Tabaran Institute of Higher Education ISSN 2476-5880
International Journal of Language Testing
Special Issue, 2023: *Advanced Psychometric Methods in Language Testing*
Research Note

Detecting Measurement Disturbances: Graphical Illustrations of Item Characteristic Curves

Bolathan Yessimov[1], Rasha Abed Hussein[2], Aisha Mohammed[3], Aalaa Yaseen Hassan[4], Ahmed M. Hashim[5], Salma Saad Najeeb[6], Yusra Mohammed Ali[7], Ahmed Salim Abdullah[8], Al Khateeb Nashaat Sultan Afif[9]

Abstract
Measurement disturbances refer to any conditions that affect the measurement of some psychological latent variables, which result in an inaccurate interpretation of item or person estimates derived from a measurement model. Measurement disturbances are mainly attributed to the characteristics of the person, the properties of the items, and the interaction between the characteristics of the person and the features of the items. Although numerous researchers have detected measurement disturbances in different contexts, too little attention has been devoted to exploring measurement disturbances within the context of language testing and assessment, especially using graphical displays. This study aimed to show the utility of graphical displays, which surpass numeric values of infit and outfit statistics given by the Rasch model, to explore measurement disturbances in a listening comprehension test. Results of the study showed two types of outcomes for examining graphical displays and their corresponding numeric fit values: congruent and incongruent associations. It turned out that graphical displays can provide diagnostic information about the performance of test items which might not be captured through numeric values.

Keywords: Graphical displays; item characteristic curves; measurement disturbances; model-data fit

1. Introduction

Rasch model is a probabilistic model used to analyze categorical data as a logistic function of the trade-off between the examinee's abilities and the item difficulty. The higher the ability of the person, in comparison to the difficulty of the item, the more likely that person is to give a correct answer. Studies have indicated that individual differences can be accurately measured on a linear continuum in the Rasch model (Rasch, 1960/1980; Wright & Douglas, 1977). A unique feature of the model is that item difficulty parameters can be estimated without

[1] Institute of Natural Sciences and Geography, Abai Kazakh National Pedagogical University, Almaty, Kazakhstan, 13, Dostyk Av., 050010 Almaty, the Republic of Kazakhstan; ORCID 0000-0003-0313-1391
[2] Al-Manara College For Medical Sciences, Misan, Iraq
[3] College of Education, Al-Farahidi University, Baghdad , Iraq
[4] Al-Nisour University College, Baghdad, Iraq
[5] Department of English Language, Mazaya University College/Iraq
[6] Al-Esraa University College, Baghdad, Iraq
[7] Department of Medical Laboratory Technics, Al-Zahrawi University College, Karbala, Iraq
[8] English Department, AlNoor University College, Nineveh, Iraq
[9] People's Friendship University of Russia, Moscow, Russia

considering person ability measures, and person ability parameters can be estimated without considering item difficulty calibrations (Effatpanah & Baghaei, 2021). This feature is known as specific objectivity (Rasch, 1960/1980). Based on this feature, it is thus possible to accurately estimate item difficulties and person ability. However, as measurement is prone to error, measurement disturbances must be taken into account (Abdulridah Dhyaaldian et al., 2022; Baghaei, 2021; Firoozi, 2021; Smith, 1991; Tabatabaee-Yazdi et al., 2021). Winsteps, the software used to run the Rasch model, allows for plotting empirical item characteristics curves (ICCs) against the true probability curve. This allows for a straightforward method to make a comparison between test items for measurement disturbances.

2. Review of Literature

Measurement disturbances refer to any conditions that affect the measurement of some psychological latent variables, which result in an inaccurate interpretation of item or person estimates (Schumacker, 2015). As argued by Smith (1985), measurement disturbances can be categorized into three classes. The first class includes disturbances that are attributed to the characteristics of the person such as fatigue, boredom, illness, cheating, plodding, and start-up. The second class consists of disturbances attributed to the interaction between the characteristics of the person and the features of the items such as item type, item content, item bias, and guessing. Within the framework of the Rasch model, there are only two conditions for identifying the interaction between the person and the items of a given test (Schumacker, 2015, p. 77): (1) person ability: the amount of the trait or ability possessed by the person, and (2) person ability: the amount of the trait required to give a correct response to a given test item. Any other conditions that go beyond these two conditions are viewed as measurement disturbances. Finally, the third class involves disturbances attributed to the properties of the items. Since these three conditions affect the measurement process and are sources of construct-irrelevant variances, measurement disturbances are considered as a validity concern which may compromise the interpretation and use of item and person measures (AERA, APA, NCME, 2014; Afsharrad et al., 2023).

