Detecting Careless Responding to Assessment Items in a Virtual Learning Environment Using Person-fit Indices and Random Forest

Sanaz Nazari, Walter L. Leite, and A. Corinne Huggins-Manley
University of Florida
sanaznazari@ufl.edu

ABSTRACT

Careless responding and keeping students motivated for different tests have been common problems in many areas, especially in education. This study's objective was to demonstrate a novel approach to detect careless responding using person-fit indices developed within the field of psychometrics combined with a random forest. The data used was obtained from various tests in the Math Nation virtual learning platform. The result of person-fit indices as previously used measures of careless responding as well as the result of a random forest classifier to capture careless responding were compared by Receiver Operating characteristic (ROC) analysis and the area under the curve (AUC). The result showed that random forest combined with person-fit indices outperformed person-fit indices directly in detecting careless responding. Some important applications of this method for applied researchers are discussed in the conclusion section.

Keywords

Careless responding, Person-fit index, Random forest classifier, Virtual learning environment

1. INTRODUCTION

Paulhus (1991) defined response bias as "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content." Response biases are particularly important as they become a threat to the validity of conclusions and lead to measurement error (e.g., Meade & Craig, 2012). Three forms of biases frequently addressed in psychological and educational studies are acquiescence, careless responding, and social desirability bias (e.g., Cheung & Rensvold, 2000; Leite & Nazari, 2020; Wise, 2015). In formative assessments administered within virtual learning environments (VLE; Weller, 2007), it is hard to keep students motivated across different tests. Lack of motivation to complete formative assessments may result in careless responding, which in turn may produce biased estimates of student ability. This study focuses on careless responding, which occurs when students do not put much effort and thought into answering an assessment item (Voss & Vangsness, 2020). Wise and Kong (2005) have shown that careless responding can be associated with disengagement in an assessment. Other common names frequently used for this type of responding are non-effortful responding, inattentive responding, rapid guessing behavior, and examinee motivation (e.g., Rios & Soland, 2020; van Barneveld, 2007; Wise & DeMars, 2010; Wise, 2015; Wise and Kong, 2005). Throughout this paper, we use the "careless responding" label.

Researchers have proposed several different methods to identify careless responding. Attention check items including directed response items (e.g., "Please answer 'disagree' when responding to this item"), bogus items (e.g., "I do not understand a word of English," Meade & Craig, 2012), maximum longstring index, even-odd consistency, and Mahalanobis distance have been utilized to capture carelessness (e.g., Voss & Vangsness, 2020; Niessen et al. 2016). Self-report surveys have also been administered after the main test (e.g., Voss & Vangsness, 2020; Niessen et al. 2016; Wise, 2015), and response time effort and different types of time thresholds (Wise, 2017; Wise, 2015; Rios & Soland, 2020) are computed. Other researchers (e.g., Niessen et al., 2016; Patton et al., 2019) have used person-fit indices such as the number of Guttman errors and the standardized loglikelihood to detect careless responding. Since there is a wide range of available person-fit indices in the literature and each index captures misfitting persons from a different perspective, we opted for investigating person-fit indices.

Several nonparametric and parametric person-fit statistics (see Table 1) detect response biases that can be employed to detect carelessness (e.g., Karabatsos, 2003). Most of these indices were developed to target dichotomous items, and some of them have been extended to test polytomous items. Under the item response theory (IRT) framework, parametric person-fit indices assess the model fit at the individual level to examine the meaningfulness of obtained test scores (Embretson & Reise, 2013). In a way, the consistency of individuals' item response vectors is examined based on an IRT model by these indices (Embretson & Reise, 2013), which is the main concern when attempting to flag particular individual responses that may have been careless.

Although previous research on using person-fit indices to detect careless responding has provided promising results (Karabatsos, 2003), the current study attempted to improve the performance of person-fit indices by using multiple indices simultaneously. This is similar to considering multiple predictors in the model, which is possible by machine learning classifiers. Therefore, the present study sought to enhance capturing careless responding by using random forest (Breiman, 2001; Fernández-Delgado et al., 2014).). The study's research question is: Does the use of person-fit indices as features in a random forest to detect careless responding to items in formative assessments of a VLE outperform the use of person-fit indices by themselves? This

research question was answered with an analysis of responses to multiple 10-item formative assessments within the Math Nation VLE (Lastinger Center for Learning & University of Florida, 2019).

