ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Editorial; 2024; 15(1);1-4

EDITORIAL

Adapting the Person Fit Analysis: Ideas on Detecting Person Misfit in Computerized Adaptive Testing

Beyza Aksu DÜNYA*

Highlights

- Most test accountability stakes occur at the individual level (Walker & Engelhard, 2016) so person fit analysis is an important part of documenting validity evidence.
- Much of the available research on person fit in Computerized Adaptive Testing (CAT) utilized traditional person fit statistics for detecting person misfit.
- Among the studied approaches, cumulative sum (CUSUM) procedures have been found powerful in CAT but when the parameters of the underlying statistical model are known before and after the change in response string (which doesn't hold in most CAT applications).
- A comprehensive approach with multiple indicators of person fit may be needed.

In this editorial chapter, I aim to summarize findings on person fit analysis in computerized adaptive testing (CAT) from prior research and discuss potential avenues for further research. In item response theory (IRT) applications, person fit quantifies fit of a response pattern to the model (Bradlow & Weiss, 2001, p. 86). Person misfit refers to unexpected response patterns by individuals. There are many potential reasons of misfit including special knowledge (Sinharay, 2016), cheating, guessing (Meijer, 1996), fatigue (Swearingen, 1998), warming up (Meijer, 1996), or faking (Ferrando & Anguiano-Carrasco, 2012). Evaluation of misfit is a significant step for addressing discrepancies within the measurement model. When IRT models are used, evidence of model fit which involves person fit analysis results should be reported (Standard 4.10; AERA, APA & NCME, 2014) as validity evidence to enhance score interpretations. Once misfitting items are identified, corrective steps such as item revision or removal can be implemented. For examinees who exhibit misfit, additional exploration can be undertaken to pinpoint behaviors that might necessitate adjustments to the test program or corrective interventions for particular examinees.

Although IRT estimates are robust to model-data misfit and many control mechanisms, involving both statistical (i.e., standardized log-likelihood index) and graphical approaches (i.e., person response plots), are available to detect person misfit, respondents in real test administrations may respond to items in unique and unstudied way (Walker & Engelhard, 2016). In addition, available misfit measures are specifically designed for fixed-item tests and have lower power when used with adaptive testing (van Krimpen-Stoop & Meijer, 1999, Meijer & van Krimpen-Stoop, 2010, Robin, 2002). This comes from two advantageous features of CAT that is item selection mechanisms which result in shorter tests and

* Assoc. Prof. Dr., Bartın University, Faculty of Education, Bartın-Turkiye, baksu@bartin.edu.tr, ORCID ID: 0000-0003-4994-1429

To cite this article:
Dünya, B.A. (2024). Adapting the person fit analysis: ideas on detecting person misfit in computerized adaptive testing, *15*(1), 1-4. https://doi.org/10.21031/epod.1461703

modest spread of item difficulties for an examinee (Meijer & van Krimpen-Stoop, & E.M.L.A. 2009, p. 32). In CAT, an item selection mechanism based on maximum information is utilized as part of the testing algorithm. This algorithm chooses items from an item pool that best match to the examinee's ability level. It aims to minimize the administration of items that are significantly too easy or too difficult for that examinee. Consequently, every examinee is presented with a unique test comprising items that are targeting for the examinee's ability level. Paradoxically, adaptive nature of CAT reduces the traditional sources of person misfit, while it poses a challenge for the detection of person misfit. In CAT, likelihood of inappropriate item selection that is too hard or too easy for a particular respondent is minimized. However, a person's responses should still be checked for fit to the IRT model chosen to calibrate parameters. Since different sets of items are drawn from an item pool with item parameters considered to be known, person fit checks in CAT, which may be absent in the item pool development stage, should provide additional quality check for data-model fit (Walker and Engelhard, 2016).

To address this concern attached to CAT applications, researchers have developed adaptive test specific person fit statistics and tested their misfit detection power (Hendravan, Glas & Meijer, 2005; McLeod & Lewis, 1999; van Krimpen-Stoop, 2000). A handful person fit indices that performed well in CAT depend on the CUSUM approach (i.e., LARD by Bradlow and Weiss, 2001; iterative upper and lower CUSUM by van Krimpen-Stoop & Meijer; 2000, 2001, 2002). This approach was found particularly successful at identifying abrupt shifts in response patterns, attributed to issues like decreased attention, speededness, or item preknowledge. CUSUM-based statistical process control mechanisms are found the most useful especially when the parameters of the underlying model before and after the change are known (Montgomery, 2013), which is not the case for CAT. Researchers addressed this shortcoming of CUSUM-based fit statistics for detecting person fit by proposing change-point based fit statistics (Tests for change point- TFCP;). Similar to the CUSUM approach, the logic of tests for change point (TFCP) is to find the point where the model parameters underlying a sequence of responses have changed in some fashion. This approach was tested for its usefulness for CAT since item parameters within an item pool are assumed to be known, whereas person parameters are not (Sinharay, 2016). Although TFCP-based fit statistics were found powerful in detecting unexpectedly abrupt change in response string, potential reasons of person misfit is not limited to this in CAT. An abrupt change in response strings can occur due to various reasons, such as initial warming up, speededness/fatigue or loss of attention through the end, or specialized content knowledge (Smith and Plackner, 2010) on a series of items during the test. Yet, these indicators might not always serve best in identifying misfit within a CAT context. For instance, to detect misfit caused by test fraud, including item memorization, pre-existing item knowledge, or item parameter drift, alternative approaches to diagnosing misfit may be required. Alternatively, Walker and Engelhard (2016) proposed a two step-approach for person misfit detection that integrates person response functions (PRF, Trabin & Weiss, 1979) to person fit statistics. Their approach enables to further investigate reason and location of misfit. Another piece of graphical evidence could be grounded in the adaptive nature of CATs. As the CAT progresses to later stages, variability in ability estimates is expected to decrease. Plotting the ability estimates against the sequence of item administration and drawing a line through these estimates can offer further visual insight into person misfit. Ideally, in a typical CAT administration, the slope of this line should approach to zero, indicating stabilization in the ability estimation process. Otherwise, a deviation from this pattern would signal a person's misfit and warrants further investigation.