Using a variety of item response theory (IRT) models, numerous researchers have used graphical approaches to identify measurement disturbances in different contexts. For instance, under the framework of the Rasch model, Schumacker (2015) made a comparison between empirical and theoretical item response functions (IRFs) to detect unexpected response patterns that might reflect a kind of measurement disturbance. He found that visual displays offer a diagnostic way of identifying measurement disturbances or systematic patterns of misfit across different levels of the construct that are unobservable through the use of model-data fit statistics, including the outfit and infit indices widely employed in the Rasch model. Wind and Schumacker (2017), in the rater-mediated context, conducted a study to show the utility of graphical ways to detect measurement disturbances for raters. The results showed that visual illustrations can be employed to explore measurement disturbances for raters. They also corroborated the findings of Schumacker's (2015) study in which graphical displays have diagnostic value for exploring measurement disturbances that are not captured through the use of model-data fit indices. In a similar vein, Effatpanah and Baghaei (2022) recently applied the

Kernel smoothing IRT (Ramsay, 1991) as a non-parametric IRT model to examine the scoring patterns of raters. They argued that graphical displays obtained from the Kernel smoothing IRT have the potential to explore various measurement disturbances and rater effects.

Although previous studies have shown that graphical displays permit the detection of different measurement disturbances related to test items beyond the information given by the widely used item fit statistics, too little attention has been paid to exploring measurement disturbances within the context of language testing and assessment. The current study seeks to demonstrate the utility of visual illustrations to detect measurement disturbances in a listening comprehension test that goes beyond numeric values of infit and outfit statistics given by the Rasch model.

3. Method

3.1.Participants and Setting

Participants in this study were 191 intermediate English as a foreign language (EFL) students in the College of Education, Al-Farahidi University, Baghdad, Iraq. There were 109 female and 82 male students. The ages of these participants ranged from 19 to 41 (M = 23.86; SD = 4.59).

3.2.Instrumentation

To assess the listening comprehension ability of the students, an English listening comprehension test was administered. The test consisted of 20 multiple-choice items. The score of the test varied from 0 to 19 (M = 9.75; SD = 3.62). Reliability coefficients of the test were estimated using Cronbach alpha, and a value of 0.70 was obtained which is acceptable.

4. Results

The WINSTEPS computer package Version 3.73 (Linacre, 2009a) was used for the analyses. Table (1) displays item measures, fit statistics, and corrected item-total correlation. The left column gives the number of items. Columns two and three present item difficulties in logits and their error of measurement, respectively. As can be seen, Items 18 and 20 were the most difficult items with difficulties of 2.24 and 1.72, respectively, and Items 6 and 7 were the easiest with difficulties of -1.77 and -1.62, respectively. Columns four and five provide INFTIT and OUTFIT means squares statistics, respectively. These two indices indicate to what extent the items represent the single underlying latent trait being measured. As given in Table (1), except for Items 18 and 20, all INFIT and OUTFIT values were within the ideal range from 0.5 to 1.5 (Linacre, 2009b). The last column shows the corrected item-total correlations obtained from SPSS. It is calculated to measure the strength of the relationship between each item score and the total score of the test. Higher positive values for the item-total correlation indicate *that the item is discriminating well* between low- and high-performing examinees. Among the twenty items of the test, Items 1, 19, and 7 are the most discriminating items, and Items 20, 10, and 18 are the least discriminating items.