Table 1: Used Person-fit indices in PerFit package

Person-fit index	Function	Author
Nonparametric		
Personal point-biserial correlation	r.pbis	Donlon & Fischer (1968)
Caution statistic	C.Sato	Sato (1975)
Modified caution statistic	Cstar	Harnisch & Linn (1981)
Number of Guttman errors	G	van der Flier, 1977
Normalized Guttman errors	Gnormed	van der Flier, 1977
Agreement statistic	A.KB	Kane & Brennan (1980)
Disagreement statistic	D.KB	Kane & Brennan (1980)
Dependability statistic	E.KB	Kane & Brennan (1980)
U3 statistic	U3	van der Flier (1980)
Standardized normal U3	ZU3	van der Flier (1982)
Norm conformity index	NCI	Tatsuoka & Tatsuoka (1982, 1983)
HT statistic	Ht	Sijtsma (1986), Sijtsma and Meijer (1992)
Parametric		
Standardized normal loglikelihood	lz	Drasgow et al. (1985)
Corrected lz	lzstar	Snijders (2001)

1.1 Theoretical Framework

A comparison of 36 person-fit indices was performed by Karabatsos (2003) to examine the strength of fit indices to detect five types of aberrant responding (cheating, careless responding, lucky guessing, creative responding, and random responding). Results showed that Ht (Sijtsma, 1986) and then U3 (Van Der Flier, 1982) person-fit indices produced the highest area under the curve (AUC) to detect all types of aberrant responding, and other indices such as Guttman errors (Meijer, 1994) and lz (Drasgow, Levine, & Williams, 1985) indicated acceptable AUC. Additionally, in a simulation study, Artner (2016) investigated five well-known indices to detect guessing, cheating, careless behavior, distorting, and fatigue in responses and found that Ht, Cstar, and U3 performed better than OUTFIT and INFIT. Recently, another comparative study (Beck, Albano & Smith, 2018) measured response time and mentioned person-fit statistics to detect inattentive responding. Again, Ht was found to have the highest AUC. However, a major limitation of using person-fit indices for detecting careless responding to formative assessment items is that they only classify the entire response vector for a quiz as careless or not. However, it may be that only part of the responses to an assessment was careless, such as the last few items due to respondent fatigue. To overcome this limitation, we propose the random forest approach that utilizes person-fit indices as predictors of careless responding.

Over the last decade, as educational datasets have become larger, alongside substantial increases in computation speed, researchers have shown interest in using more complex machine learning classifiers. The random forest was first introduced by Breiman (2001) to overcome the problems of boosting (i.e., fitting trees to bootstrap-resampled data, e.g., Shapire et al., 1998) and bagging (i.e., bootstrap aggregating: fitting trees to reweighted data; Breiman, 1996) by adding another layer of randomness to bagging. The advantage of the random forest is that it chooses the best predictor among a subset of randomly selected predictors for splitting a node, while in a standard tree, the best predictor is chosen among all variables (Liaw & Wiener, 2002). By using this strategy, the random forest is robust against overfitting, and it outperforms other classifiers, such as discriminant analysis, support vector machines, and neural networks in some situations (Breiman, 2001).

One major advantage of combining the random forest with person-fit indices to detect careless responding is that classification of responses as careless or not can be done at the item response level rather than the person-level. Therefore, for the vector of responses a student provides for a formative assessment, only some responses may be classified as careless. In other words, carelessness can be viewed less as a static person feature on a test and more as a feature of the interaction between a particular person and a particular assessment item. Another advantage of the proposed method is that it can simultaneously use multiple person-fit indices, allowing optimal use of each index's unique sensitivity to careless responding.

