Using Person Fit and Person Response Functions to Examine the Validity of Person Scores in Computer Adaptive Tests

A. Adrienne Walker and George Engelhard Jr.

Treating the inferences from all test scores as if they are equally trustworthy is problematic. Test score inferences from persons who do not exhibit adequate model-data fit may not provide accurate inferences about their knowledge, skills, or abilities. Aggregate level person fit analyses provide validity evidence that can be used to inform test score interpretation and use in general. Individual person fit procedures provide further validity evidence that can be used to inform test score interpretation for specific test-takers.

In addition to providing valuable validity information, individual person fit analyses are important for testing practice because most test accountability stakes occur at the individual test-taker level. For computer adaptive tests (CAT), the information obtained from an individual person fit analysis is even more relevant. Traditional item quality-checking and aggregate person quality-checking procedures have restricted utility in CAT because each test-taker can potentially receive a different set of items. Individual person fit analysis in CAT can provide a customized quantification of how well a person's responses accord with the model used to generate his or her achievement level. With increases in the numbers of CAT being administered (Chang and Ying 2009), it is important to study procedures that can provide additional and post-test validity evidence.

The purpose of this study is to explore person fit in a computer adaptive test using a two-step procedure that incorporates statistical and graphical techniques. First, person fit statistics were used to statistically quantify misfit. Then, person response functions (PRF, Trabin and Weiss 1979) were used to graphically depict misfit. This general approach to examining fit has been explored using paper–pencil

A.A. Walker (✉)
Division of Educational Studies, Emory University,
North Decatur Building, Suite 240, Atlanta, GA 30322, USA
e-mail: angela.adrienne.walker@emory.edu

G. Engelhard Jr.
The University of Georgia, Athens, GA, USA

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_24

tests (Emons et al. 2005; Nering and Meijer 1998; Perkins et al. 2011; Ferrando 2014; Walker et al. 2016) and it seems extendable and useful for exploring person fit in CAT. The research question that guides this study is: Do person response functions in conjunction with person fit statistics have the potential to detect and inform researchers of misfit in CAT?

Background

Many methods exist for examining person fit in the context of paper-pencil tests (Karabatsos 2003; Meijer and Sijtsma 2001). By contrast, the research examining person fit in CAT is sparse (Meijer and van Krimpen-Stoop 2010; van Krimpen-Stoop and Meijer 1999, 2000). Much of the research conducted on person fit in CAT uses traditional paper–pencil person fit statistics for detecting person misfit. Researchers have reported that the sampling distributions of some traditional person fit statistics do not hold to their theoretical distributions in CAT, and this makes the interpretation of misfit difficult (Glas et al. 1998; McLeod and Lewis 1999; Nering 1997; van Krimpen-Stoop and Meijer 1999).

Some researchers such as McLeod and Lewis (1999), Meijer (2005), and van Krimpen-Stoop and Meijer (2000), have proposed and evaluated adaptive test-specific person fit statistics. The results have been mixed. Meijer (2005) reported that the detection power of an adaptive test-specific person fit statistic was higher than the detection power of other person fit methods in CAT, which included traditional person fit statistics. van Krimpen-Stoop and Meijer (2000) reported similar detection rates for their adaptive test-specific person fit statistic as was found for traditional person fit statistics in paper-pencil tests, but McLeod and Lewis (1999) reported that their adaptive test-specific person fit statistic was not powerful for detecting misfit in an adaptive test.

Some of these same researchers have promoted using a different statistical framework for conceptualizing person fit in CAT because the items on each adaptive test cover a different range of difficulty. Bradlow et al. (1998) and van Krimpen-Stoop and Meijer (Meijer and van Krimpen-Stoop 2010; van Krimpen-Stoop and Meijer 2000, 2001) introduced the cumulative sum procedure, CUSUM, for detecting person misfit in CAT. Both sets of researchers argue that the CUSUM procedure is useful for person misfit detection in CAT.

In summary, previous research suggests that procedures used to detect person misfit in CAT are not well-understood. An approach that provides researchers with more than one piece of person fit information may be needed to best understand person *mis*fit in CAT. The two-step approach in this study uses both statistical and graphical information to examine and illustrate person fit.

