

Detection Power of Multilevel Latent-Trait Differential Person Functioning: A Monte Carlo Comparison with Conventional Person Misfit Statistics

Lihshing Wang,¹ Wei Pan,¹ and Haiyan Bai²

(1) University of Cincinnati, 51 Corry Boulevard, Cincinnati, OH 45221-0049, U.S.A.

(2) University of Central Florida, 4000 Central Florida Boulevard, Orlando, FL 32816, U.S.A.

Paper presented at the 2008 International Meeting of Psychometric Society, Durham, New Hampshire.

Abstract

Research on aberrant response patterns that deviate from modeled prediction has received considerable attention in the psychometric literature. A recent line of inquiry uses the multilevel modeling approach for the detection of differential person functioning in latent trait models in a hierarchical data structure. Preliminary Monte Carlo studies comparing the performance of this innovative approach with conventional person misfit statistics have yielded encouraging results. However, systematic evidence is lacking in demonstrating its power performance and robustness across misfit conditions. The present study extends previous research by simulating three different types of person misfit: spuriously high, spuriously low, and random guessing, with gender bias as the group-level effect. The proposed approach along with three other conventional indices was then applied to the simulated data to compare their relative performance in recovering those misfitting simulees and detecting gender bias. The entire experiment was replicated by Bootstrap sampling to evaluate its stability performance. Preliminary results suggest that power performance of the proposed method is either superior or comparable to that of the conventional indices. Suggestions for future research to manipulate potential confounding or moderating factors are discussed.

Introduction

As latent trait models continue to populate the mainstream psychometric literature, a major line of research centers on fitting the postulated model to the observed data (e.g., Smith, 2000; Stone & Zhang, 2003). Because model-data misfit may stem from either item misfit or person misfit, or both, identifying the source of inconsistency between the posited model and the observed data is crucial for maximizing the utility of latent trait methodology (Molenaar & Hoijtink, 1996; Rudas & Zwick, 1997).

The existing literature on model-data fit focuses primarily on item misfit, commonly known as *item bias* or *differential item functioning* (DIF) (Camilli & Shepard, 1994; Holland & Wainer, 1993). In recent years, however, person misfit research has received considerable attention in major measurement journals (e.g., Karabatsos, 2003; Meijer & Sijtsma, 2001; Petridou & Williams, 2007). According to a methodology review by Meijer and Sijtsma (2001), about forty statistical indices have been proposed to date, of which thirty-six indices were systematically compared

by Karabatsos (2003) on their performance in detecting different types of deviant patterns.

Although a wealth of person misfit research has accumulated over the past two decades, conventional person misfit indices suffer from one or more of the following limitations: (a) They assume that the item response parameters are fixed and known when estimating person misfit. (b) Traditional ordinary-least-squares or maximum-likelihood person misfit indices tend to be very unstable when test length is short or when a different set of items is administered to different subjects (Reise, 2000). (c) The sampling distribution for the null model of the person misfit statistic is often unknown or deviant from the nominal distribution, and the uncertainty of the item and person parameters is unaccounted for (Glas & Meijer, 2003). (d) They attempt to assess the fit of a latent trait model at the individual examinee level. (e) Comparison across groups is typically performed without due regard for the multilevel structure underlying the data.

In order to simultaneously accommodate the above limitations, Reise (2000) proposed using the multilevel logistic regression model for the detection of person misfit. With items nested within persons and persons nested within groups, the multi-group item response data matrix exhibits a hierarchical clustering of data structure that lends itself well to the multilevel modeling (MLM) methodology (Adams, Wilson, & Wu, 1997; Fox & Glas, 2001; Swanson, Clauser, Case, Nungester, & Featherman, 2002; Woods, 2008). Applying MLM to the response data, we can assess *differential person functioning* (DPF) (Johanson & Alsmadi, 2002) that manifests itself as person misfit at the individual level and group bias at the aggregate level. Specifically, the *person response slope* (PRS), which is the slope of a monotonically nonincreasing person response function that relates examinee response probability to item difficulty (Trabin & Weiss, 1983), can be modeled both at the individual level and group level for estimating the scalability of the person parameter. In effect, a PRS indicates the degree of consistency between item difficulty and response probability, given a fixed latent trait value.