2. METHODS

2.1 Participants

The sample for this study consisted of item responses from 14474 students obtained from the Algebra 1 section of Math Nation during the time that face-to-face instruction in schools in the state of Florida was canceled due to the threat of COVID-19 infection and replaced by online instruction. More specifically, the period that responses were collected was from March 18 to May 31, 2020. Math Nation focuses on preparing students for taking the high-stakes Algebra 1 End-of-Course exam required for high school graduation, usually administered in May by the Florida Department of Education. Because of COVID-19, the Algebra 1 End-of-Course exam was canceled, removing an important motivator for students to use Math Nation. However, Math Nation usage spiked during this period because teachers could use it as a resource to teach algebra while students were attending classes virtually. Therefore, this sample is well-suited for studying careless responding because students were making heavy use of a VLE and its assessment features, and yet they did not have the pressure of practicing for the high-stakes achievement test at the end of the school year, and they were engaging in schooling from home, which may expose them to many distractions.

2.2 Measures

There are a total of 10 sections in Math Nation, which corresponds to major concepts in Algebra, such as linear functions and quadratic functions. Each section has between 6 and 12 formative assessments with 3 items and one formative assessment with 10 items randomly drawn uniquely for each student from an item pool. The students can take these assessments as many times as they like. This study's data included responses to 40 items of the item pool of the 10-item formative assessment of "Section 9: One-Variable Statistics" of Math Nation. We chose this section because it had the highest number of responses to items during the time period of interest. We chose to focus on the 10-item assessment rather than the 3-item ones because teachers frequently ask students to complete the 10-item assessment before moving forward to the next section of Math Nation. Once completed, students have the option to review their answers and watch solution videos for each item.

The items used in this study had multiple formats (e.g., single choice, multiple-choice, constructed response) but were scored dichotomously (i.e., correct, or incorrect). The items were written to mimic the content and format of items on the statewide Algebra 1 End-of-Course exam and have been found in our research project to correlate strongly to scores on that exam. Difficulty and discrimination parameters for the items in the study were taken from the estimates reported in a previous study (Xue et al., 2020), which used the 2-parameter logistic (2PL) item response theory model (IRT; Birnbaum, 1968).

2.3 Analysis

In the current study, we compared the effectiveness of 14 nonparametric and parametric person-fit statistics (see Table 1) as well as the performance of the random forest classifier to capture carelessly answered responses. These person-fit indices are available in the PerFit (Tendeiro, Meijer, & Niessen, 2016) package of R statistical software (R Core Team, 2018). To obtain nonparametric person-fit indices, no IRT model is required. However, for the parametric lz and lzstar, the item parameters (i.e., difficulty and discrimination) for a 2PL IRT model were obtained from Xue et al. (2020) and used to estimate the student ability parameters. The 2-PL model formula is

$$P(Y_{is} = 1|\theta_s) = \frac{exp^{1.7a_i(\theta_s - b_i)}}{1 + exp^{1.7a_i(\theta_s - b_i)}},$$

where P indicates probability, θ is a latent trait of ability, the subscript i indicates an item, Y is an item response, b is item difficulty, a is item discrimination, the subscript s indicates a student, and 1.7 is a scaling constant. This formula gives us the probability of correct response (Y=1) for item i and person s conditional on the individual's ability.

All 14 person-fit indices were computed using the *PerFit* (Tendeiro, Meijer, & Niessen, 2016) package for all selected individuals in the sample and the result of the person-fit indices (person-fit scores) was saved to be used as part of the predictors for the random forest. Additional predictors were item difficulty and discrimination estimated parameters, the number of items answered by each student, ranging from 10 to 40 items, and the number of correct items answered (from zero to 10). Within "Section 9: One-Variable Statistics", students could respond to all available 40 items by taking section tests multiple times. Only the responses for students who answered at least 10 items were retained for the analysis. The total number of answered items out of 40 were available in the dataset and used as a predictor of carelessness in the random forest.

Random forest was implemented with the *randomForest* (Liaw & Wiener, 2002) package in R. To identify carelessness in responses for the model training, the time taken by each student to answer each item was recorded, and by comparing the graph of the frequency of correct responses versus incorrect responses across time for each item, empirical cutoffs were determined. In most cases, the point in time where the frequency of incorrect responses was decreasing, and the frequency of correct responses started to increase, was chosen as a cutoff for carelessness. For example, for item two, a cutoff of eight seconds was determined (see blue dashed line in Figure 1). Therefore, student responses to item two that were recorded in less than eight seconds were coded as careless.

