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

Evaluating Measurement Invariance in the IELTS Listening Comprehension Test

Maabreh Hatem Ghaleb*[1], Osama Wael Suleiman[2], Aisha Mohammed[3], Zaid Hilal Abed Alqiraishi[4], Hamid Khalaf Mutar[5], Yusra Mohammed Ali[6], Firas Zakariya Hadi[7], Akram Ali Anber[8], John Emaimo[9], Liudmila Georgievna Karandeeva[1]

Abstract
Measurement invariance (MI) refers to the degree to which a measurement instrument or scale produces consistent results across different groups or populations. It basically shows whether the same construct is measured in the same way across different groups, such as different cultures, genders, or age groups. If MI is established, it means that scores on the test can be compared meaningfully across different groups. To establish MI mostly confirmatory factor analysis methods are used. In this study, we aim to examine MI using the Rasch model. The responses of 211 EFL learners to the listening section of the IETLS were examined for MI across gender and randomly selected subsamples. The item difficulty measures were compared graphically using the Rasch model. Findings showed that except for a few items, the IELTS listening items exhibit MI. Therefore, score comparisons across gender and other unknown subgroups are valid with the IELTS listening scores.

Keywords: Differential Item Functioning, IELTS, Measurement Invariance, Rasch Model

1. Introduction

Measurement invariance (MI) is important for ensuring that comparisons between groups are valid and reliable. If a measurement instrument does not exhibit measurement invariance, it may be biased and produce inaccurate results when used with different groups. Measurement invariance within item response theory refers to the extent to which the measurement properties of a test or scale remain consistent across different groups or populations. Specifically, it refers to the degree to which the same underlying construct is being measured in the same way across different groups, such as different cultural or linguistic groups, gender, age, or educational levels (Bond et al., 2020).

[1]* Peoples' Friendship University of Russia, Moscow, Russia. Email: maabrekh_kh@rudn.ru; ORCID: https://orcid.org/0000-0003-4851-6420
[2] English Department, Al-Noor University College, Nineveh, Iraq
[3] College of Education, Al-Farahidi University, Baghdad, Iraq
[4] College of Education, The Islamic University in Najaf, Iraq
[5] English Language and Literature Department, Al-Mustaqbal University College, Babylon, Iraq
[6] Department of Medical Laboratory Technics, Al-Zahrawi University College, Karbala, Iraq
[7] English Department, Mazaya University College, Iraq
[8] Al-Esraa University College, Baghdad, Iraq
[9] Federal School of Dental Technology & Therapy, Enugu, Nigeria

Within the Rasch measurement (Rasch, 1960/1980) methodology, MI refers to the property that the relationship between an individual's ability level and their response to an item is consistent across different groups or conditions. Specifically, it means that the Rasch model parameters (item difficulty and person ability) are equivalent across groups, indicating that the items are measuring the same construct in each group and that differences in item responses can be attributed to differences in ability levels rather than group membership. This is important for ensuring that scores obtained from a test or questionnaire are comparable across different populations or contexts.

Measurement invariance is essential because it allows researchers to make meaningful comparisons between groups and ensure that differences observed between groups are not due to measurement bias. In order to establish measurement invariance, researchers typically use statistical techniques such as confirmatory factor analysis (CFA) and differential item functioning (DIF) analysis to test whether the items on a test function similarly across different groups.

MI is also directly related to Rasch's specific objectivity (Rasch, 1977). Specific objectivity refers to the property of the Rasch model that ensures that the estimates of items and person parameters are independent of the particular sample of items or persons used in the analysis. This means that if two different samples of items or persons are analyzed using the Rasch model, the resulting estimates of item and person parameters will be equivalent, as long as the samples are measuring the same construct. This property is important because it allows for comparisons across different samples and ensures that results are not biased by specific characteristics of a particular sample (Baghaei & Doebler, 2019; Baghaei et al., 2019).

