Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 11, No. 1, March 2021

ISSN 2476-5880

A Mokken Scale Analysis of an English Reading Comprehension Test

Mona Tabatabaee-Yazdi[1]*, Khalil Motallebzadeh[1], Purya Baghaei[2]

Received: 10 December 2020 Accepted: 27 February 2021

Abstract
Reading comprehension in English, as one of the most central skills, has a vital role in the process of learning English as a Foreign/Second Language. The current study used the Mokken Scale Analysis (MSA), a probabilistic-nonparametric approach to item response theory (IRT), to determine the unidimensionality and scalability of a 20-item reading comprehension test administered on 300 EFL university students in the Iranian context. The results showed no major concerns in terms of item scalability. Monoton Homogeneity Model (MHM) fitted all the items of the test very well as measured by the scalability coefficients and restscore groups method. Considering the IIO, it was concluded that the ordering of items according to their mean is invariant across examinees although H^T was small. Dimensionality analysis results using the AISP showed that the test is unidimensional providing evidence of the validity of the test in measuring a single ability dimension.

Keywords: Double monotonicity Model; Nonparametric Item-response theory; Monoton Homogeneity Model; Mokken-scale analysis; Reading comprehension

1. Introduction

1.1. Reading Comprehension

Tests, surveys, and questionnaires as the main educational research instruments are frequently used throughout education for the learners' assessments of "courses, programs, and clerkships as well as for student self-assessments and patient satisfaction" (Palmgren, Brodin, Nilsson, Watson, & Stenfors, 2018, p. 2). A well-prepared test or questionnaire is considered to be a suitable tool for

[1] *English Department, Tabaran Institute of Higher Education, Mashhad, Iran. tabatabaee.mona@gmail.com
[2] English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran.

 ISSN 2476-5880

the measurement of examinees' latent traits or abilities. Measurement and assessment of different variables are ubiquitous in education and psychology, and other social sciences. It is because measurement allows scientists to describe individuals' differences in terms of the phenomenon in question (Bryman, 2004).

One of the most important tests used in educational settings throughout the world is tests of reading comprehension ability. Reading comprehension in English, defined as the skill to read, comprehend, process, and recall what has just been read (Ganie & Rangkuti, 2019; Roohani Tonekaboni, Ravand, Rezvani, 2021), has a critical role in the process of learning English as a Foreign/Second Language (Motallebzadeh, & Tabatabaee-Yazdi, 2016). Reading comprehension is also the foundation of all phases of language learning instruction (Mikulecky, 2008). Researchers (Ganie & Rangkuti, 2019; Osikomaiya, 2021; Li et al., 2021) stated that if the reading skill is mastered, transfer of knowledge to other school subjects is facilitated and this helps learners to not only tackle the difficulties of the English language comprehension but also of other school subjects (Baghaei, & Ravand, 2019; Tabatabaee-Yazd, & Baghaei, 2018).

In contrast with the first language (L1) readers, second language (L2) students begin learning to read with no preliminary language bases that are present for L1 readers. First language readers often start reading with a simple implied knowledge of grammatical structures and vocabulary, and different cognitive abilities (Flanagan & Ortiz, 2001; McGrew & Flanagan, 1998; Woodcock, McGrew, & Mather, 2001). Accordingly, one of the main psychological difficulties of reading comprehension is that learners must not only read to deduce the meaning of each word they also should link the meaning of words into the whole perception of the coherent sentence (Jamilah, 2021). Therefore, a deeper thinking process is needed about the words that may just have read or seen by the readers on the text for the first time (Setiyadi, Kuswendi, & Ristiyana, 2019).

According to Hemati and Baghaei (2020), English reading comprehension ability as one of the main language skills has been the focus of much research, and scholars attempt to shed light on the various aspects of this complex language skill. Consequently, this study aimed to provide a pragmatic approach to explore an English reading comprehension test from IRT perspectives, using Mokken Scale Analysis (MSA; Mokken, 1971).