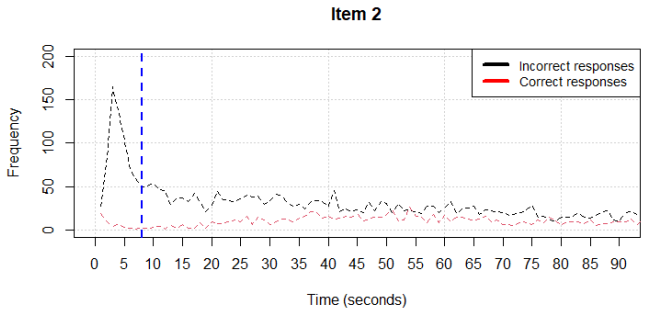

Figure 1: Frequency graph of correct responses versus incorrect responses across time for item 2.

To evaluate predictions by person-fit indices and random forest, we obtained the Receiver Operating Characteristic (ROC) curve. The ROC curve compares sensitivity against the specificity of a predictor for dichotomous data (Hanley, & McNeil, 1982). Sensitivity is the ability to identify true careless responding (true positive rate), and specificity is correctly identifying those items not careless (true negative rate). Usually, on ROC curves, "1-specificity" is demonstrated, which identifies the Type I Error rate (false positive rate; Hanley, & McNeil, 1982). The AUC of a ROC curve ranges from 0.5 to 1 from the identification line (i.e., the diagonal on the ROC) and represents the accuracy of the predictor or the feature, and different values of it can be interpreted to show the strength of a test. Generally, AUC that is 0.90-1 indicates an outstanding test, 0.80-0.90 is considered excellent, 0.70-0.80 is an acceptable one, and the AUC of about 0.5 suggests no discrimination (Hosmer, Lemeshow, & Sturdivant, 2013).

Creating ROC plots for person-fit indices requires the "true" careless responses (see above for flagging careless responses with time cutoffs) to compare them with estimated careless/non-careless responses by person-fit indices and calculate AUC. For person-fit indices, we defined a person as careless if he/she has responded to at least one item out of 10 carelessly based on the specified cutoffs. Regarding random forest ROC, true labels for carelessness were calculated per item response, and each student received 10 true labels of careless/non-careless for 10 answered items.

3. PRELIMINARY RESULT

Person-fit indices have been used as a measure to detect careless responding in previous studies (e.g., Niessen et al., 2016; Patton et al., 2019). In this study, all 14 available person-fit indices in the PerFit package were calculated for all students in the data, and the result are shown in Table 2. Most of these person-fit indices indicated a very poor AUC around 50%, which suggests no discrimination between careless/non-careless students. Therefore, only indices with the three best AUC were retained for later comparisons: 1) Guttman errors with 54.7%, 2) agreement statistic with 69.1%, and 3) dependability statistic with 63.5%. It is noteworthy that even the best performing index of agreement statistic with 69.1% AUC is not discriminating enough to be considered as an acceptable classifier of careless/non-careless students based on Hosmer, Lemeshow, and Sturdivant (2013) proposed cutoff of 70% AUC.

Table 2: AUC of 14 person-fit indices

Person-fit index	PerFit function	AUC (%)
Nonparametric		
Personal point-biserial correlation	r.pbis	50.2
Caution statistic	C.Sato	50.3
Modified caution statistic	Cstar	52.4
Number of Guttman errors	G	54.7
Normalized Guttman errors	Gnormed	49.9
Agreement statistic	A.KB	69.1
Disagreement statistic	D.KB	51.6
Dependability statistic	E.KB	63.5
U3 statistic	U3	52.4
Standardized normal U3	ZU3	50.1
Norm conformity index	NCI	50.1
HT statistic	Ht	50.4
Parametric		
Standardized normal loglikelihood	lz	50.1
Corrected lz	lzstar	50.1

The random forest has the advantage over person-fit indices in that it can use multiple person-fit indices as predictors but also include other predictors. The random forest included the three person-fit indices with the highest AUC and four additional predictors: estimated item difficulty and item discrimination parameters, number of correct responses, and number of items taken within the section. AUC of ROC illustrated that random forest with the set of specified predictors improved the classification and achieved the AUC of 77.2%, which is an acceptable test to distinguish between careless/non-careless responses (see Figure 2). The random forest also outperformed the best person-fit agreement statistic by about 8%.