Invariance is also related to unidimensionality principle which is one of the core assumptions of all item response theory models including the Rasch model (Baghaei, 2021). Unidimensionality means that the test must measure a single latent trait. Examining unidimensionality is important in Rasch measurement because it ensures that the items in a test or questionnaire are measuring only one construct or trait. If the items are measuring multiple constructs, then the results of the analysis may be inaccurate and unreliable. Unidimensionality is also important for ensuring that the test or questionnaire is valid and can accurately measure what it intends to measure. By examining unidimensionality, researchers can identify problematic items and remove them from the analysis, leading to a more accurate and reliable measurement of the construct being studied.

The principle of measurement invariance is the basis of some of the most fundamental approaches for checking the overall fit of the Rasch model including Andersen's likelihood ratio test (Andersen, 1973) and Martin-Löf's test (Martin-Löf, 1973). Baghaei et al. (2017) also developed a descriptive fit statistic based on the invariance of the measures.

Following Baghaei (2010) and Ravand and Firoozi (2016), the current study aimed to examine the invariance of item and person measures in the IELTS listening section. Graphical methods were employed to check if the IELTS listening items remain invariant across subsamples and if

the test is unidimensional. Graphical methods provide very informative means to evaluate item quality (Dhyaaldian et al., 2023; Effatpanah & Baghaei, 2022/2023; Yessimov et al., 2023)**.**

2. Method

2.1. Instrument

The listening comprehension section of a retired version of the IELTS (International English Language Testing System) was employed for the purposes of this study. The test contained 40 binary items in four different formats of multiple-choice, gap-filing, table completion, and map completion. The first 20 items of the test were used for the sake of this demonstration.

2.2. Participants

The test was administrated to 211 undergraduate students of English as a foreign language at Al-Noor University College, Nineveh, Iraq. Participants (127 female and 84 male) aged from 19 to 38 (M=22.56, SD=3.97), and the test was administered as an approximate predictor of their IELTS band score. Participation in the test was voluntary and test takers gave written consent to be tested. Participants were provided with diagnostic feedback and their IELTS band score in return for their cooperation.

3. Analyses and Results

3.1. Item Measure and Fit

The 20 dichotomously scored listening items were analyzed with the unidimensional Rasch Model (Rasch, 1960/1980). Winsteps Rasch computer program (Linacre, 2023) was used to estimate the model. Table 1 shows the fit statistics and item difficulty parameters. As Table 1 shows, all the items have acceptable infit and outfit mean square values. The acceptable range was set from .70 to 1.30 (Wright & Linacre, 1994). Furthermore, the point-measure correlations are all positive and high.

Table 1.
Item Measure and Fit Values

Item	Measure	SE	Infit MNSQ	Outfit MNSQ	Point-Measure Cor.
1	-2.86	.25	1.06	.77	.45
2	-2.74	.25	.78	.51	.55
3	-2.20	.22	1.16	1.21	.43
4	-1.80	.20	1.08	1.02	.49
5	-1.89	.21	1.01	.78	.53
6	-1.41	.19	.83	.78	.61
7	-1.16	.19	.96	.87	.57
8	-.93	.18	1.01	.91	.56
9	.15	.17	.98	.96	.57