1.2. Mokken Scale Analysis

In educational and psychological measurement, raw total scores are used to order the testees from the most to the least able, or from the most anxious to the least anxious. Considering the underlying assumptions of item response theory (IRT), raw scores are ordinal scale data meaning that only the order of the respondents can be determined, not their differences in strengths or weaknesses. However, to consider raw scores as ordinal the response patterns have to follow an axiom named *transitivity* (Sijtsma & Molenaar, 2002, as cited in Baghaei, in press) which means that if a testee answers a more difficult question correctly, then an easier item should have been answered correctly, too.

 ISSN 2476-5880

The non-parametric item response theory (NIRT) models that are collectively known as Mokken Scale Analysis (MSA; Mokken, 1971) aim to assess dichotomous (binary) or polytomous (ordinal) unidimensional scales. Comparing with parametric IRT models, non-parametric models often concentrate on less restrictive assumptions about the data by focusing on in-depth model fit examination and data investigation to comprehend the test, the items, and the respondents (Junker & Sijtsma, 2001). Therefore, MSA can be applied when designing or constructing multi-item questionnaires. MSA provides several statistical tools, including a psychometric method of data reduction, for constructing scales to measure persons and items regarding personality and cognitive attributes; and also for discovering the interrelationship between items and latent traits. These methods test whether the data meet the axiom of an ordinal scale. If NIRT models fit the data, it is assumed that the testees can be ordered using their total scores (Baghaei, in press). Thus, a higher item score indicates a higher attribute level. Many scholars (e.g., Meijer & Baneke, 2004; Meijer, Tendeiro, & Wanders, 2015; Wind, 2017) asserted that NIRTcan be considered as an efficient approach in settings that the underlying response processes are not comprehensible well, such as emotional variables.

MSA evolved from the Guttman scaling model, believing that the items in a scale are hierarchically ordered meaning that the items of a test are ordered by their degree of difficulty. Therefore, the Guttman scaling model is deterministic since it does not allow for any random errors (Palmgren et al., 2018). MSA can be used in both confirmatory (for a number of items supposed to form a scale) and in an exploratory approach (for the analysis of a number of items to discover whether the items set up one or more scales). Both approaches use the same criteria grounded in two models known as *Monotone Homogeneity Model (MHM)* and the *Double Monotonicity Model (DMM)* **(**Mokken, 1971). Both models show a person ordering but only the DMM implies an invariant item ordering (Sijtsma, & van der Ark, 2017).

1.2.1. Dichotomous Response Mokken Scale Analysis

1.2.1.1.Monotone Homogeneity Model (MHM).

Under MHM three common basic assumptions of all IRT models should be checked. *Unidimensionality* which means the test should measure just one single trait; *local independence* that means the items should not depend on each other after conditioning out the latent variable, meaning that answers to an item should not be influenced by answers to other items; and *monotonicity* which accounts for the fact that there should be an increasing or at least a constant relationship between the latent trait and the probability of a correct response (Wind, 2017).

The MHM does not limit the shape of the IRFs; as long as the IRFs are non-decreasing, they fit the MHM. Thus, IRFs in MHM may take a shape that does not essentially match with parametric IRT models. If the MHM holds, respondents can be ordered with their raw total scores. Moreover, when data fit the MHM assumptions, the learners' ordering on the latent trait is invariant across subsets of items. Thus, the model can be similar to the two-parameter logistic model in parametric IRT (Wind, 2017).

 ISSN 2476-5880

1.2.1.2.Double Monotonicity Model (DMM)

Mokken's (1971) DMM, besides considering the three assumptions of MHM, requires another assumption known as *invariant item ordering (IIO)* or *non-intersecting Item Response Functions (IRF)* which means that individual items response functions should not intersect with any other item response functions, therefore, if this assumption holds, the items are ordered the same for all examinees (Wind, 2017) which provide evidence for invariant ordering of both items and learners.