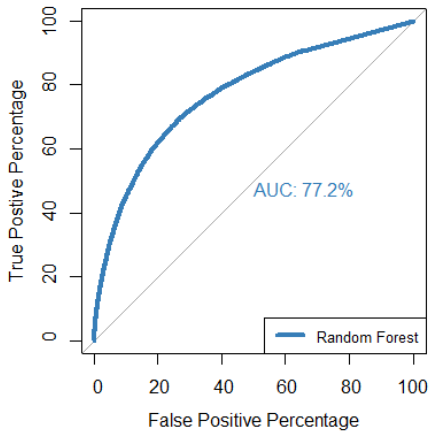

Figure 2: AUC of ROC plot for random forest classifier

4. DISCUSSION AND CONCLUSSION

The study objective was to compare the detection of careless responses between person-fit indices by themselves and random forest, including fit indices and other predictors. From the obtained results, we can conclude that random forest more accurately predicts careless responding, and this research took a methodological step forward in automatic identification of careless responding. By introducing different predictors, multiple dimensions are available, and a random forest can be used to investigate different parts of the data. In addition, careless responding can be conceptualized and examined as an item-person interaction rather than a static person feature. If desired, the random forest results can be aggregated to the person level (e.g., the proportion of responses from each person that were flagged as careless), allowing additional information at the person level. At this level, students can be labeled as careless responders or non-careless.

Person-fit indices as independent measures of carelessness may face some problems regarding ROC analysis. Sometimes, the true positive rates against false positive rates swing, cancel out, and end up in a very low AUC around 50% that is no discrimination. However, this issue does not occur in the random forest because of the way trees are constructed. Person-fit indices (and other predictors) can be removed from the random forest when they do not help improve classification accuracy.

One limitation of this study is that we relied on time cutoffs of each item response to create the criterion for careless responding. Like all options for "true" carelessness in responses, one could argue that our time-based flags of "true" careless responses are fallible in their own ways. However, empirical thresholds or time cutoffs have been used many times in previous research to detect careless responding (e.g., Wise & Kong, 2005; Wise & DeMars, 2010; Wise, 2015; Wise, 2017; Rios & Soland, 2020). Alternative criteria could come from surveys of students after they complete each formative assessment.

The result of this research will eventually be available as a trained random forest model to be used for applied researchers to detect careless responding in their data. They could enter the raw data to an R package to be developed in the future, and for the number of items in their test, they obtain a prediction of whether each answer is careless or non-careless. Then, within the context of their research, they can decide how they would like to aggregate and interpret carelessness (i.e., at the person level or at the item level) and make decisions according to the available results.

5. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

6. REFERENCES

[1] Artner, R. (2016). A simulation study of person-fit in the Rasch model. *Psychological Test and Assessment Modeling*, *58*(3), 531-563. Retrieved from https://lirias.kuleuven.be/retrieve/523896

[2] Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-Fit as an Index of Inattentive Responding: A Comparison of Methods Using Polytomous Survey Data. *Applied Psychological Measurement*, *43*(5), 374-387.. https://doi.org/10.1177%2F0146621618798666

[3] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, Contributed

chapter. *Statistical theories of mental test scores*, Chapters-17.