In a preliminary exploration of this novel MLM approach to DPF, Wang, Reise, Pan, and Austin (2004) demonstrated its feasibility and utility with both empirical and simulated data. In another simulation study to explore the power of the PRS in a non-MLM context for detecting different types of misfit, Emons, Sijtsma, and Meijer (2004) found the detection rates to range from near zero to 99%, compared to the detection rates of largely between 45% and 95% using conventional person misfit indices (Karabatsos, 2003). This large variance in the performance of different person misfit statistics is due in part to the different misfit conditions being simulated in those studies. Clearly, unequivocal evidence is lacking in demonstrating the superiority of the MLM-DPF approach over other conventional approaches in detecting misfitting or aberrant respondents under a specific misfit condition. Without such evidential support, the significance of this novel methodology cannot be firmly established.

The purpose of this study is to evaluate the performance of multilevel modeling approach to the detection of differential person functioning under the latent trait framework. Specifically, this paper seeks to accomplish two objectives: (a) By comparing the misfit detection rate of the MLM-DPF approach with three conventional person misfit indices, this study demonstrates the effectiveness of the proposed methodology in recovering computer-simulated aberrant respondents with known and varying degrees of spuriously high scores. (b) By modeling the variation

of person misfit at the individual level in a higher-order model with group membership as an explanatory factor, this study demonstrates the utility of the MLM-DPF approach for detecting response bias due to systematic confounding effects from population group membership.

Method

Data Simulation

Sample size and test length. The simulated data matrix consisted of 6000 subjects responding to 60 dichotomous items to approximate a typical large-scale state test data. Such a large sample size eliminates the need for multiple replications typically required in a Monte Carlo simulation study (Serlin, 2000). A test length of 60 provided sufficient test reliability and satisfactory standard error of measurement in measuring a latent construct of a single dominant dimension.

Latent trait and item difficulty distribution. The MULTISIM software available from Assessment Systems (2006) was used to generate the 6000 simulees with a unit normal latent trait distribution and 60 items with a uniform item difficulty distribution with the difficulty parameter ranging from -2.0 to +2.0. This unit normal latent trait distribution approximates most real-life data, and the uniform item difficulty distribution ensures that simulees at the two extremes would result in stable parameter estimates. These ability and item parameters were then entered into the MULTILOG program and fit to the one-parameter logistic model to generate the simulated item response data matrix.

Response aberrancy. Two types of response aberrancy are commonly observed in the literature: *spuriously high* correct response rate and *spuriously low* correct response rate. A typical example of a spuriously high correct response scenario is examinees with low latent trait levels copying answers from examinees with high latent trait levels, yielding unexpectedly high correct response rates on highly difficult items. An example of a spuriously low correct response scenario is an observer/scorer error that leaves previously learned items blank in a longitudinal observation of student progress, which are then erroneously scored as incorrect along with all other missing responses, resulting in misfitting response vectors. Because spuriously low scores have been found to be more difficult to detect than spuriously high scores (Emons, Sijtsma, & Meijer, 2004, p. 32), the results from detecting this particular type of under-performing aberrancy may suffer from the floor effect of underestimating the power performance of all the indices and thus obscuring the detection differences. For the purpose of this study, therefore, the spuriously high correct response scenario was chosen to illustrate the feasibility of the proposed methodology.