Theoretical Framework

In computer adaptive testing, tests are comprised of items that are individually selected to target the achievement for each test-taker by using a series of rule-based algorithms. These algorithms are heavily reliant on item response theory (IRT) to implement, and the banked item parameters are considered to be fixed and known. The Rasch model (Rasch 1960/1980) is theoretically compatible with an adaptive test procedure because person estimates from a Rasch-calibrated item bank are statistically equivalent across all customized tests, regardless of difficulty (Bond and Fox 2015).

Responses to adaptive test items are stochastic, but a person's responses should still accord with the model chosen to calibrate the items and generate the final score. In other words, adequate person fit should be observed. Person response functions show the relationship between the person's probability of giving the correct response and the difficulty of the items to which she or he responds (Trabin and Weiss 1979). Rasch-based PRF provide a clear graphical illustration of what good person fit looks like. By visually comparing this theoretical (Rasch) function with an empirical function, which is created from the person's observed responses, evidence of misfit can be seen.

According to the Rasch model, the probability of a person correctly answering a dichotomously scored item (where 1 denotes a correct response and 0 denotes an incorrect response) is

$$\Pr(X_i = 1|\theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad (1)$$

In the model, θ_n represents the achievement level of person n and δ_i represents the difficulty level of item i.

Method

In operational testing situations, some person misfit is expected to exist, but the amount and type of person misfit is unknown. For the exploratory analysis planned in this study, the amount of misfit needed to be controlled, but the type of misfit did not. So, to produce a realistic simulated scenario, adaptive test data simulated to fit the Rasch model were generated. It was expected that some of the simulated person responses would misfit the model by chance. We classified each simulated person as fitting or misfitting the model using three person fit statistics: Outfit MSE (Wright and Stone 1979), Infit MSE (Wright and Masters 1982), and Between fit MSE, Bfit-P (Smith 1985). These will be described later. Then, we created and visually examined person response functions of two groups of test-takers: test-takers whose responses fit the model and test-takers whose responses did not fit the model.

Data Generation

To simulate an adaptive test, five hundred items were generated to represent a unidimensional item bank calibrated with the Rasch model using the catR package (Magis and Raiche 2012) for the R platform. These items were uniformly distributed over the logit range of −5.00 to 5.00. Next, using the same package (and the 500 items), dichotomous item responses were generated for 5000 test-takers drawn from a standard normal distribution, $\theta \sim N(0, 1)$. To simulate a dichotomous item response, a random number from a uniform distribution, $U(0, 1)$ was compared to the probability of giving the correct response computed from the known θ and δ, and the Rasch model (Eq. 1). When the probability of giving the correct response was greater or equal to the random number, the dichotomous response was set to 1; otherwise, it was set to 0 (Harwell et al. 1996).

Achievement level estimation was calculated with maximum likelihood procedures. A new provisional achievement estimate was computed after every response starting after three randomly selected items with approximate difficulties of −2.00, 0.00, and 2.00 were administered. The next item (i.e., item 4 and beyond) was selected based on the item's proximity to the current provisional estimate of achievement. This item selection process is the same as maximum information selection (Thissen and Mislevy 2000) when the Rasch model is used (Magis and Raiche 2012). No content coverage or item exposure constraints were placed on the item selection, and the test was stopped after 40 items were administered. To check that the maximum likelihood achievement estimator did not severely bias the results, we also performed the adaptive test process using the weighted maximum likelihood estimator (Warm 1989), which has been shown to correct for estimation bias when the number of items is small.

Data Analysis

The analyses for examining person fit were conducted using the final achievement estimates ($\hat{\theta}$) yielded from the adaptive test procedure, the known item difficulty values (δ) and the dichotomously scored item responses. First, the three person fit statistics designed to detect misfit to the Rasch model were computed: Outfit, Infit, and Bfit-P. The Outfit and Infit person fit statistics are useful for detecting random disturbances to the model, such as what may be produced by random guessing behavior or careless responding behavior (Smith and Plackner 2010). The Bfit-P person fit statistic is good for detecting more systematic disturbances, such as what may be produced by persons who run out of time, are sub-experts or have a deficiency in a content domain, or are slow to warm-up to the test (Smith and Plackner 2010). The formulations of these fit statistics are included below:

$$\text{Outfit } MSE_n = \frac{1}{L}\sum_{i=1}^{L}\frac{(X_{ni} - E_{ni})^2}{V_{ni}} \tag{2}$$

$$\text{Infit } MSE_n = \frac{\sum_{i=1}^{L}(X_{ni} - E_{ni})^2}{\sum_{i=1}^{L} V_{ni}} \tag{3}$$

$$\text{Bfit-P } MSE_n = \frac{1}{(J-1)}\sum_{j=1}^{J}\frac{\left(\sum_{i\in j}^{Lj} X_{ni} - \sum_{i\in j}^{Lj} E_{ni}\right)^2}{\sum_{i\in j}^{Lj} V_{ni}} \tag{4}$$

In these formulations, X_{ni} is the observed response for person n on item i (either 0 or 1), E_{ni} is the expected response for person n on item i based on the estimated achievement level (i.e., a probability calculated from the model), V_{ni} is the variance, i.e., $E_{ni}(1 - E_{ni})$, and L is the number of items on the test (Smith 1985). For the Bfit-P statistic, J is the number of item subsets and L_j is the number of items in each subset (Smith 1985).

The values for the three fit statistics can range from 0.00 to ∞ and are assumed to approximate a chi-squared distribution (Wright and Stone 1979; Smith 1991; Smith and Hedges 1982). The expected values of Outfit, Infit, and Bfit-P are 1.00 when the data fit the Rasch model. For Infit and Outfit, two types of response patterns are discordant with the Rasch model: muted and noisy response patterns. Fit statistics greater than 1.00 signify *noisy* fit. Noisy response patterns indicate that the response data are too unruly to be governed by the model. Substantively, this may indicate random responding or person dimensionality. Fit statistics less than 1.00 signify *muted* fit, which may suggest dependency in the data. Substantively, this may indicate item exposure (cheating) or very slow, methodical responding. In most person fit research, the concern is with identifying noisy response patterns (Reise and Due 1991), or those patterns with values substantially higher than 1.00. This concern was the focus of this study as well. Thus, the term *misfitting* referred to extreme person fit values located in the upper tail of the person fit statistic distribution.

In this study, the Bfit-P statistic was computed based on item administration order. In the context of this study, a large Bfit-P value would suggest that the person's performance on the first, middle, and/or last subsets of items do not accord with the model, rather than suggesting general misfit over the entire response pattern. The three subsets of items were as follows: items 4–15 (n = 12), items 16–27 (n = 12), and items 28–40 (n = 13). The first three items were omitted from the Bfit-P analysis because they were used to start the computer adaptive test.

Statistical Person Fit Analysis

Like in real testing situations, the responses that truly misfit the Rasch model in this study were unknown. A method to classify each person as fitting or misfitting the model was needed. This is done by identifying a critical person fit value and using it to classify persons as fitting or misfitting. One procedure that has been used and recommended in the recent literature uses a pre-set Type I error rate (i.e., α) and derives the critical value by simulating multiple sets of test data and taking the average fit statistic value at the pre-set point on the distribution (van Krimpen-Stoop and Meijer 2000; Lamprianou 2013; Petridou and Williams 2007). In this study, the process for establishing the threshold values for categorizing misfit followed a similar process to that reported in van Krimpen-Stoop and Meijer (2000), and used five replicated computer adaptive data sets with 10,000 persons each.

Graphical Person Fit Analysis

Fourteen persons who did not fit the Rasch model according to their fit statistics were chosen to illustrate person response functions. These persons were chosen because (a) they represented a range of estimated achievement levels, (b) they had only one of the three statistics flagged, and (c) they had a large fit statistic. The rationale for this last decision was because the response vectors in this study were simulated to fit the Rasch model, so by choosing the persons that produced large fit values, we were selecting those persons who were most likely to truly misfit. For comparison, 11 persons whose response patterns *fit* the Rasch model were also chosen. They were selected to represent a range of achievement levels.

