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

Distractor Analysis in Multiple-Choice Items Using the Rasch Model

Omarov Nazarbek Bakytbekovich[1], Aisha Mohammed[2], Ammar Muhi Khleel Alghurabi[3], Hajir Mahmood Ibrahim Alallo[4], Yusra Mohammed Ali[5], Aalaa Yaseen Hassan[6], Lyazat Demeuova[7], Shvedova Irina Viktorovna[8], Bekenova Nazym[9], Al Khateeb Nashaat Sultan Afif[8]

Abstract

The Multiple-choice (MC) item format is commonly used in educational assessments due to its economy and effectiveness across a variety of content domains. However, numerous studies have examined the quality of MC items in high-stakes and higher-education assessments and found many flawed items, especially in terms of distractors. These faulty items lead to misleading insights about the performance of students and the final decisions. The analysis of distractors is typically conducted in educational assessments with multiple-choice items to ensure high-quality items are used as the basis of inference. Item response theory (IRT) and Rasch models have received little attention for analyzing distractors. For that reason, the purpose of the present study was to apply the Rasch model, to a grammar test to analyze items' distractors of the test. To achieve this, the current study investigated the quality of 10 instructor-written MC grammar items used in an undergraduate final exam, using the items responses of 310 English as a foreign language (EFL) students who had taken part in an advanced grammar course. The results showed an acceptable fit to the Rasch model and high reliability. Malfunctioning distractors were identified.

Keywords: Distractor analysis, Item response theory, Multiple-choice items, Rasch model

1. Introduction

Multiple-choice (MC) items are commonly used in educational assessments due to their capacity for measuring many knowledge, competencies, and skills in an objective and efficient manner (Gierl et al., 2017). MC items are highly reliable, administered easily, and scored objectively (Rodriguez, 2016). They are also used for formative and diagnostic purposes and are able to assess a wide range of knowledge and course materials. Such advantages have made MC items appropriate for a variety of purposes ranging from high-stakes exams to classroom achievement testing, despite some potential limitations including guessing and unintentionally

[1] Non-profit Joint Stock Company Semey Medical University, Department of Hospital Surgery. Republic of the Kazakhstan, 071400, Semey City, Abai Street, 103. ORCID: 0000-0003-3262-1410
[2] College of Education, Al-Farahidi University, Baghdad, Iraq
[3] College of Education, The Islamic University in Najaf, Iraq
[4] English Department, Ahl-Al-Bayt University, Kerbala, Iraq
[5] Department of Medical Laboratory Technics, Al-Zahrawi University College, Karbala, Iraq
[6] Al-Nisour University College, Baghdad, Iraq
[7] Institute of Natural Sciences and Geography, Abai Kazakh National Pedagogical University, Almaty, Kazakhstan, ORCID 0000-0003-0313-1391
[8] People's Friendship University of Russia, Moscow, Russia
[9] Candidate of Biological Sciences, Docent, Institute of Natural Sciences and Geography, Abai Kazakh National Pedagogical University, Almaty, Kazakhstan. ORCID: 0000-0001-5586-6125

exposing test takers to incorrect information. Consequently, because of these advantages, enhancing the quality of MC items is of paramount importance.

MC items contain the stem, alternative responses, and further information (e.g., figures, passages, and tables) which are essential for responding to the given item. The stem consists of a question that test takers must respond to, whereas the response options contain several alternative responses with a correct option, keyed option, and one or more distractors, known as plausible but incorrect options.

Distractors are used to discriminate between low and high-performing test takers, that is, test takers who have acquired the required knowledge to respond correctly to the item from those who have not mastered the content. Thus, the distractors of MC items should include a set of sensible but wrong options on the basis of test takers' common misconceptions which allows for measuring the learning status of students in a specific content domain (Rashidi, & Safari, 2014; Shin et al., 2019). Numerous researchers have suggested guidelines for developing MC items and distractors (Haladyna & Rodriguez, 2013; Haladyna et al., 2002; Huntley & Welch, 1993; McLeod et al., 2003; Moreno et al., 2015; Mosier & Price, 1945).