Person misfit data. To simulate the person misfit data illustrating cheating, the following steps were observed: (a) Of the 6000 examinees, 2237 were first identified as low-ability examinees with latent trait levels ranging from -2.0 to -0.5. Of these 2237 low-ability examinees, 1080 examinees were then randomly selected to be simulees with response aberrancy. This accounts for 18% of the total 6000 examinees or 54% of the 2237 low-ability examinees, yielding a mean trait value of -1.24. This percentage of aberrant respondents was judged appropriate because previous research suggested that there are essentially no differences in the detection rates for misfit percentages between 5% and 25% (Karabotsas, 2003, p. 286). (b) Of the 60 items, 12 (or 20% of the 60 items) were randomly selected from the items with high difficulty levels ranging from .50 to 2.0, yielding a mean item difficulty level of 1.39. The item

responses of the 1080 cheating examinees on these 12 items were then replaced by the item responses of another group of randomly selected 1080 examinees with high latent trait levels ranging from .5 to 2.0, which are matched to the item difficulty levels of .5 to 2.0 described above. By replacing the 1080 randomly selected low-ability examinees' responses with the 1080 randomly selected high-ability examinees on those 12 randomly-selected high-difficulty items, the cheating effect was simulated to mimic the scenario of low-ability examinees copying answers from high-ability examinees.

Bootstrap replications. To test the stability of the resulting statistics, we further selected five random subsamples of 1000 examinees each from the 6000 examinees, of which 200 examinees were misfitting simulees. This sample size was judged to be reasonably manageable with optimal statistical power for detecting aberrancy (Cohen, 1998). Each 1000-by-60 item response matrix with 200-by-12 misfitted responses was entered into HLM (Raudenbush, Bryk, & Congdon, 2004) for multilevel modeling of person misfit.

Model Framework

A three-level MLM model was used to compute the PRS as an estimate for person misfit and to explore the extent to which differential person functioning can be explained by population group membership. The three-level MLM model is as follows:

Level 1 (*Item level*):

$$\text{Ln}[P_{ijk}/(1 - P_{ijk})] = \pi_{0jk} + \pi_{1jk}b_i \quad (1)$$

Level 2 (*Person level*):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}\theta_{jk} + r_{0jk} \quad (2)$$

$$\pi_{1jk} = \beta_{10k} + r_{1jk} \quad (3)$$

Level 3 (*Group level*):

$$\beta_{00k} = \gamma_{000} + u_{00k} \quad (4)$$

$$\beta_{01k} = \gamma_{010} + u_{01k} \quad (5)$$

$$\beta_{10k} = \gamma_{100} + u_{10k} \quad (6)$$

where P_{ijk} is the person response probability on the i th item for the j th person in the k th group; b_i is the item difficulty index of the i th item; π_{0jk} and π_{1jk} are the logistic regression coefficients for the i th item for the j th person in the k th group; θ_{jk} is the person latent trait for the j th person in the k th group; β_{00k} , β_{01k} , and β_{10k} are the regression coefficients in the k th group; r_{0jk} and r_{1jk} are the residuals for the j th person in the k th group; γ_{000} and γ_{010} are the intercepts; and u_{00k} and u_{01k} are the residuals in the k th group. Of particular interest is the slope of b_i (i.e., π_{1jk}), which is the PRS proposed in this study as an index of person misfit. The smaller the absolute value of the PRS, the flatter the person response curve, the more serious the person misfit. If the variance of u_{10k} (τ_{10}^2) is significantly different from zero, then it can be said that group membership explains the variation in differential person functioning well. The HLM V6.0 software (Raudenbush, Bryk, & Congdon, 2004) was used for the MLM-DPF analyses to produce the PRS person misfit index.

Conventional Person Misfit Indices

Three conventional person misfit indices were obtained by entering the simulated data to the WPerfit program (Ferrando & Lorenzo, 2000): the standardized log-likelihood index (I_z) (Drasgow, Levine, & Williams, 1985), the standardized

extended caution index (ECI_{z}) (Tatsuoka, 1984), and the chi-square statistic of discrepancy between expected and observed person response curves (χ_D^2) (Trabin & Weiss, 1983). In a systematic comparison of 36 person misfit indices, χ_D^2 was demonstrated to be second best in overall detection performance and rank first among the 25 parametric statistics with simulated data (Karabatsos, 2003).