Expected and empirical person response functions were created for these 25 test-takers. The expected PRF were created by plotting the expected probabilities, which were computed by inserting the final achievement estimate into Eq. 1 as θ_n and the item difficulty parameters as δ_i. The empirical PRF were created by smoothing the test-taker's original dichotomous responses, using an iterative Hanning procedure (Velleman and Hoaglin 1981). Specifically, the dichotomous responses to the items (y_i, first ordered by item difficulty) were transformed to continuous values between 0.00 and 1.00, s_i, by using the formula:

$$s_i = \left(y_{i-1} + 2y_i + y_{i+1}\right)/4. \tag{5}$$

In this formula, the first and last responses (i.e., y_1 and y_n) are left as-is. For the responses to items y_2 through y_{n-1} (items 2 through 39 in this study), s_i replaces the observed responses, y_i. The response y_i receives a weight of two and the two responses adjacent to y_i on each side receive a weight of 1.00.

The goal of the Hanning procedure was to obtain an adequate smooth, so that the final response function was useful, while still preserving the original response

pattern. To achieve this goal with dichotomous data, the procedure was repeated a number of times. Following Engelhard (2013), we iterated the procedure one time for each raw score point. That is, a person who obtained a raw score of 20 out of 40 had his or her responses smoothed 20 times.

The final smoothed values represented the *empirical* function, and were plotted on the same coordinate space as the expected function. Thus, for each person (N = 25), two PRF were created and superimposed on top of each other. The visual inspection of the PRF focused on the within-person congruence between the expected response function and the empirical response function.

Results

Before the person fit analyses commenced, the estimated achievement levels from the adaptive test procedure were evaluated for accuracy. Specifically, two sets of simulated results were evaluated: one that used maximum likelihood estimation and one that used weighted maximum likelihood estimation. The difference between the estimated and true achievement levels using these two methods were negligible, with the mean bias and mean absolute bias being discrepant at the thousandth decimal place or smaller. The achievement levels obtained from the maximum likelihood procedure were retained and used for the subsequent person fit analyses.

Statistical Person Fit Analysis

The threshold values for the fit statistics were established via simulation using a pre-set Type I error rate of 0.05. The resulting threshold values were Outfit = 1.166, Infit = 1.050, and Bfit-P = 2.997. Using these values, each test-taker was categorized as either fitting or misfitting the model three times, once for each person fit statistic. Person fit values greater than the threshold were defined as misfitting and person fit values less than the threshold were defined as fitting. The percentages of test-takers flagged for misfit for each fit statistic was 4.6 % for Outfit, 4.3 % for Infit, and 5.4 % for Bfit-P.

Graphical Person Fit Analysis

Person response functions for 14 misfitting test-takers and 11 fitting test-takers were evaluated. All 25 PRF were distinctive, however, similar characteristics were noticed. In the PRF, the x-axis represents the item difficulty continuum. The y-axis represents the probability of giving the correct response to the items, $[\Pr(x = 1)]$. The dichotomously scored items, i.e., the raw scored responses, are shown by the

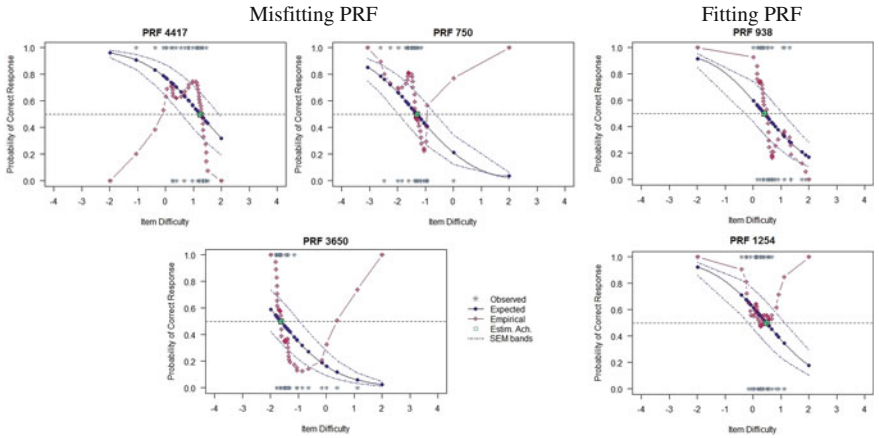

Fig. 1 Three person response functions illustrating misfit by Outfit. Two person response functions illustrating adequate fit are included for comparison. Note: The reference line included at Pr(x = 1) = 0.50 is where the item difficulty and person achievement level are equal. The *square* indicates the achievement estimate, $\hat{\theta}$, for these persons

asterisks, where a response of 1 means the test-taker gave the correct response to the item and a response of 0 means the test-taker gave the incorrect response to the item. The empirical response function created by the Hanning procedure in Eq. 5 is shown by the diamond line. The Rasch expected response function is shown by the circle line.