Figure 1 illustrates the ICCs for eight items (e.g., 2, 5, 8, 10, 15, 18, 19, and 20) of the test. Two major outcomes were observed while investigating the numeric and graphical

indicators of measurement disturbances. As Figure 1a demonstrates, the first outcome was the presence of a congruence between the numeric indices and graphical displays. For example, for Items 8 and 15, the values of INFIT and OUTFIT MNSQs were within the ideal boundary of 0.5 and 1.5, indicating the best fitting items. Their graphical displays also showed no considerable measurement disturbances. In each plot, the empirical ICCs are shown by a blue line and the theoretical ICCs by a red curve. The upper and lower 95% confidence intervals are shown by the lines on the two sides of the ICC. For items 8 and 15, the empirical ICCs are within the intervals and close to the logistic S-shaped expected ICC. Similarly, the analysis of empirical ICCs of Items 18 and 20, as the worst fitting items, and their corresponding fit statistics values showed that there is a congruency between graphical displays and numeric values. As Figure 1b present, the ICCs of Items 18 and 20 indicate that the empirical ICCs have deviated from the expected S-shaped curve, and there is a very large gap between the two lines showing confidence intervals. The same scenario was also true for Items 1, 4, 6, 7, 12, and 17 when the results of numeric values and graphical displays were congruent.

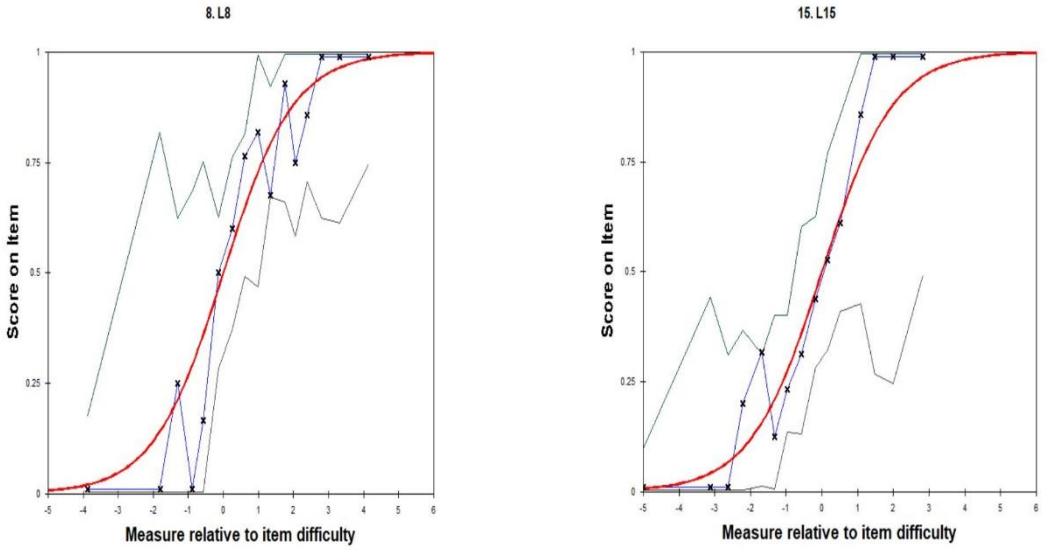
Table 1
Item Measures, Fit Statistics, and Corrected Item-Total Correlation for the Listening Test

Items	Item Difficulty	Model S.E.	INFIT MNSQ	OUTFIT MNSQ	Corrected Item-Total Correlation
1	0.00	0.16	0.82	0.76	0.489
2	0.03	0.16	1.00	1.08	0.276
3	0.74	0.17	0.87	0.94	0.398
4	0.33	0.16	1.08	1.04	0.204
5	0.23	0.16	1.14	1.18	0.129
6	-1.77	0.20	0.84	0.64	0.447
7	-1.62	0.19	0.84	0.68	0.449
8	-0.75	0.17	1.01	1.02	0.266
9	-0.97	0.17	1.09	1.14	0.189
10	-1.12	0.17	1.24	1.38	0.012
11	-0.51	0.16	1.04	1.08	0.227
12	0.31	0.16	0.93	0.93	0.345
13	0.05	0.16	0.87	0.86	0.425
14	0.74	0.17	0.97	0.92	0.309
15	0.57	0.16	1.01	1.04	0.252
16	0.82	0.17	1.18	1.26	0.056
17	-0.86	0.17	0.87	0.79	0.426
18	2.24	0.23	1.10	1.95	0.038
19	-0.17	0.16	0.81	0.83	0.489
20	1.72	0.20	1.21	1.62	-0.021