The IIO property is crucial in creating hierarchical scales. According to Palmgren et al. (2018. p.5), "if these four assumptions are not excessively violated, higher sum scores are seen as corresponding to higher values on the latent trait, suggesting that respondents can be reliably ordered on the latent trait by their sum scores".

1.2.2. Polytomous Response Mokken Scale Analysis

To work on rating scales data, MHM and DMM formulation of polytomous data was proposed by Molenaar in (1982, 1997). Although these models work under the same underlying assumptions as Mokken's (1971) dichotomous formulations, the polytomous model assumptions are assessed for each item both at the overall item level and within rating scale categories (Wind, 2017). Therefore, for polytomous ratings, *monotonicity* means when levels of student progress within a rating scale item or for a particular rater increase, the cumulative likelihood for a rating in or above each rating scale category should also be nondecreasing. *Unidimensionality* also means that just one single trait should be measured by the scale. Like dichotomous models, *local independence* for polytomous models means that after taking into account learners' ability, learners' responses to the items should be statistically independent. *IIO* for polytomous rating scales is proposed for assessing whether item response functions intersect.

2. Method

2.1.Participants and Setting

Three hundred BA students (65.7% Females, 34.3% Male) majoring in English as a foreign language (EFL) from different universities in Iran were recruited for the study. The test was administered in one online session during normal class periods in the academic year of 2020. All the academic classes in 2020 were held online due to the COVID-19 pandemic, thus, the test was prepared and administered using Google form. All participants were native speakers of Farsi**.**

2.2.Instrumentation

Participants' reading comprehension was measured using a reading comprehension test consisting of 20 four-option multiple-choice items that were selected from the reading comprehension section of one of the official past, 2011, Iranian University Entrance Examination (IUEE). The test contained three passages with lengths of 428, 466, and 504 words. The Cronbach's alpha reliability of the reading test was 0.94. The mean and standard deviation of the sample on the test were 6 and 6.23, respectively.

 ISSN 2476-5880

3. Results

3.1. Scalability assessment and monotonicity

The *R* package Mokken (V 3.0.6, van der Ark, 2021) was run to analyze the data. The Mokken scalability coefficients including item-pair scalability (H_{ij}), item scalability (H_j), and total scalability (H) were examined. According to Mokken (1971), a scale is considered weak if $.30 \leq H < .40$, a medium scale if $0.40 \leq H < 0.50$, and strong if $H \geq 0.50$; the values of H_{ij} must be greater than zero or non-negative which indicates a non-negative relation between an item and the latent trait. Thus, all Hij coefficients (accordingly, all pairwise interactions) must be positively related, and items must be properly homogeneous with other items. These assumptions can lead to the development of instruments that coincide to more accurate values of reliability and homogeneity than tools related to conventional classical test theory (CTT) reliability analysis (Van Schuur, 2003). Moreover, H_j values have to be checked and items should be omitted if their coefficient $H_j < 0.30$. Items make a Mokken scale if $H_j \geq 0.30$. If H_j values are smaller than .30, they do not fit the MHM well. In this case, it is possible to revise the test, and reanalyze the data. However, if it is not possible to revise the test, systematic removal or replacement of the misfitting items should be conducted.

Table 1.
Discrete item scalability and monotonicity

Items	Hj	Monotonicity			
		Active comparison (ac)	Violations (vi)	Significant violations (zsig)	crit
1	.58	6	0	0	0
2	.60	6	0	0	0
3	.52	6	0	0	0
4	.61	3	0	0	0
5	.59	6	0	0	0
6	.61	3	0	0	0
7	.56	3	0	0	0
8	.59	6	0	0	0
9	.60	3	0	0	0
10	.63	3	0	0	0
11	.54	6	0	0	0
12	.62	3	0	0	0
13	.55	6	0	0	0
14	.73	1	0	0	0
15	.57	6	0	0	0
16	.55	3	0	0	0
17	.64	6	0	0	0
18	.60	3	0	0	0
19	.62	3	0	0	0
20	.58	3	0	0	0

*Scale H=0.591

 ISSN 2476-5880

According to Table 1, the results showed that H_{ij} and H_j were positive for all the items. Moreover, $H_j > .30$ was achieved for all the items of the reading comprehension test (Table 1). Besides, H was .591 which indicated the scale is strong. These results suggest that each of the items contributes to a meaningful overall ordering of participants in terms of reading comprehension ability. Therefore, the scale can be safely treated as unidimensional and can order the respondents using the total scale on the reading ability scale.