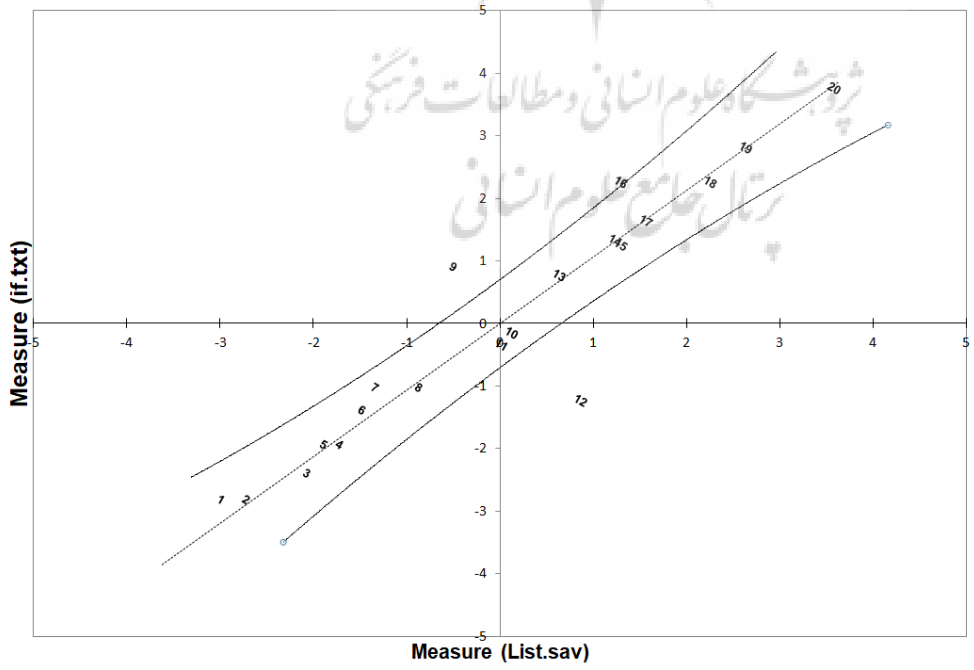
10	0.00	.17	1.01	1.12	.55
11	-.15	.17	.98	.84	.58
12	-.09	.17	1.09	1.15	.52
13	.69	.18	.98	1.00	.55
14	1.27	.18	1.10	1.59	.45
15	1.27	.18	.94	.79	.55
16	1.70	.19	1.06	.99	.47
17	1.59	.19	1.00	1.20	.49
18	2.24	.21	.96	.91	.46
19	2.69	.24	.88	.60	.48
20	3.64	.31	1.10	3.38	.28

3.2.Invariance across Sex

The invariance of the item parameters across sexes was evaluated graphically. The sample was divided into two partitions based on sex, i.e., female and male. The item parameters were separately estimated in these two subsample of the data and then they were brought onto the same scale. The item parameters from the two subsections were cross plotted against each other and 95% quality control lines based on the joint standard errors of the parameters were drawn (Wright & Stone, 1979). Figure 1 shows the cross plot of the item parameters. As can be seen in the figure, all the items, except Items 9 and 12, fall within the 95% quality control lines. This is an indication that these two items do not function similarly across males and females.

Figure 1.