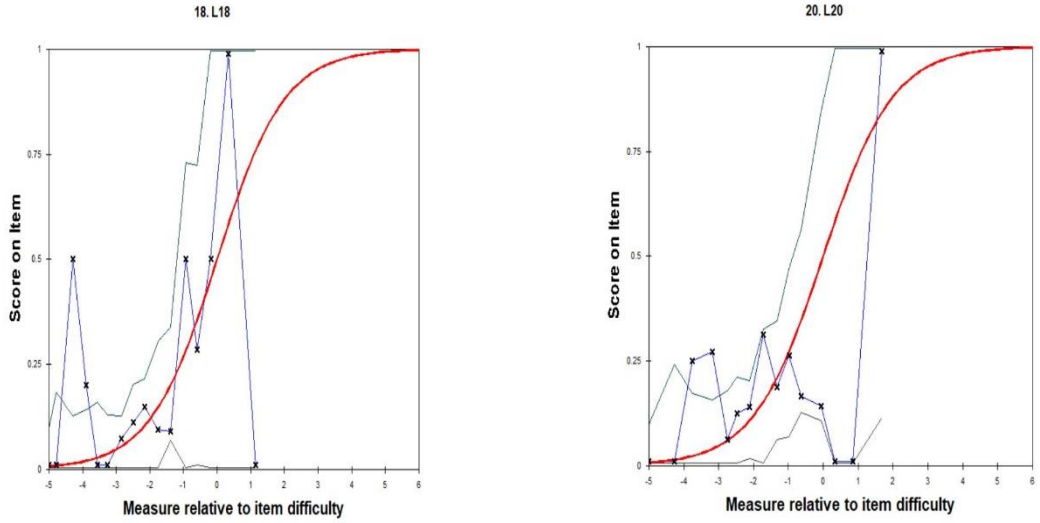
Note. S.E. = Standard error of measurement; MNSQ = Mean Squaure

On the other hand, another major outcome was when there is no congruence between numeric values and graphical displays. For example, fit statistics values for Items 2 and 19 show the adequate fit of the items to the model; however, the analysis of empirical and theoretical ICCs indicated significant measurement disturbances that suggest further examination (See Figure 1c). In the same way, as depicted in Figure 1d, for Items 5 and 10, the numeric values and graphical displays showed a disagreement because the ICCs of the two items deviated from the theoretical ICC at some points, although most parts of the ICCs were within the ideal boundary. It is thus possible to estimate item and person parameters on a unidimensional continuum. The same scenario was true for Items 3, 9, 11, 13, 14, and 16 when the results of numeric values and graphical displays were incongruent.

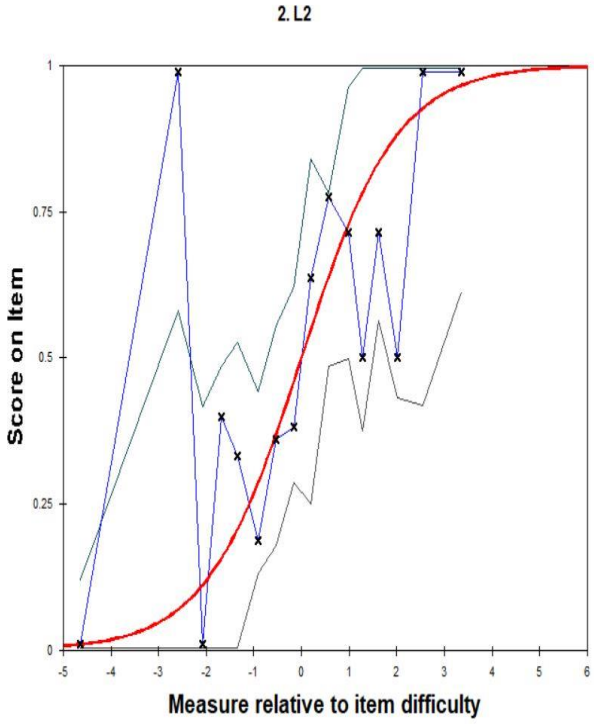
Figure 1
Empirical and Theoretical Item Characteristic Curves for Eight Items of the Test