Because the use of implausible distractors leads to misleading insights about the performance of students and contaminating critical decisions, distractor analysis should be implemented to ensure that all answer options perform well. Distractor analysis is the process of examining the functioning of wrong answers against the correct answer for MC items on a test. As argued by Wolf and Smith (2007, p. 209), distractor analysis indicates to what extent the responses to the distractors are in concord with the expected cognitive operations upon which the distractors were constructed. Assessing the quality of MC items can be typically conducted in two ways. The first way is to use professional judgment processes such as content guidelines, style and format, writing the stem, and writing options (Haladyna & Rodriguez, 2013). The second way is to utilize statistical methods to identify item properties and utilize the statistics to specify whether an item can be appropriately inserted into the test for assessing the performance of test takers (Downing, 2006).

Over the past few decades, a great deal of research has been conducted to analyze the effect of distractors (e.g., Baghaei & Dourakhshan, 2016; Baghaei & Amrahi, 2011a; Hohensin & Baghaei, 2017; Brown & Abdulnabi, 2017; Shin et al., 2019a, 2019b, to name a few; See Lions et al., 2022, for a comprehensive review). Although previous studies have provided valuable insight into the analysis of response options, especially distractors, in MC items, there has been little research on the use of item response theory (IRT) and the Rasch model (Rasch, 1960) for assessing the quality of response options, especially for evaluating instructor-constructed MC tests. To fill this gap, this study aims to use the Rasch model to examine the quality of response options of an instructor-constructed MC grammar test. Rasch model is a probabilistic model used to make a prediction about the outcome of encounters between persons and a set of items (Aryadoust et al., 2021; Baghaei & Amrahi, 2011b). In the Rasch model, the probability of getting an item right is a function of person ability and item difficulty. More particularly, the Rasch model makes a logistic function of the discrepancy between item difficulty and person ability. A person with a higher ability has more probability to give a correct answer to the item. A distinctive feature of the Rasch model is that it creates a ruler-like device, as an interval scale with logits units (or log-odds units), to plot items and persons

on the same latent trait continuum on which the position of persons and items indicates the ability and difficulty measures, respectively.

2. Method

2.1 Participants

Participants in this study were 310 English as a foreign language (EFL) undergraduate students at Ahl-Al-Bayt University, Kerbala, Iraq. There were 199 (64.2%) male and 111 (35.8%) female students.

2.2 Instruments

To assess the grammar ability of the students, the researchers constructed a four-option MC test which includes 10 items. Students had to fill in the blanks by choosing the correct word or phrase among the four available options. The test was administered as part of a final exam in a grammar course in eight parallel classes. The score of the test ranged from 2 to 9, with a mean of 2.57 (*SD* = 1.709). Using Cronbach alpha, the reliability coefficients of the test was calculated, and a value of 0.42 was obtained, indicating the low reliability for the scale.

3. Results

3.1. Item Characteristics

For this study, we used WINSTEPS computer package, Version 3.73 (Linacre, 2009) to fit the test data to the Rasch model. Table 1 provides Rasch measurement results, including item difficulty measures, their standard error of measurement, fit statistics, and point-measure correlation. The second column gives the difficulty of the test items, indicating the location of items on the latent trait continuum and are explained in logits (or log odd-units). Item analysis produced an item difficulty varied from -1.23 to 1.17 logits, with separation and item reliability coefficients of 4.26 and 0.95, respectively. Person estimates also varied from -3.69 to 2.37, with separation and person reliability coefficients of 0.51 and 0.21, respectively. Person and item reliability coefficients indicate the precision of the test in measuring person performance and the difficulty of the items (Linacre, 2009). Person reliability with higher values shows that the test differentiates well among students with different ability levels (Bond & Fox, 2015). Separation is the ratio of true variance to error variance and shows the number of students' classifications and the items' hierarchy (Linacre, 2009). Column "Standard error" presents the extent to which item difficulty measures were
precisely estimated.

Columns three and four demonstrate the results of infit and outfit mean-square (MNSQ) for assessing the quality of the items. As argued by Linacre (2002), outfit MNSQ is sensitive to outliers, whereas infit MNSQ is sensitive to abnormal behavior of items close to persons' measures. Values within the range of 0.5 to 1.50 are considered ideal (Linacre, 2009). In this study, all the test items have fit values within the acceptable range, suggesting that items can represent the intended underlying latent trait very well.

The last column shows point-measure correlations for all the test items. Point-measure is similar to item-total correlation or item discrimination in classical test theory and indicates the extent to which observed scores agree with the expected latent trait. As can be seen, all values

are positive, suggesting that the patterns of the difficulty of test items conform to the Rasch model (Linacre, 2009).