Overall, reviewing the available literature on person fit in CAT, it appears there remains significant room for research, particularly in light of recent advancements in CAT research, such as multistage testing. The points below highlight essential areas for further investigation and aims to offer a foundation for researchers interested in exploring this field more deeply:

- Specific Challenges in Measuring Affective Constructs with CATs: CAT applications of psychological constructs can yield unique challenges in person-fit analysis, such as biases linked to social desirability. Developing indices specially designed for the nature of affective CATs, which address the varied reasons for person misfit in assessing psychological constructs, could be a viable approach.

- Holistic Fit Indices for Multi-scale CAT: Utilizing CAT to evaluate individuals across a range of dimensions, from cognitive abilities and personality traits to specific skill sets, is known for its precision and efficiency on individual scales (Maurelli & Weiss, 1981). A composite fit index that considers the interrelationships and collective performance across scales could increase the CAT's effectiveness, ensuring a holistic assessment of person fit.

- Lastly, investigating person fit within multistage CAT applications can offer a promising avenue for research, especially in light of recent studies such as Sideridis, Ghamdi & Zamil (2023), which compare the effectiveness of multistage CAT and traditional CAT. Their findings highlight a notable divergence in theta scores for high-ability examinees within multistage CAT frameworks, despite generally supporting multistage CAT's role in enhancing measurement accuracy. This discrepancy highlights the necessity for further exploration into how different multistage CAT designs handle misfit detection, particularly in scenarios involving high and low-ability examinees.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing.

Bradlow, E. T., & Weiss, R. E. (2001). Outlier measures and norming methods for computerized adaptive tests. Journal of Educational and Behavioral Statistics, 26(1), 85-104. https://doi.org/10.3102/10769986026001085

Chen, J., & Gupta, A. K. (2012). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance* (2nd ed.). Springer.

Csörgő, M., Horváth, L., & Szyszkowicz, B. (1997). Integral tests for suprema of Kiefer processes with application. Statistics & Risk Modeling, 15(4), 365-378. https://doi.org/10.1524/strm.1997.15.4.365

Ferrando, P. J., & Anguiano-Carrasco, C. (2012). Response Certainly, Conscienciousness, and Self-concept Clarity as antecedents of Acquiescence: A prediction model. Anuario de Psicología, 42(1), 103-112. https://psycnet.apa.org/record/2014-14293-007

Hendrawan, I., Glas, C. A., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. Applied Psychological Measurement, 29(1), 26-44. https://doi.org/10.3390/educsci10110324

Maurelli, V. A., & Weiss, D. J. (1981). Factors Influencing the Psychometric Characteristics of an Adaptive Testing Strategy for Test Batteries. http://files.eric.ed.gov/fulltext/ED212676.pdf

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. Applied Psychological Measurement, 23(2), 147-160. https://doi.org/10.1177/01466219922031275

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

Meijer, R.R., van Krimpen-Stoop, E.M.L.A. (2009). Detecting Person Misfit in Adaptive Testing. In: van der Linden, W., Glas, C. (eds) Elements of Adaptive Testing. Statistics for Social and Behavioral Sciences. Springer, New York, NY. https://doi.org/10.1007/978-0-387-85461-8_16

Robin, F. (2002). Investigating the relationship between test response behavior, measurement and person fit. In annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Sideridis, G., Ghamdi, H., & Zamil, O. (2023). Contrasting multistage and computer-based testing: score accuracy and aberrant responding. Frontiers in Psychology, 14, 1288177. https://doi.org/10.3389/fpsyg.2023.1288177

Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. Journal of Educational and Behavioral Statistics, 41(5), 521-549. https://doi.org/10.3102/1076998616658331

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). 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. https://apps.dtic.mil/sti/tr/pdf/ADA080933.pdf

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

van Krimpen-Stoop, E.M.L.A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. Applied Psychological Measurement, 23, 327-345. https://psycnet.apa.org/doi/10.1177/01466219922031446

Walker, A. A., & Engelhard, G. (2016). Using person fit and person response functions to examine the validity of person scores in computer adaptive tests. In Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings (pp. 369-381). Springer Singapore. http://dx.doi.org/10.1007/978-981-10-1687-5