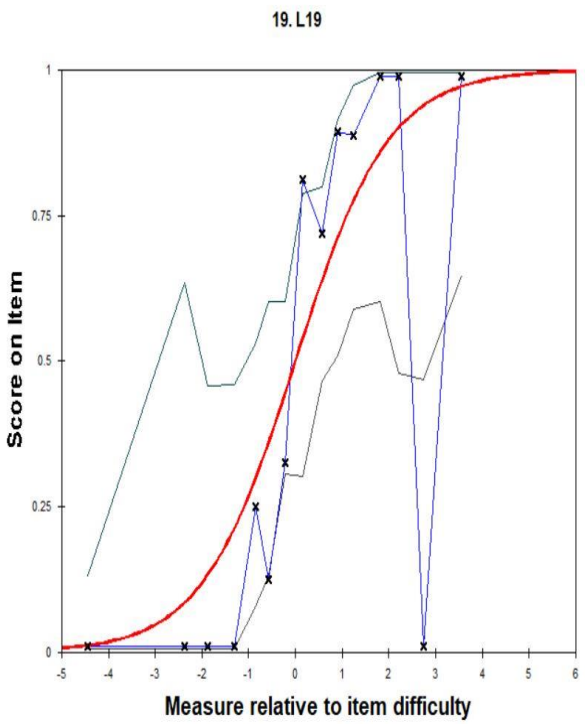


(a)



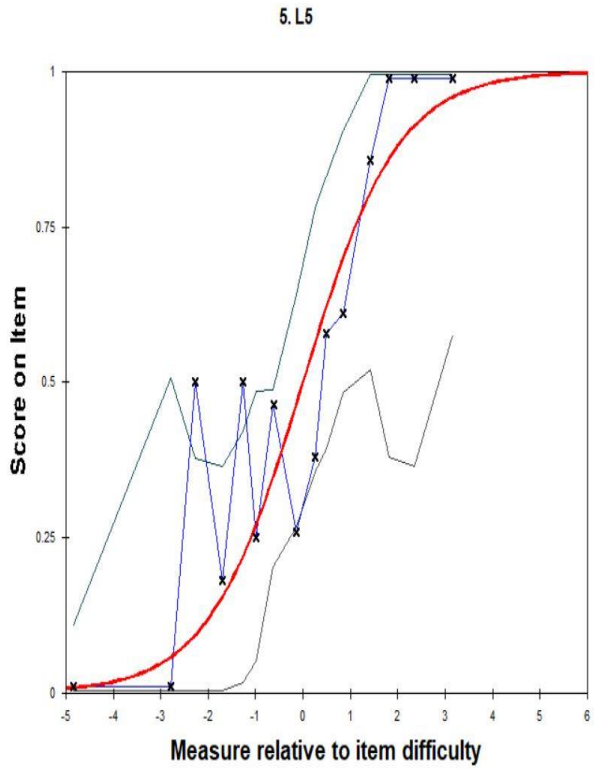
(b)

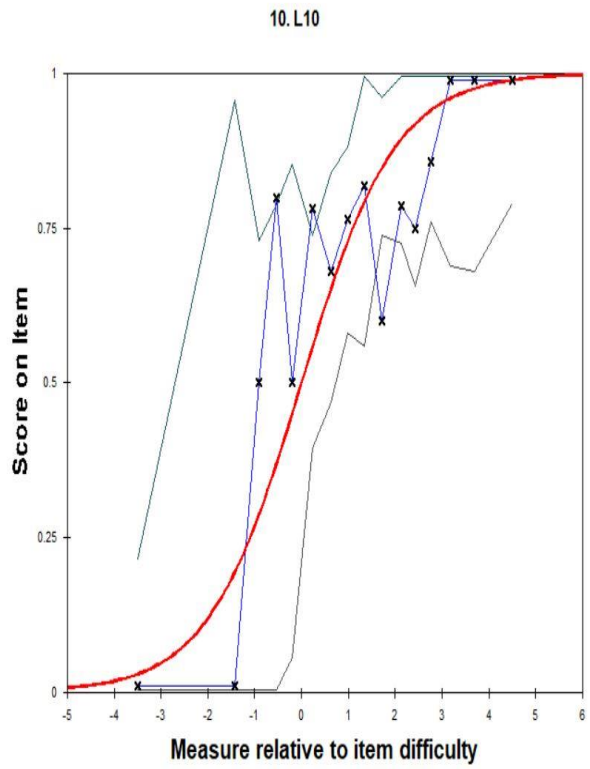

2. L2
Score on Item
Measure relative to item difficulty


19. L19
Score on Item
Measure relative to item difficulty

(c)

(d)