To aid in the interpretation of person misfit, three additional elements are included in the person response function plots. The estimated achievement level for the person is denoted by a square. The dotted lines on either side of the Rasch-expected function represent the SEM bands. These bands are calculated by plotting the Rasch probabilities for the estimated achievement level plus and minus two standard errors of measurement, i.e., $\Pr(x = 1|\hat{\theta} \pm 2 * \text{SEM})$. Finally, a horizontal reference line is drawn where the probability of the simulated person giving the correct response is 0.50, which is the location of the achievement estimate in the Rasch model.

Figure 1 highlights characteristics of misfit in the computer adaptive test as detected by Outfit. PRF 938 and 1254 illustrate adequate model fit and PRF 4417, 3650, and 750 illustrate Outfit misfit. One common observation for the persons flagged as misfitting in Fig. 1 is the large unexpected correct (or incorrect) response at the end (or beginning) of the response pattern. For instance, for Person 4417, the diamond point located at item difficulty −2.00 illustrates that this person gave an incorrect answer to this item and the circle point at the same location illustrates that the model expected the person to give a correct answer.

A second common observation for the persons flagged as misfitting by Outfit is the unexpected responses in the middle of the functions. There are some peaks and dips in the empirical response function located in the middle of the item difficulty

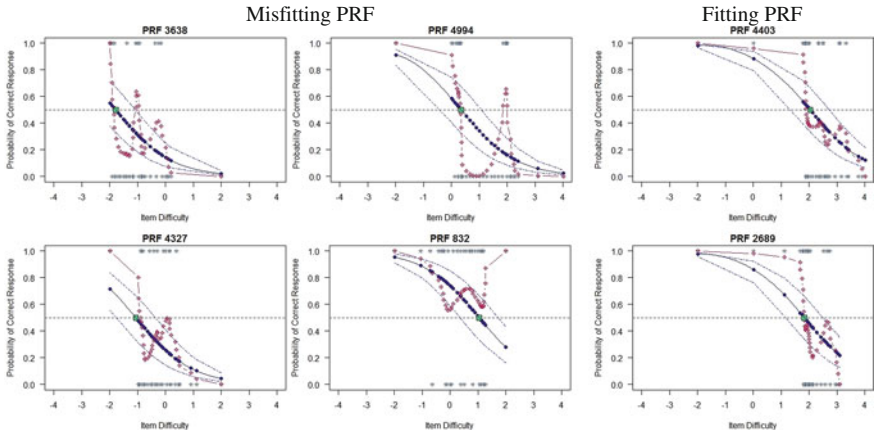

Fig. 2 Four person response functions illustrating misfit by Infit. Two person response functions illustrating adequate fit are included for comparison. Note: The reference line included at Pr (x = 1) = 0.50 is where the item difficulty and person achievement level are equal. The *square* indicates the achievement estimate, $\hat{\theta}$, for these persons

continuum, near the estimated achievement level, and a segment of the empirical function extends beyond the SEM bands. These visual characteristics of the PRF are consistent with judgments of person misfit, as suggested by the person fit statistics.

It is noted that the empirical function for Person 938, who was categorized as fitting the model, also extends beyond the SEM bands around the achievement estimate, and the empirical function for Person 1254 exhibits an unexpected response at the end of the responses. The difference is that for these fitting persons, *either* peaks and dips *or* an extreme unexpected response is present in the empirical function, but not both. For the misfitting persons, severe dips and peaks of the empirical function are seen *in addition to* an unexpected response at the end (or beginning).