Table 1.
Item Difficulty Measures, Fit Indices, and Point-measure Correlation

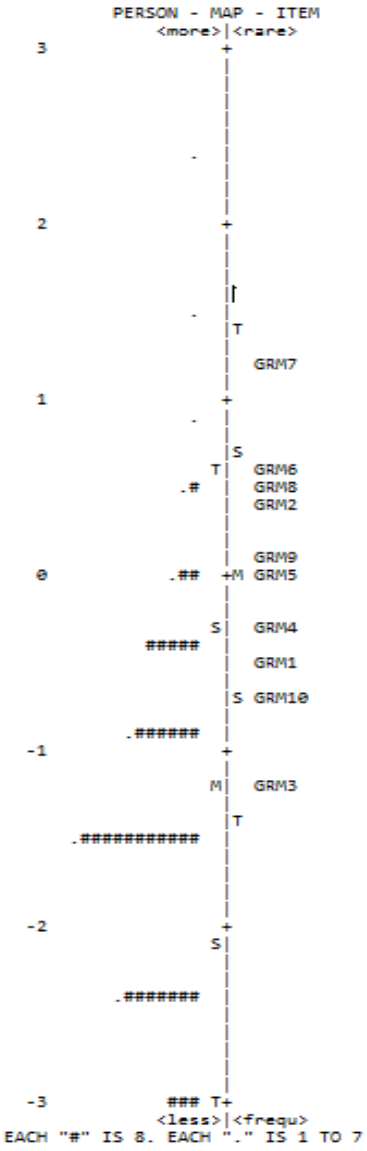
Items	Item Difficulty	Standard Error	Infit MNSQ	Outfit MNSQ	Point-measure Correlation
1	-0.52	0.13	0.92	0.94	0.48
2	0.36	0.16	0.91	0.84	0.45
3	-1.23	0.13	1.04	1.04	0.43
4	-0.34	0.14	1.11	1.09	0.34
5	0.02	0.15	0.84	0.79	0.52
6	0.62	0.17	1.07	1.18	0.28
7	1.17	0.20	1.15	1.22	0.18
8	0.53	0.16	0.99	0.91	0.37
9	0.15	0.15	0.98	1.02	0.39
10	-0.73	0.13	1.02	1.02	0.42
Mean	0.00	0.15	1.00	1.01	-
SD	0.68	0.02	0.09	0.13	-

Note. MNSQ = Mean-square; SD = Standard Deviation

Figure 1 depicts the positions of items and persons on the same scale, referred to as the Wrigh map. This indicates a unique feature of the Rasch model which allows to express person ability estimates and item difficulty estimates on the scale which is expressed in logits. The expected latent trait is represented by a dotted line. Persons and items on top of the metric are students with higher abilities and more difficult items, and those at the bottom of the metric are students with lower abilities and easier items. The map demonstrates that items cover a relatively wide range of difficulty, from -1.23 logits (Item 3; *SEM* = 0.13) to 1.17 logits (Item 7; *SEM* = 0.20). Person ability measures vary from -3.69 to 2.37. Overall, the Wright map shows that the level of the test items' difficulties was higher than the ability of level of the students. Therefore, a number of easier items should be added to the test to cover the ability levels of all students.

Figure 1.

Wright Map of the Distribution of Items and Persons on the Latent Trait



3.2. Checking Unidimensionality

We also checked the unidimensionality of the test. Unidimensionality is an important assumption in IRT and Rasch models. This indicates that all items of a test should measure only a single latent trait. To examine whether the test is unidimensional, we used principal component analysis of linearized Rasch residuals (PCAR). Residuals are considered as differences between predictions of the Rasch model and the observed data and are part of the data the model has not accounted for. Residuals are expected to be randomly distributed and uncorrelated (Linacre, 2009). The lower the values of the residuals, the better the data fit the Rasch model. The analysis of PCAR revealed that the model explains 19.9% variance of the actual data; person measures explain 6.6%, and item measures explain 13.3%. Although the observed model variance agrees with the expectations of the model (19.7%), 80.1% of the variance isnot explained. The first component accounts for 11.6% of the unexplained variance