5. L5
Score on Item
Measure relative to item difficulty


10. L10
Score on Item
Measure relative to item difficulty

5. Conclusion

Using the Rasch model, this study aimed to show the graphical illustrations to identify measurement disturbances in a listening comprehension test. The results of the current study converge with the findings of the study conducted by Wind and Schumacker (2017) who found two kinds of outcomes for the investigation of graphical displays and their corresponding numeric fit values within a rater-mediated assessment context. The first type was when the results of numeric values and graphical displays were congruent, and the second type was the presence of incongruent conclusions between numeric values and graphical illustrations. As Wind and Schumacker (2017, p. 7) recommend, because "Rasch fit statistics often mask patterns in residuals", it is highly recommended for routine investigations of model-data fit to include visual illustrations which can provide diagnostic information about the performance of test items which might not be observable through several numeric summaries of model-data fit.

Even though graphical analyses can provide much more diagnostic information for detecting measurement disturbances which exceed the use of numeric values of fit statistics, most practitioners and researchers are still dependent on numeric model-data fit indices to explore misfitting items and persons. According to Meijer et al. (2015) "[t]here seems to be a great reluctance by specially trained psychometricians to use graphs. We often see fit statistics and large tables full of numbers that certainly do not provide more information than graphs" (p. 89). Further research is required to illustrate the effectiveness of visual displays in identifying measurement disturbances in educational measurement and language testing, in particular.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no specific funding for this work from any funding agencies.

References

Abdulridah Dhyaaldian, S. M., Al-Zubaidi, S. H., Mutlak, D. A., Neamah, N. R., Ali Albeer, A. A. M., Hamad, D. A., Al Hasani, S. F., Musa Jaber, M., & Ghaleb Maabreh, H. (2022). Psychometric Evaluation of Cloze Tests with the Rasch Model. *International Journal of Language Testing*, *12*(2), 95-106. https://doi.org/10.22034/ijlt.2022.157127

Afsharrad, M., Pishghadam, R., & Baghaei, P. (2023). A comparison of the added value of subscores across two subscore augmentation methods. *International Journal of Language Testing*, *13*(Special Issue), 109-125. https://doi.org/10.22034/ijlt.2023.386592.1234

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

Baghaei, P. (2021). *Mokken scale analysis in language assessment*.Waxmann Verlag.

Effatpanah, F., & Baghaei, P. (2021). Cognitive components of writing in a second language: An analysis with the linear logistic test model. *Psychological Test and Assessment Modeling*, *63*(1), 13-44. URL:https://www.psychologie-aktuell.com/journale/psychological-test-and-assessment- modeling/currently-available.html

Effatpanah, F., & Baghaei, P. (2022). Exploring rater quality in rater-mediated assessment using the non-parametric item characteristic curve estimation. *Psychological Test and Assessment Modeling, 64*(3), 216-252. URL:https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_2022-3/PTAM__3-2022_2_kor.pdf

Firoozi, F. (2021). Mokken Scale Analysis of the Reading Comprehension Section of the International English Language Testing System (IELTS). *International Journal of Language Testing*, *11*(2), 91-108. URL: https://www.ijlt.ir/article_138059.html

Linacre, J. M. (2009a). *WINSTEPS Rasch Measurement* (Version 3.73) [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2009b). *A user's guide to WINSTEPS*. Winsteps.com

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2015). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 85-110). Routledge.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630. https://doi.org/10.1007/BF0229449

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, 1960. (Expanded edition, The university of Chicago Press, 1980).

Schumacker, R. E. (2015). Detecting measurement disturbance effects: The graphical display of item characteristics. *Journal of Applied Measurement, 16*, 76-81. URL:http://jam-press.org/abst2015.htm

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*(3), 433-444. https://doi.org/10.1177/001316448504500301

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, *51*(3), 541-565. https://doi.org/10.1177/0013164491513003

Tabatabaee-Yazdi, M., Motallebzadeh, K., Baghaei, P. (2021). A Mokken scale analysis of an English reading comprehension test. *International Journal of Language Testing*, *11*(1), 132-143. https://www.ijlt.ir/article_130373.html

Wind, S. A., & Schumacker, R. E. (2017). Detecting measurement disturbances in rater mediated assessments. *Educational Measurement: Issues and Practice*, *36*(4), 44-51. https://doi.org/10.1111/emip.12164

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement*, *1*(2), 281-295. https://doi.org/10.1177/014662167700100216