Cross Plot of Items across Sex

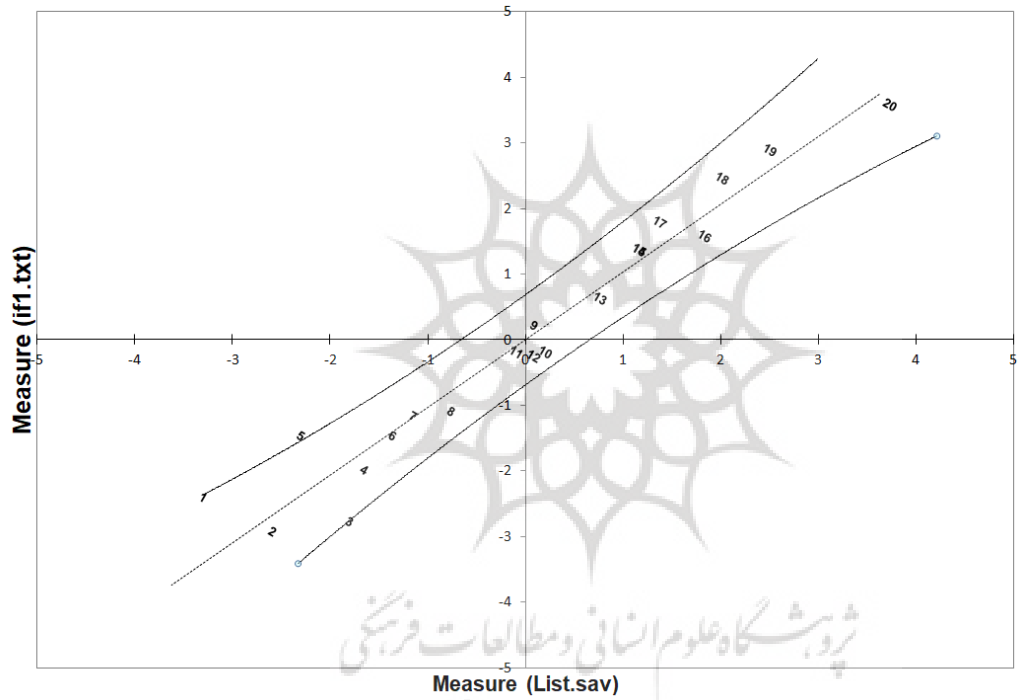


3.3. Invariance across Random Partitioning

In the next step, the sample was divided into subsets randomly and the item parameters were estimated separately and were brought onto the scale. Figure 2 shows the cross plot of the item parameters. As Figure 2 shows, all the items are behaving as expected and fall within the 95% confidence quality control lines. Only Item 5 looks to be a bit on the border but it is not a cause for concern.

Figure 2.

Cross Plot of Items across Random Subsamples

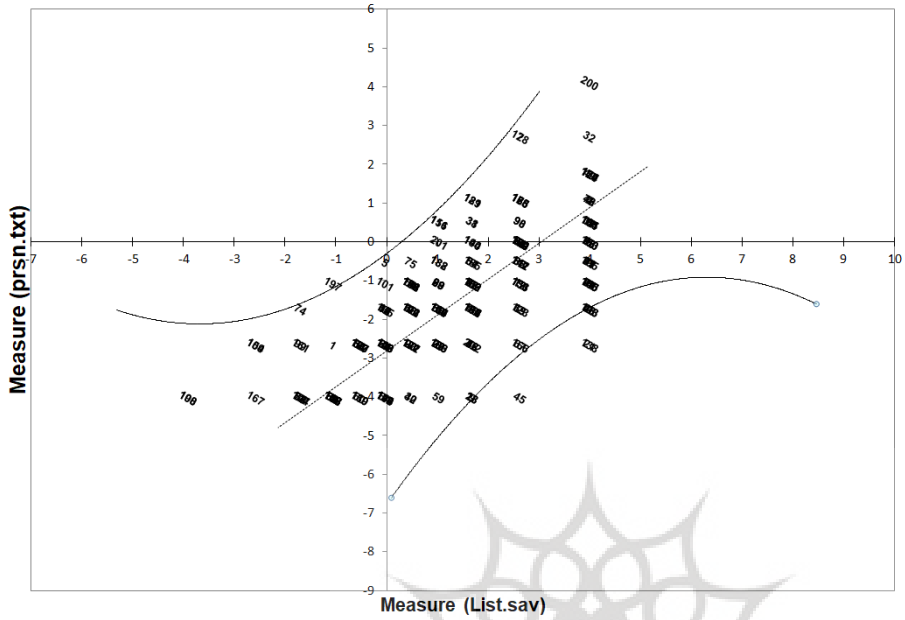


3.4. Invariance of Person Parameters across Subsets of Items

In the final stage, the items were divided into two subsets. The first 10 items made up Subset 1 and the second 10 items made up Subset 2. Examinees' ability parameters were estimated on the basis of these two subsets separately and then were brought onto the same scale using a shift constant. The item parameters from the two subsets were cross plotted against each other. The 95% quality control lines were constructed to indicate the degree to which the pairs of person estimates are allowed to diverge from each other. Figure 3 shows the outcome. As the figure shows, almost all the person parameters fall within the 95% quality control lines. Only a few persons fall outside the lines. According to Wright and Stone (1979), if 95% of the estimates are within the lines, the measurement has been invariant.

Figure 3.

Cross Plot of Persons across Subsets of Items



4. Discussion and Conclusion

Constant item difficulty estimates (measurement error should be considered) across subsamples of the data is an indication that the construct is constant for the two subsamples. Non-invariant item parameters across subsamples are a sign that the construct changes across the subsamples. If this occurs, it means that the test does not function similarly for the two groups and it is not clear what the test measures. Such a test does not allow comparison of all the examinees on the same scale. According to Wright et al. (2000), if item difficulty does not remain constant across relevant subsets of the sample, the difficulty would not have a practical meaning.