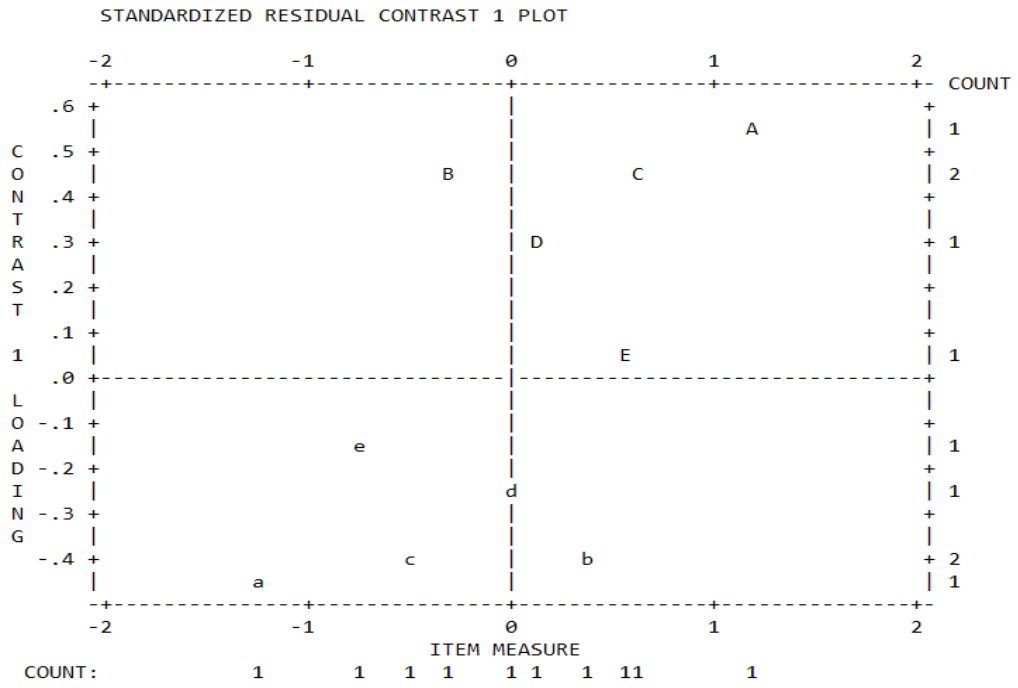
with the eigenvalue equal to 1.4, which is smaller than 2. As Figure 2 shows, the residuals of the test items do not form identifiable patterns. Items scatter across the map, indicating the unidimensionality of the test (Linacre, 2009).

3.3. Checking Response Options

Furthermore, Table 3 summarizes response option statistics for the data. Column one indicates the item number. The second and third columns demonstrate the data code (response options) and answer key for each item, respectively. Column three shows the frequency and percentage of each option. As an illustration, Option 4 in Item 1 has the largest proportion of responses, suggesting that a large number of students have selected this option and have given a correct answer to the item. The fourth and fifth columns represent the average measure of students' ability, who have selected each option, and the standard error of the mean measure of the sample of students who selected each option, respectively.

Figure 2.

Representation of the Standardized Residual Contrast in the Grammar Test



In relation to average ability, we expect that correct answers should have higher average abilities. In other words, the average mean for the correct option should be constantly greater than other options, especially distractors. The values in the table support this expectation. Column seven shows the mean of the outfit mean-squares related to response options. As can be seen, except for Option 3 in Item 6, Option 3 in Item 7, and Option 2 in Item 8 (represented by an asterisk * in the table), all values are within the acceptable range of 0.50 to 1.50 (Linacre, 2009). The last column finally indicates the point-measure correlation between options and students' total Rasch person measures on the test. Correlations should be positive for the correct

options and negative for the distractors to indicate that students who select the wrong options have lower measures compared to those students who select the correct options. Except for Distractor 3 in Item 6, Distractor 1 in Item 7, Distractor 2 in Item 8, and Distractor 1 in Item 9 (represented by asterisks* in the table), all distractors of the test items have negative values.

Table 3.

