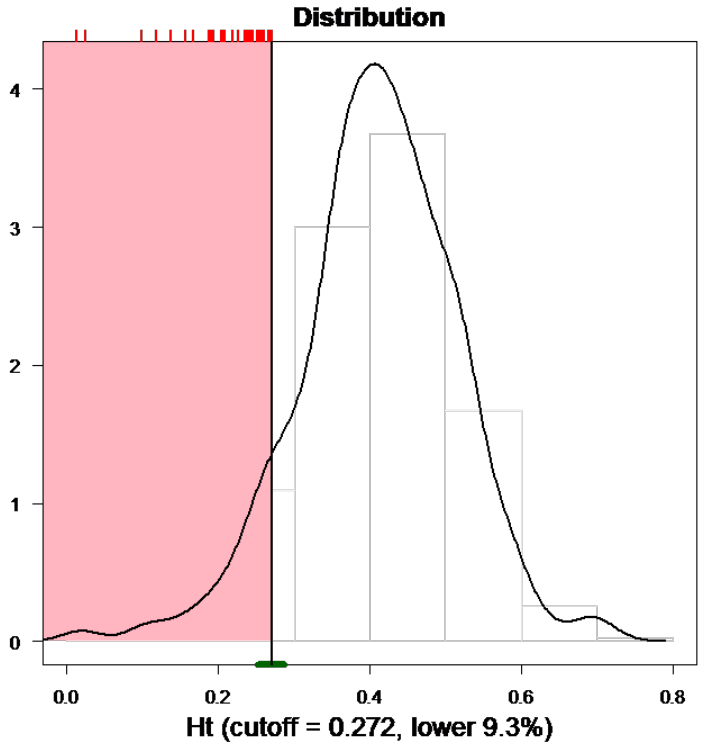
Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J: L. Erlbaum Associates.

Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224–247. doi:10 . 1177 / 0146621607302479

Emons, W. H., Meijer, R. R., & Denollet, J. (2007). Negative affectivity and social inhibition in cardiovascular disease: evaluating type-d personality and its assessment using item response theory. *Journal of psychosomatic research*, *63*(1), 27–39. doi:10.1016/j.jpsychores.2007.03.010

Emons, W. H., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-

Figure 5 ■ Distribution of H^T values in the data with cutoff value



response functions. *Psychological Methods*, *10*, 101–119. doi:10.1037/1082-989X.10.1.101

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*, 133–146. doi:10.1111/j.1745-3984.1981.tb00848.x

Kane, M. T. & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, *4*, 105–126. doi:10.1177/014662168000400111

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement In Education*, *16*, 277–298. doi:10.1207/S15324818AME1604_2

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, *4*, 269–290. doi:10.2307/1164595

Meijer, R. R. (1994). The number of guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*, 311–314. doi:10 . 1177 / 014662169401800402

Meijer, R. R. (1996). Person-fit research: an introduction. *Applied Measurement in Education*, *9*, 3–8. doi:10 . 1207/s15324818ame0901_2

Meijer, R. R., Egberink, I. J., Emons, W. H., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: an illustration with harter's self-perception profile for children. *Journal of Personality Assessment*, *90*, 227–238.

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

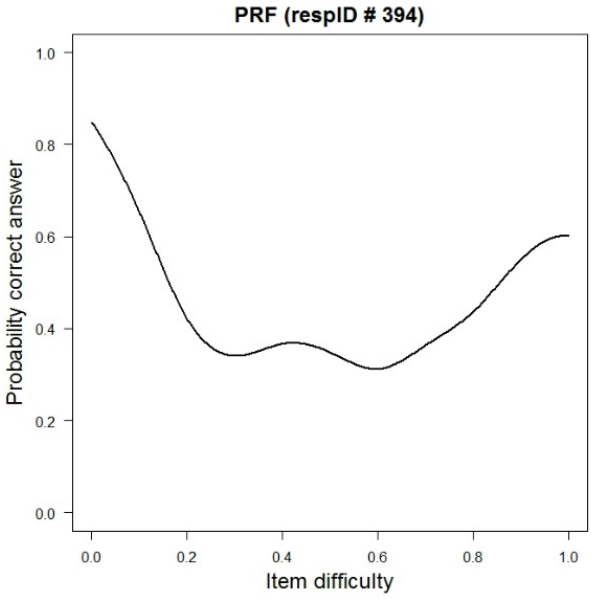
Meijer, R. R. & Tendeiro, J. N. (2014). *The use of person-fit scores in high-stakes educational testing: how to use them and what they tell us* (tech. rep. No. 14-03). Law School Admission Council, Research Report.