Table 1 shows that there is no violation of monotonicity for any of the items as illustrated by the vi column which means the ordering of participants with their total scores is warranted. Crit value is used to assess the violation of monotonicity if there is any. According to Molenaar and Sijtsma (2000), crit statistic < 40 is considered as not extremely violating monotonicity; therefore, the items can be safely incorporated in any Mokken scale, but crit values > 40 show violation of monotonicity, and the most serious items (highest violations) should be deleted before rerunning the analysis.

3.2. Non-intersection of IRFs

For assessing the IIO, the transposed Mokken scale coefficient (H^T) was studied. $H^T < .3$ shows that the item ordering is inaccurate, $0.30 \leqslant H^T < 0.40$ means low accuracy, $0.40 \leqslant H^T < 0.50$ indicates medium accuracy, and $H^T \geqslant 0.50$ suggests high accuracy (Ligtvoet, Van der Ark, te Marvelde, & Sijtsma, 2010). The Crit value (Sijtsma & Molenaar, 2002; Van Schuur, 2011) is used as a measure of the effect size of the violation of IIO. Crit < 40 shows minor violation; 40 ⩽ Crit < 80 indicates nonserious violation but the violated items should be reviewed by the researcher; and Crit ⩾ 80 signifies serious violation.

The H^T value in this study is shown to be 0.22 which indicates the item ordering is not much accurate. The values of the (zsig) column in Table 2 show that five items (items 3, 8, 11, 16, and 20) violated the IIO assumption, although according to the crit values and the backward item selection method only two items of 3 and 11 should be removed. According to (zsig) column one significant violation was observed for Item 8, 11, 16, and 20, and two significant violations were observed for item pairs involving Item 3. These violations indicate that the difficulty of these five items is not invariant across the range of the ability, meaning that they do not retain a constant level of difficulty across the ability continuum. Therefore, these items can be considered as potential candidates for removal.

ISSN 2476-5880

Table 2.
IIO and Backward step removal of items violating IIO

Items	Hj	IIO			Backward Selection			
		Active comparison (ac)	Violations (vi)	Significant violations (zsig)	Crit	Step1	Step2	Step3
17	.64	55	0	0	0	0	0	0
13	.55	57	3	0	11	0	0	0
15	.57	57	2	0	0	0	0	0
8	.59	56	3	1	36	1	1	0
11	.54	57	5	1	**45**	1	1	NA
2	.60	56	3	0	11	0	0	0
1	.58	57	2	0	2	0	0	0
20	.58	57	4	1	29	1	0	0
16	.55	57	2	1	20	1	0	0
3	.52	57	4	2	**42**	2	NA	NA
18	.60	57	2	0	11	0	0	0
5	.59	57	2	0	7	0	0	0
9	.60	57	0	0	0	0	0	0
6	.61	57	1	0	2	0	0	0
7	.56	57	3	0	12	0	0	0
4	.61	57	1	0	3	0	0	0
19	.62	57	1	0	0	0	0	0
10	.63	57	2	0	4	0	0	0
12	.62	57	2	0	4	0	0	0
14	.73	57	0	0	0	0	0	0

*H^T=0.22

Accordingly, Figure 2 shows the violation of the IIO and nonintersection of Item 3 and Item 5 and 18, and the intersection between item-pairs 11 and 2 and 20. Once the two items were removed from the scale, H^T improved to .23, and the test scalability H to .594, indicating a stronger scale. Thus, it can be concluded that the DMM fitted to the data after removing two items.