Summary of Response Option Statistics for the Grammar Data

Items	Data Code (Response Options)	Correct Answer	Count (%)	Average Ability	S.E. Mean	Outfit MNSQ	Point-measure Correlation
1	1	0	69 (22%)	-1.86	0.12	0.8	-0.24
	3	0	80 (26%)	-1.74	0.11	0.9	-0.20
	2	0	22 (7%)	-1.74	0.19	0.8	-0.09
	0	0	36 (12%)	-1.54	0.16	1.0	-0.06
	4	1	103 (33%)	-0.62	0.10	1.0	0.48
2	3	0	91 (29%)	-1.73	0.11	0.9	-0.21
	0	0	44 (14%)	-1.65	0.15	0.9	-0.10
	4	0	16 (5%)	-1.53	0.34	1.4	-0.03
	2	0	99 (32%)	-1.50	0.09	0.9	-0.08
	1	1	60 (19%)	-0.37	0.12	0.8	0.45
3	1	0	40 (13%)	-1.97	0.17	0.9	-0.21
	0	0	66 (21%)	-1.88	0.13	1.0	-0.24
	2	0	28 (9%)	-1.66	0.20	1.1	-0.08
	4	0	31 (10%)	-1.60	0.18	1.1	-0.07
	3	1	145 (47%)	-0.87	0.08	1.0	0.43
4	4	0	35 (11%)	-2.29	0.17	0.6	-0.30
	3	0	36 (12%)	-1.56	0.20	1.4	-0.06
	0	0	65 (21%)	-1.52	0.13	1.2	-0.07
	1	0	81 (26%)	-1.42	0.11	1.2	-0.03
	2	1	93 (30%)	-0.81	0.09	1.1	0.34
5	2	0	96 (31%)	-1.78	0.09	0.7	-0.25
	3	0	54 (17%)	-1.64	0.14	0.9	-0.11
	0	0	33 (11%)	-1.64	0.19	1.1	-0.08
	4	0	52 (17%)	-1.62	0.14	0.9	-0.10
	1	1	75 (24%)	-0.36	0.10	0.8	0.52
6	1	0	56 (18%)	-1.87	0.14	0.7	-0.21
	2	0	23 (7%)	-1.70	0.15	0.6	-0.08
	0	0	116 (37%)	-1.68	0.09	0.8	-0.22
	3	0	65 (21%)	-0.82	0.14	1.9*	0.26*
	4	1	50 (16%)	-0.67	0.15	1.2	0.28
7	2	0	17 (5%)	-1.86	0.24	0.7	-0.11
	4	0	37 (12%)	-1.70	0.19	0.9	-0.11

	0	0	149 (48%)	-1.64	0.08	0.8	-0.24
	1	0	74 (24%)	-0.81	0.12	1.8*	0.29*
	3	1	33 (11%)	-0.79	0.16	1.2	0.18
	0	0	127 (41%)	-1.76	0.09	0.8	-0.30
	3	0	29 (9%)	-1.60	0.21	1.6*	-0.07
8	4	0	14 (5%)	-1.57	0.31	1.0	-0.04
	2	0	87 (28%)	-1.25	0.11	1.2	0.07*
	1	1	53 (17%)	-0.47	0.12	0.9	0.37
	4	0	18 (6%)	-1.92	0.20	0.6	-0.12
	3	0	16 (5%)	-1.89	0.20	0.6	-0.11
9	0	0	149 (48%)	-1.70	0.08	0.8	-0.29
	1	0	58 (19%)	-1.16	0.13	1.3	0.09*
	2	1	69 (22%)	-0.57	0.13	1.0	0.39
	1	0	20 (6%)	-2.04	0.18	0.6	-0.16
	0	0	42 (14%)	-1.86	0.18	1.1	-0.17
10	2	0	47 (15%)	-1.76	0.14	0.9	-0.15
	3	0	86 (28%)	-1.57	0.11	1.1	-0.11
	4	1	115 (37%)	-0.77	0.09	1.0	0.42

Note. S.E. = Standard Error; MNSQ = Mean-square

4. Conclusion

This study set out to evaluate the quality of an instructor-constructed MC grammar test. The Rasch model was utilized to check data-model fit and analyze distractors. The results of fit indices, item difficulty measures, and point-measure correlations showed that the structure of the observed data conforms to the predictions of the Rasch model. The result of the reliability coefficient for items was higher than the acceptable value of 3, and the separation value was greater than 2, suggesting the representativeness of the test items. However, person reliability coefficients and separation values were low, indicating that the test could not accurately measure person performance and identify students' classifications. This could be attributed to the very low standard deviation of scores (*SD* = 1.709), that is, data points were mainly below the mean. The unidimensionality of the test was also investigated using PCAR. It turned out that the eigenvalue of the first contrast was lower than 2, indicating the unidimensionality of the test.

Finally, the analysis of response options showed that all correct responses had greater average abilities, that is, the average mean for distractors was lower than the correct options. The values of Outfit MNSQ revealed that apart from the options of three items (e.g., Option 3 in Item 6, Option 3 in Item 7, and Option 2 in Item 80, all values were within the ideal range of 0.50 to 1.50. Except for some distractors, the values of point-measure correlations were also negative.