Molenaar, I. W. (1991). A weighted loevinger h-coefficient extending mokken scaling to multicategory items. *Kwantitatieve Methoden*, *12*, 97–117.

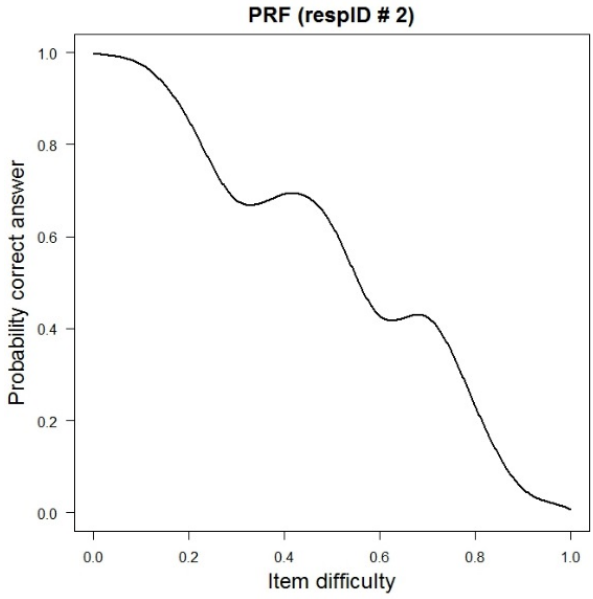
Mousavi, A. & Cui, Y. (2013). Evaluate the performance of lz and lz* of person fit: a simulation study. In *Poster presented at the graduate student research session of the annual conference of national council on measurement in education*. San Francisco. USA.

Figure 6 ■ Person response function for case number 394 (misfitting; panel a) and case number 2 (panel b).

(a)



(b)



Müller, J. M., Hasselbach, P., Loerbroks, A., & Amelang, M. (2015). Person-fit statistics, response sets and survey participation in a population-based cohort study. *Psihologija*, *48*(4), 345–360. doi:10.2298/PSI1504345M

Nering, M. L. & Meijer, R. R. (1998). A comparison of the person response function to the lz person-fit statistic. *Applied Psychological Measurement*, *22*, 53–69. doi:10.1177/01466216980221004

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*, 3–38.

Sato, T. (1975). *The construction and interpretation of s-p tables*. Tokyo: Meiji Tosho.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, *7*, 131–145.

Sijtsma, K. & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in mokken's non-irt-based irt model. *Applied Psychological Measurement*, *16*, 149–157. doi:10 . 1177 / 014662169201600204

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

Sijtsma, K. & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, Calif: SAGE Publications.

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

Tatsuoka, K. K. & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, *7*, 215–231. doi:10.2307/1164646

Tatsuoka, K. K. & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, *20*, 221–230. doi:10.1111/j.1745-3984.1983.tb00201.x