Figure 2 highlights characteristics of misfit in the computer adaptive test as detected by Infit. PRF 4403 and 2689 illustrate fit to the model. PRF 3638, 4327, 4994, and 832 illustrate misfit to the model. Here, the common observation for the persons flagged as misfitting is the extreme discrepancies between the expected and empirical functions around the probability of 0.50 or the estimated achievement level. For instance, large dips and peaks are displayed for Persons 3638, 4327, and 4994 instead of a steady decline in the probabilities. A substantial portion of the empirical functions for these misfitting persons extend beyond the SEM bands. By contrast, the empirical function for Persons 4403 and 2689 (who fit the model) exhibit smaller dips and peaks, which barely fall beyond the SEM bands. For Person 832, the empirical function does not approach the probability of 0.50. These visual characteristics of the PRF are consistent with judgments of person misfit and suggest that more than one achievement estimate is plausible.

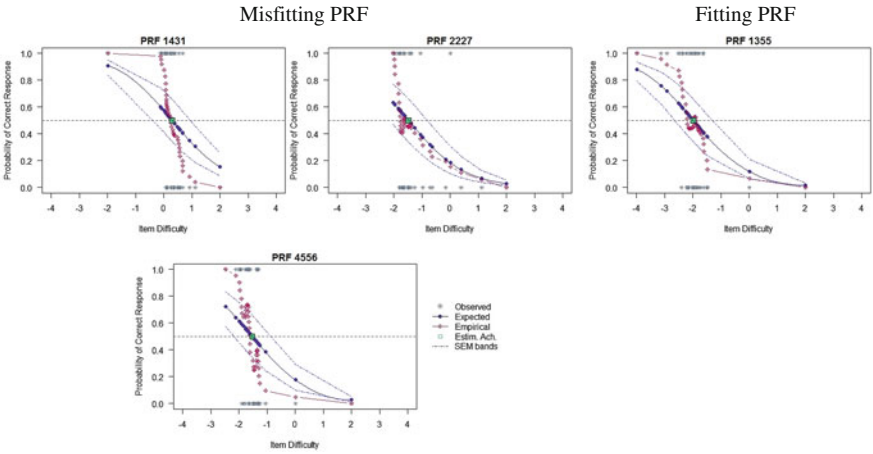

Fig. 3 Three person response functions illustrating misfit by Bfit-P. One person response function illustrating adequate fit is included for comparison. Note: The reference line included at Pr (x = 1) = 0.50 is where the item difficulty and person achievement level are equal. The *square* indicates the achievement estimate, $\hat{\theta}$, for these persons

Figure 3 highlights characteristics of misfit in the computer adaptive test as detected by Bfit-P. Person response function 1355 illustrates fit to the model and PRF 1431, 4556, and 2227 illustrate misfit to the model. Here, there is no obvious or consistent characteristic that shows misfit. For Persons 1431 and 4556, the empirical functions appear steep, but they do not show major unexpected deviations (i.e., dips and peaks) from the expected PRF. For Person 2227, the empirical function follows the expected function well. The characteristics of these PRF are not consistent with judgments of person misfit, as suggested by the person fit statistics.

Discussion and Conclusion

In this study, the simulated persons fit the model generally well, but there was some evidence of individual person misfit. This observation highlights the need for conducting individual person fit analyses in practice. Although achievement estimates from IRT models have been shown to be fairly robust to model-data misfit in paper-pencil tests (Adams and Wright 1994; Sinharay and Haberman 2014) and CAT (Glas et al. 1998), test-takers in real situations may respond to items in unique and unstudied ways that may threaten the inferences of their scores. Moreover, in CAT where the item parameters are considered known, and where each test-taker may receive a different set of items, evaluating individual person fit can provide

vital information about model-data fit that may be absent from pre-test and post-test quality checks.

The advantage of using a two-step, statistical and graphical, procedure for examining individual person fit in CAT is that it allows for a statistical quantification about individual person fit (i.e., person fit statistic) to be further informed by a visual inspection of the actual response pattern (e.g., empirical person response function). Given the skepticism regarding the use of existing person fit statistics in an adaptive test context (Glas et al. 1998; McLeod and Lewis 1999; Nering 1997; van Krimpen-Stoop and Meijer 1999), this additional information is warranted.