An important finding of the study is that the data had an adequate fit to the Rasch model. This diverges from previous studies (e.g., Brown & Abdulnabi, 2017; Shin et al., 2019) which indicated that the structure of the Rasch model is not appropriate for evaluating MC items due to its strict assumptions. Because MC items have the possibility of guessing, it is rational to

use models which can identify the effect of the chance performance. Three-parameter logistic IRT (3-PL IRT; Lord, 1980) model has the potential to estimate pseudo-guessing parameter. The present study, however, has demonstrated that the Rasch model can detect items and response options with low discrimination.

The analysis of distractors allows instructors and test developers to grasp why students generate wrong answers which results in diagnostic inferences about test performance. Instructors can also detect the content areas that require instructional improvement and give feedback to students for further remedial instruction (Arefsadr et al., 2022).

One limitation of this study is that a small sample size (N = 310) was used for analyzing the test. Future studies can use large sample sizes to investigate whether the results of this study, especially the good fit of the data to the Rasch model, can be replicated. Besides, linear logistic test modeling (Baghaei & Kubinger, 2015; Fischer, 1973) can be used to examine if decision processing in option selection imposes any construct-irrelevant cognitive load on examinees (Embretson & Wetzel, 1987).

References

Arefsadr, S., Babaii, E., & Hashemi, M. R. (2022). Why IELTS candidates score low in writing: Investigating the effects of test design and scoring criteria on test-takers' grades in IELTS and world Englishes essay writing tests. *International Journal of Language Testing*, *12*(2), 145-159. https://doi.org/10.22034/ijlt.2022.157131

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Baghaei, P., & Dourakhshan, A. (2016). Properties of single-response and double-response multiple-choice grammar items. *International Journal of Language Testing*, *6*(1), 33 49. URL: https://www.ijlt.ir/article_114425.html

Baghaei, P., & Amrahi, N. (2011a). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53*, 192-211.

Baghaei, P., & Amrahi, N. (2011b). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2*, 1052-1060.

Baghaei, P. & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20*, 1-11.Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd Ed.)*. Routledge.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. H. Downing & M. Thomas (Eds.), *Handbook of test development* (pp. 3-25). Lawrence Erlbaum Associates.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*, 175 193. https://doi.org/10.1177/014662168701100207

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374. doi: 10.1016/0001-6918(73)90003-6

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive Review. *Review of Educational Research, 87*(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple choice tests matter? *Psicologica, 38*, 93-109.

Huntley, R. M., & Welch, C. J. (1993, April). *Numerical answer options: Logical or random order?* Paper presented at the annual of meeting of the American Educational Research Association, Atlanta, GA.

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

Lions, S., Monsalve, C., Dartnell, D., Paz Blanco, M., Ortega, G., & Lemarié, J. (2022). Does the response options placement provide clues to the correct answers in multiple-choice tests? A systematic review. *Applied Measurement in Education, 35*(2), 133-152. https://doi.org/10.1080/08957347.2022.2067539

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

McLeod, I., Zhang, Y., & Yu, H. (2003). Multiple-choice randomization. *Journal of Statistics Education, 11*(1), -7. Retrieved from http://ww2.amstat.org/publications/jse/v11n1/mcleod.html

Moreno, R., Martínez, R. J., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*(4), 388-394. https://doi.org/10.7334/psicothema2015.110

Mosier, C. I., & Price, H. G. (1945). The arrangement of choices in multiple choice questions and a scheme for randomizing choice. *Educational and Psychological Measurement, 5*(4), 379-382. https://doi.org/10.1177/001316444500500405

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. Danish Institute for Educational Research.

Rashidi, N., & Safari, F. (2014). Does the type of multiple-choice item make a difference? The case of testing grammar. *International Journal of Language Testing, 4*(2), 175-186. URL: https://www.ijlt.ir/article_114398.html

Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. Raymond, and T. Haladyna (Eds.), *Handbook of test development* (2nd Ed., pp. 259-273). Routledge.

Shin, J., Bulut, O., & Gierl, M. J. (2019a). The effect of the most-attractive-distractor location on multiple-choice item difficulty. *The Journal of Experimental Education, 88*(4), 643-659. https://doi.org/10.1080/00220973.2019.1629577

Wolf, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement, 8*(2), 204-234. URL: http://jampress.org/abst2007.htm