[4] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140. https://doi.org/10.1007/BF00058655

[5] Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[6] Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of cross-cultural psychology*, *31*(2), 187-212. https://doi.org/10.1177%2F0022022100031002003

[7] Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*(1), 105-113. https://psycnet.apa.org/doi/10.1177/001316446802800110

[8] 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*(1), 67-86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

[9] Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

[10] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The journal of machine learning research*, *15*(1), 3133-3181. https://doi.org/10.1117/1.JRS.11.015020

[11] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[12] Harnisch, D., & Linn, R. (1981). Analysis of Item Response Patterns: Questionable Test Data and Dissimilar Curriculum Practices. *Journal of Educational Measurement*, *18*(3), 133-146. Retrieved October 15, 2020, from http://www.jstor.org/stable/1434737

[13] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[14] Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, *4*(1), 105-126. https://doi.org/10.1177%2F014662168000400111

[15] Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298. https://doi.org/10.1207/S15324818AME1604_2

[16] Lastinger Center for Learning, & University of Florida. (2019). *Algebra Nation*. Retrieved 9/20/2019 from http://lastingercenter.com/portfolio/algebra-nation-2/

[17] Leite W.L., Nazari S. (2020) Marlowe-Crowne Social Desirability Scale. In: Zeigler-Hill V., Shackelford T.K. (eds) Encyclopedia of Personality and Individual Differences. Springer, Cham. https://doi.org/10.1007/978-3-319-24612-3_45

[18] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22. Retrieved from https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

[19] Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437. http://dx.doi.org/10.1037/a0028085

[20] Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311-314. https://doi.org/10.1177%2F014662169401800402

[21] Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, *44*(3), 309-341. https://doi.org/10.3102%2F1076998618825116

[22] Paulhus, D. L. (1991). Measures of personality and social psychological attitudes. In J. P. Robinson & R. P. Shaver (Eds.), *Measures of social psychological attitudes series* (Vol. 1, pp. 17–59). San Diego: Academic. https://doi.org/10.1016/C2013-0-07551-2

[23] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/.

[24] Rios, J. A., & Soland, J. (2020). Parameter Estimation Accuracy of the Effort-Moderated Item Response Theory Model Under Multiple Assumption Violations. *Educational and Psychological Measurement*, 0013164420949896. https://doi.org/10.1177%2F0013164420949896

[25] Sato, T. (1975). The construction and interpretation of SP tables. Tokyo, Japan: Meiji Tosho.

[26] Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, *26*(5), 1651-1686. https://doi.org/10.1214/aos/1024691352

[27] Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, *7*(22), 131-145. Retrieved from https://research.tilburguniversity.edu/files/1030745/COEFFICI.PDF

[28] Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16*(2), 149-157. https://doi.org/10.1177%2F014662169201600204

[29] Snijders, T. A. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331-342. https://doi.org/10.1007/BF02294437

[30] Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, *7*(3), 215-231. https://doi.org/10.3102%2F10769986007003215

[31] Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, 221-230.

[32] Tatsuoka, K., & Tatsuoka, M. (1983). Spotting Erroneous Rules of Operation by the Individual Consistency Index. *Journal of Educational Measurement*, *20*(3), 221-230.

Retrieved October 15, 2020, from http://www.jstor.org/stable/1434713

[33] Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1-27. http://dx.doi.org/10.18637/jss.v074.i05

[34] van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement*, *31*(1), 31-46. https://doi.org/10.1177%2F0146621606286206

[35] Van der Flier, H. (1977). Environmental factors and deviant response patterns. *Basic problems in cross cultural psychology, Amsterdam: Swets & Seitlinger*.

[36] Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties*. Swets & Zeitlinger.

[37] Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*(3), 267-298. https://doi.org/10.1177%2F0022002182013003001

[38] Voss, N. M., & Vangsness, L. (2020). Is Procrastination Related to Low-Quality Data?. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12355

[39] Weller, M. (2007). *Virtual learning environments: Using, choosing and developing your VLE*. Routledge.

[40] Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237-252. https://doi.org/10.1080/08957347.2015.1042155

[41] Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52-61. https://doi.org/10.1111/emip.12165

[42] Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27-41. https://doi.org/10.1080/10627191003673216

[43] Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

[44] Xue, K., Huggins-Manley, A. C., & Leite, W. L. (2020). Semi-supervised Learning Method for Adjusting Biased Item Difficulty Estimates Caused by Nonignorable Missingness under 2PL-IRT Model In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.), *Proceedings of The 13th Conference of Educational Data Mining*. https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_217.pdf