Detection and Accuracy Rates

The *detection rate* computes the percentage of misfitting examinees being correctly identified as such (i.e., true positives). For the PRS, the cut-off value is $-.9638$, which is the one-tailed critical t -value at a significance level of $.05$ for testing the null hypothesis that the PRS is less than -1.0 in a logistic latent trait model (Lord & Novick, 1968). Thus, if an observed value of the PRS is greater than $-.9638$, the corresponding individual would be identified as person misfit to the test. The critical values for l_z and ECI_{z} are -2.00 and 2.00 respectively (Ferrando & Lorenzo, 2000). For the chi-square statistic χ_D^2 , the critical value is 16.92 with 9 degrees of freedom at a significance level of $.05$ and 10 strata. The method that produces the highest detection rate would be judged as demonstrating the highest statistical power in recovering simulees with known aberrancy. Furthermore, the *accuracy rate*, defined as the percentage of true positives plus true negatives, was also computed to provide a more comprehensive evaluation of the PRS performance.

Results and Discussion

Preliminary analyses using the proposed MLM-DPF approach showed very promising results. Figure 1 illustrates a much steeper person response curve for the non-misfitting group as was hypothesized. A flat person response curve, on the other hand, illustrates the effect of misfit.

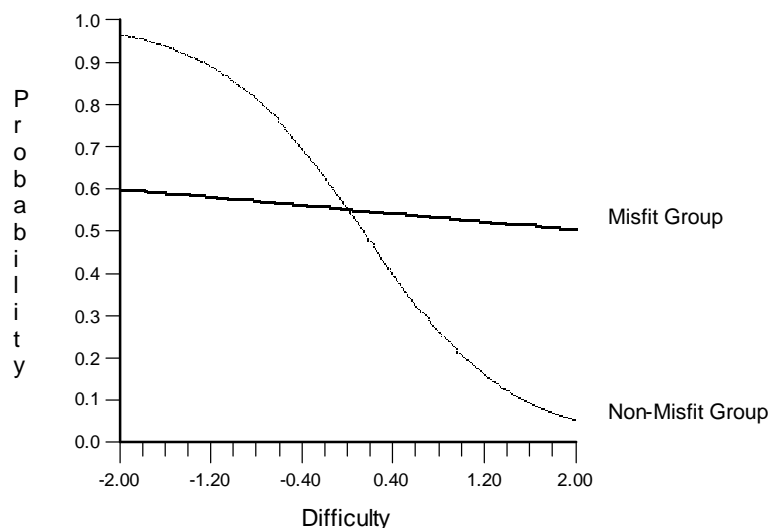


Figure 1. Person response curves for misfit ($N_1=200$) and non-misfit groups ($N_2=800$).

Correlation analysis (Table 1) showed moderate to strong mean correlations between PRS and the other three conventional person misfit indices ($|r| = .406$ to $.928$, $p < .01$). This suggests that the proposed MLM-DPF person misfit index (PRS) is largely consistent with, but not identical to, the conventional indices. Standard errors of those correlations were calculated by following the formula in Cohen, Cohen, West, and Aiken (2003).

Table 1
Mean Correlation Coefficients among Person Misfit Indices (N = 1000)

Index	Lz	ECI4z	χ^2
Lz	—		
ECI4z	-.658 (.008)	—	
χ^2	-.851 (.012)	.406 (.024)	—
PRS	-.928 (.004)	.732 (.004)	.796 (.013)

Note. Each mean correlation coefficient is based on five replications. All correlation coefficients are significant at $\alpha = .01$ level (2-tailed). The standard errors are in parentheses.

Detection analysis (Table 2) showed that PRS has a mean detection rate (98.2%, $s.e. = 0.6\%$) comparable to LZ (99.8%, $s.e. = 0.3\%$) and χ^2 (99.9%, $s.e. = 0.2\%$) but significantly better than ECI4z (52.3%, $s.e. = 2.9\%$). In addition, PRS has a mean overall accuracy rate (99.4%, $s.e. = 0.1\%$) comparable to χ^2 (99.3%, $s.e. = 0.3\%$) but significantly better than LZ (96.6%, $s.e. = 0.7\%$) and ECI4z (87.7%, $s.e. = 0.5\%$). This result suggests that PRS is comparable or superior to conventional person misfit in identifying misfitting persons, with the added advantage of detecting group bias.