One basic assumption of the Rasch model that should be tested empirically is the item-free person measurement and the person-free item calibration. If the Rasch model fits, item and person parameters should remain stable across subsets of the sample and items. This is a method of checking the overall fit of the Rasch model and unidimensionality (Baghaei, 2010; Kubinger, 2005).

In this study, 20 listening comprehension items of the IELTS were tested for invariance across sex and a random partitioning. Graphical displays showed that while the items remained invariant across random subsets, two items were non-invariant across sex. This is an indication of gender differential item functioning (DIF) and requires examination of the item contents to find out the reason. The corollary of the procedure above was replicated for the items. The items were divided into two subsets and person parameters were compared across the two subsets. The graphical displays showed that more than 95% of the examinees had constant ability parameters (considering measurement error) across the two subsets of the items. This is interpreted as the unidimensionality of the test and the overall fit of the Rasch model to the data. This is also an

empirical check of the item-free person measurement of the Rasch model. The final conclusion of the current study is that the listening module of the IELTS is unidimensional, fits the Rasch measurement model, and allows the construction of non-invariant linear measures with a stable unit (Bond, 2003).

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140. DOI: 10.1007/BF02291180

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

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.). *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research* (pp.100-112). Frankfurt/M.: Lang.

Baghaei, P. & Doebler, P. (2019). Introduction to the Rasch Poisson Counts Model: An R tutorial. *Psychological Reports*, *122*(5), 1967-1994. DOI: 10.1177/0033294118797577

Baghaei, P., Ravand, H., & Nadri, M. (2019). Is the d2 test of attention Rasch scalable? Analysis with the Rasch Poisson Counts Model. *Perceptual and Motor Skills, 126,* 70-86. DOI: 10.1177/0031512518812183

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology, 19,* 155-168.

Bond, T. G. (2003). Validity and assessment: A Rasch measurement perspective. *Metodología de las Ciencias del Comportamiento, 5*(2), 179–194.

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Dhyaaldian, S. M. A., Al-Zubaidi, S. H., Mutlak, D. A., Neamah, N. R., Albeer, M. A., Hamad, D. A., Al Hasani, S. F., Jaber, M. M., & Maabreh, H. G. (2023). Psychometric evaluation of cloze tests with the Rasch model. *International Journal of Language Testing, Special Issue*, 95-106. DOI: 10.22034/IJLT.2022.157127

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.

Effatpanah, F., & Baghaei, P. (2023). Kernel smoothing item response theory in R: A didactic. *Practical Assessment, Research, and Evaluation, 28,* 1-26.

Kubinger, K. (2005). Psychological test calibration using the Rasch model: Some critical suggestions on traditional approaches. *International Journal of Testing*, *5*(4), 377–394. DOI: 10.1207/s15327574ijt0504_3

Martin-Löf, P. (1973). *Statistiska modeller [Statistical models.] Anteckningar från seminarier lasåret 1969-1970, utarbetade av Rolf Sundberg. Obetydligt ändrat nytryck, Oktober 1973*. Stockholm: Institütet för Försäkringsmatemetik och Matematisk Statistisk vid Stockholms Universitet.

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

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-93.

Ravand, H. & Firoozi, T. (2016). Examining construct validity of the Master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing, 6*(1), 1-23.

Wright, B. D., Huber, M., O'Neill, T. & Linacre, J. M. (2000). The problem of measure invariance. *Rasch Measurement Transactions* 14(2), 745.

Wright, B. D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.

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

Yessimov, B., Abed Hussein, R., Mohammed, A., Hassan, A. Y., Hashim, A., Najeeb, S. S., Ali, Y. M., Abdullah, A.S., & Afif, N. S. (2023). Detecting measurement disturbance: Graphical illustrations of item characteristic curves. *International Journal of Language Testing, Special Issue,* 126-133. DOI: 10.22034/IJLT.2023.391731.1247