 ISSN 2476-5880

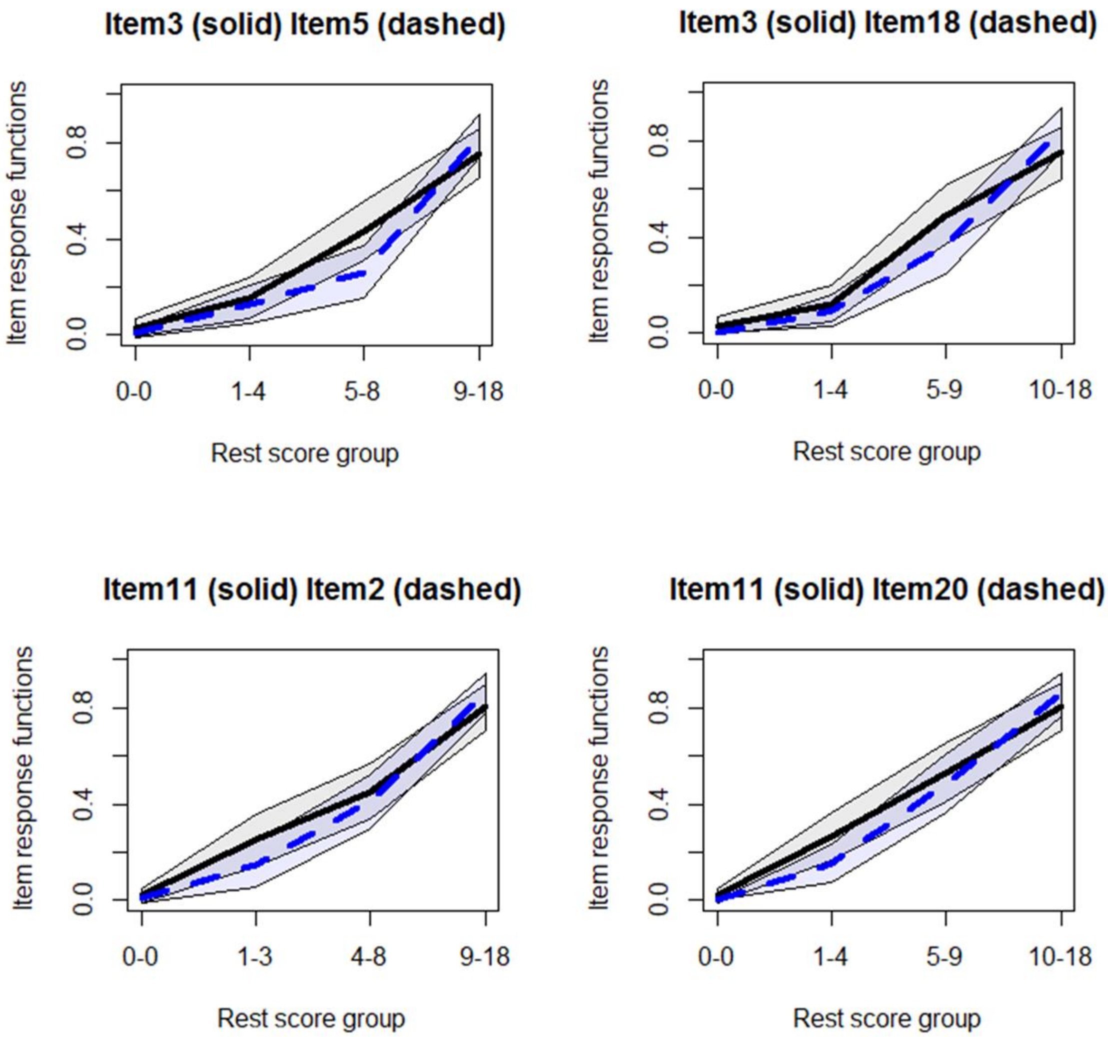

Figure 2. IRF depicting the violation of the invariant item ordering (IIO) assumption