Although the above findings are based on only five replications of random subsamples from the simulated data, the small standard errors (.001 to .029) suggest that the results are robust to sampling errors and are generalizable to the original sample.

Table 2
Mean Detection Rates and Overall Accuracy Rates of Person Misfit Indices (N = 1000)

Index	Detection Rate	Overall Accuracy Rate
Lz	.998 (.003)	.966 (.007)
ECI4z	.523 (.029)	.877 (.005)
χ^2	.999 (.002)	.993 (.003)
PRS	.982 (.006)	.994 (.001)

Note. Each mean rate is based on five replications. The standard errors are in parentheses.

Conclusions and Recommendations for Future Research

This exploratory study with simulated data has demonstrated that the proposed MLM-DPF approach to studying person misfit is both feasible and promising. The detection and accuracy rates of PRS are either comparable or superior to those of the

conventional indices, but with the added advantage of detecting group bias at the aggregate level.

Future research can follow several directions: (a) Explore the impact of different types of aberrancy, including spuriously low correct response rate and random responding. (b) Expand the simulation design to include systematic manipulation of factors such as misfit ratio, test length, item and trait distribution. (c) Investigate the impact of misfitting examinees on the accuracy of classification decisions (Hendrawan, Glas, & Meijer, 2005). (d) Apply the proposed methodology to real-life test data for the detection of response aberrancy and examine its impact on high-stakes decisions based on the misfitting data (Brown & Villareal, 2007; Lamprianou & Boyle, 2004; Petridou & Williams, 2007).

Exploration in these research frontiers helps build a strong advocacy for the importance of empirically checking data quality before high-stakes consequences are imposed on the individuals or institutions.

References

- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22(1), 47-76.
- Assessment Systems. (2006). *MULTISIM: Multidimensional Item Response Theory Analysis*. St. Paul, MN: Assessment Systems.
- Brown, R. S., & Villareal, J. C. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing*, 7(1), 1-25.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Dragow, 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.
- Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, 39(1), 1-35.
- Ferrando, P. J., & Lorenzo, U. (2000). WPerfit: A program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement*, 60(3), 479-487.
- Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66(2), 269-286.
- Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, 27(3), 217-233.
- Hendrawan, I., Glas, C. A. W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement*, 29(1), 26-44.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

- Johanson, G., & Alsmadi, A. (2002). Differential person functioning. *Educational and Psychological Measurement*, 62(3), 435-443.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277-298.
- Lamprianou, I. & Boyle, B. (2004). Accuracy of measurement in the context of mathematics national curriculum tests in England for ethnic minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement*, 41(3), 239-259.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107-135.
- Molenaar, I.W., & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education*, 9(1), 27-45.
- Petridou, A., & Williams, J. (2007). Accounting for aberrant test responses using multilevel models. *Journal of Educational Measurement*, 44(3), 227-247.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. T. (2004). *HLM6: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.
- Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, 35(4), 543-568.
- Rudas, T., & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics*, 22(1), 31-45.
- Serlin, R.C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods*, 5(2), 230-240.
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, 1(2), 199-218.
- Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40(4), 331-352.
- Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, 27(1), 53-75.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.
- Thissen, D., Chen, W.-H., & Bock, D. (2006). *Multilog 7 – Analysis of multiple-category response data*. St. Paul, MN: Assessment Systems.
- Trabin, T.E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item characteristic curve models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83-108). New York: Academic Press.
- Wang, L., Reise, S. P., Pan, W., & Austin, J. T. (2004, April). *Multilevel modeling approach to detection of differential person functioning in latent trait models*. Paper presented at 2004 annual meeting of American Educational Research Association, San Diego.
- Woods, C. M. (2008). Monte Carlo evaluation of two-level logistic regression for assessing person fit. *Multivariate Behavioral Research*, 43(1), 50-76.