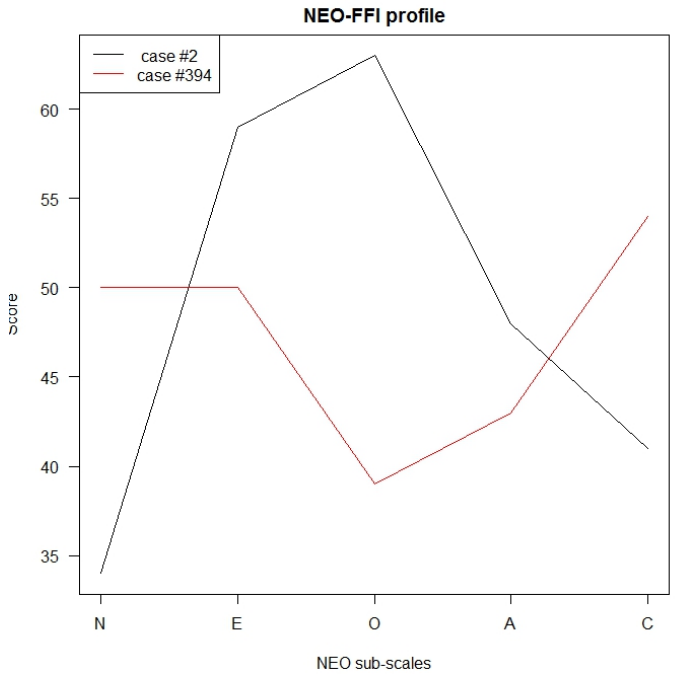
Team, R. C. (2015). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Tendeiro, J. N. & Meijer, R. R. (2014). Detection of invalid test scores: the usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51*, 239–259. doi:10.1111/jedm.12046

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (in press). Perfit: an r package for person-fit analysis in irt. *Journal of Statistical Software, in press*.

Tendeiro, J. N., Meijer, R. R., Schakel, L., & Maij-de Meij, A. M. (2013). Using cumulative sum statistics to detect inconsistencies in unproctored internet testing. *Educational and Psychological Measurement*, *73*, 143–161. doi:10.1177/0013164412444787

Figure 7 ■ Profiles of NEO-FFI scores for case #2 and case #394



van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 1–44). Amsterdam: The Netherlands.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties [comparability of individual test performance]*. Lisse: The Netherlands.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*, 267–298. doi:10 . 1177 / 0022002182013003001

Widhiarso, W. & Sumintono, B. (2016). Examining response aberrance as a cause of outliers in statistical analysis. *Personality and Individual Differences*, *98*, 11–15. doi:10.1016/j.paid.2016.03.099

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

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

Appendix

R codes used to create Figures 1 and 2:

```
As <- c(.67, 1, 1.14, 1.34, 1.27, 1.5, 1.87, 1.15, 1, .8)
Bs <- c(-2, -1.59, -.85, -.10, 0, .5, 1.2, 1.9, 2.2, 2.5)
Cs <- c(.01, .2, .15, .15, .1, .25, .2, .11, .05, .01)
Ex1 <- c(1,1,1,1,1,0,0,0,0,0)
Ex2 <- c(0,0,0,0,0,1,1,1,1,1)

Three.PLM <- function(th) {Cs + (1-Cs)/(1+exp(-(As*(th-Bs))))}
theta <- seq(-3, 3, .01)
probs <- t(sapply(theta, Three.PLM))
likelihood <- function(Ex) apply(probs,1, function(vec) {prod((vec^Ex) * (1-vec)
    ^(1-Ex))})
```

```
# Figure 1
par(mar=c(3, 4, .5, .5))
plot(theta,likelihood(Ex1),type="l",las=1, xlab="", ylab="Likelihood")
segments(.26, 0, .26, .0412, lty = 2, lwd = 2)
mtext("Ability", side = 1, line = 2, cex = 1)
mtext(expression(hat(italic(theta))), side = 1, line = .5, cex = 1.2, at = .26)

# Figure 2
par(mar=c(3, 5, .5, .5))
plot(theta, likelihood(Ex2), type="l", las=1, xlab="", ylab="")
mtext("Ability", side = 1, line = 2, cex = 1)
mtext("Likelihood", side = 2, line = 4, cex = 1)
```

Citation

Mousavi, A., Tendeiro, J. N., & Younesi, J. (2016). Person fit assessment using the Perfit package in R. *The Quantitative Methods for Psychology*, *12*(3), 232–242. doi:10.20982/tqmp.12.3.p232

Received: 1/07/2016 ~ Accepted: 13/07/2016