Because they enhance validity evidence for score interpretation and use (APA/AERA/NCME 2014), individual person fit techniques can improve computer adaptive testing practice. Person response functions require experience and some subjective judgement to interpret, but they provide complementary information about person fit. These visual representations of misfitting patterns may help researchers and other educational stakeholders understand the substantive implications of person misfit. For instance, in this study, the PRF showed *why* misfit (with the Outfit and Infit statistics) was detected and the general location. For the Bfit-P statistic, the PRF showed no substantive misfit, which suggests that more information about these person responses should be gathered before a judgement about person fit is made. Practitioners and researchers can use these two pieces of information to evaluate potential threats to a person's test score inference.

References

Adams, R. J., & Wright, B. D. (1994). When does misfit make a difference? In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 244–270). Norwood, NJ: Ablex.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*(443), 910–919.

Chang, H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics, 37*, 1466–1488.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*(1), 101–119.

Engelhard, G, Jr. (2013). Hanning (Smoothing) of person response functions. *Rasch Measurement Transactions, 26*(4), 1392–1393.

Ferrando, P. J. (2014). A general approach for assessing person fit and person reliability in typical-response measurement. *Applied Psychological Measurement, 38*, 166–183.

Glas, C. A., Meijer, R. R., & van Krimpen-Stoop, E. M. (1998). *Statistical tests for person misfit in computer adaptive testing (Research Report 98–01)*. The Netherlands: University of Twente, Faculty of Educational Science and Technology.

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101–125. doi:10.1177/014662169602000201.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person fit statistics. *Applied Measurement in Education, 16*, 227–298.

Lamprianou, I. (2013). The tendency of individuals to respond to high-stakes tests in idiosyncratic ways. *Journal of Applied Measurement, 14*(3), 299–317.

Magis, D., & Raiche, G. (2012). Random generation of response patterns under computer adaptive testing with the R Package catR. *Journal of Statistical Software, 48*(8), 1–31.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 43*, 147–160.

Meijer, R. R. (2005). *Robustness of person-fit decisions in computerized adaptive testing (Computerized Testing Report 04–06)*. Newtown, PA: Law School Admission Council.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.

Meijer, R. R., & van Krimpen-Stoop, E. M. (2010). Detecting person misfit in adaptive testing. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 315–329). New York, NY: Springer.

Nering, M. L. (1997). Distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 127–155.

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

Perkins, A., Quaynor, L., & Engelhard, G. (2011). The influences of home language, gender, and social class on mathematics literacy in France, Germany, Hong Kong, and the United States. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 4*, 35–58.

Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using multilevel models. *Journal of Educational Measurement, 44*(3), 227–247.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research, Expanded edition, Chicago: University of Chicago Press (Original work published 1960).

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217–226.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*, 23–35.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*, 433–444.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*, 541–565.

Smith, R. M., & Hedges, L. V. (1982). Comparison of likelihood ratio $\chi^2$and Pearsonian χ^2 tests of fit in the Rasch model. *Education Research and Perspectives, 9*, 44–54.

Smith, R. M., & Plackner, C. (2010). The family approach to assessing fit in Rasch measurement. In M. Garner, G. Engelhard Jr., W. Fisher Jr., & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 1, pp. 64–85). Maple Grove, MN: JAM Press.

Trabin, T. E., & Weiss, D. J. (1979). *The person response curve: Fit of individuals to item characteristic curve models*. (Research Report 79–7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 101–133). Hillsdale, NJ: Erlbaum.

van Krimpen-Stoop, E. M., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327–345.

van Krimpen-Stoop, E. M., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201–219). Dordrecht, The Netherlands: Kluwer Academic Publishers.

van Krimpen-Stoop, E. M., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199–217.

Velleman, P. F., & Hoaglin, D. C. (1981). Smoothing data. In P. Velleman & D. Hoaglin (Eds.), *Applications, basics, and computing of exploratory data analysis* (pp. 159–199). Boston, MA: Duxbury Press.

Walker, A. A., Engelhard, G. Jr., Royal, K. D., & Hedgpeth, M. W. (2016). Exploring aberrant responses using person fit and person response functions. *Journal of Applied Measurement*, *17*(2), 194–208.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

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

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