3.3.Automated item selection procedure (AISP)

Searching for item sets to evaluate the unidimensionality of the scale is done by the Mokken package through the AISP function. Practically, AISP recognizes all the items that have ample psychometric quality to be included in the final scales by identifying and excluding non- or low-discriminating items. In addition, probable unbidden or sub-attributes different from the intended attribute may be distinguished (Sijtsma, & van der Ark, 2017). It should also be noted that "AISP does not assess monotonicity, but if an IRF shows gross violations of monotonicity, this will tend to lower the item's scalability coefficient Hj and AISP will likely not select the item in a scale" (Sijtsma, & van der Ark, 2017, p. 11). As presented in Table 3, using the AISP function with a lower bound of c=0.3 ($H_J \geq c \geq 0$) showed clustering of items around just one dimension. Thus, the scale can be safely considered as a unidimensional scale.

 ISSN 2476-5880

Table 2.
Dimensionality of items

Items	Dimensions from AISP C=0.3
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1

4. Discussion and Conclusion

The Mokken Scale Analysis is an efficient way to explore the behavior of items in scales in an attempt to order persons and items by addressing the underpinning assumptions of IRT through two nonparametric models, namely, MHM and DMM. This study applied MSA to assess the psychometric properties and to determine the unidimensionality and scalability of an English language reading comprehension test developed and administered in the Iranian context. In particular, this study used the MSA to investigate whether the measurement quality of the English language reading comprehension test is adequate to order learners and make decisions according to the measurement results.

Many researchers (Sijtsma, & van der Ark, 2017; van der Ark, 2012; Watson et al., 2012) working on MSA assert that MSA can provide scholars with a detailed analysis of the scalability and dimensionality structure of items. The results of MSA analysis to explore the English reading comprehension test showed that MHM fitted all items of the test very well as measured by the scalability coefficient which specifies that the test's items can order students in terms of their reading comprehension well on the latent trait, meaning that students with a higher level of reading

 ISSN 2476-5880

comprehension ability score higher on the test. Accordingly, it allows the researchers to interpret the participants' total scores as an indicator of student ordering on the latent variable.

However, the H^T value of 0.22 specified the fact that the test items are not accurately ordered. Besides, based on the IIO assumption two items (items 3 & 11) should be removed to attain IIO. Considering the IIO, it can be concluded that the ordering of items according to their means is invariant in subgroups (Ligtvoet et al., 2010; Sijtsma & Van der Ark, 2017). Moreover, the dimensionality results using the AISP showed that the test is unidimensional.

To sum, this study applies Mokken analysis to educational tests and tries to assess the fit of the MSA models. The analysis showed that the MHM and DMM fit the test, providing evidence of the validity of the test and its successful application.

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.

References

Abdelhamid, G. S., Gómez-Benito, J., Abdeltawwab, A. T., Abu Bakr, M. H., & Kazem, A. M. (2020). A Demonstration of Mokken Scale Analysis Methods Applied to Cognitive Test Validation Using the Egyptian WAIS-IV. *Journal of Psychoeducational Assessment*, *38*(4), 493-506.

Baghaei, P. (in press). *Mokken Scale Analysis in language assessment*.

Baghaei, P. (2020). *Elements of psychometrics*. Mashhad, Iran: Jaliz Publication.

Baghaei, P., & Ravand, H. (2019). Method bias in cloze tests as reading comprehension measures. *SAGE Open*, *9*(1), 2158244019832706.

Bryman, A. (2004). *Social research methods.* New York: Oxford University.

Flanagan, D. P., & Ortiz, S. 0. (2001). *Essentials of crossbattery assessment.* New York: John Wiley & Sons.

Ganie, R., D., & Rangkuti, R. (2019). Reading comprehension problems on English texts faced by high school students in Medan. KnE *Social Sciences,* 2019, 684–694. https://doi.org/10.18502/kss.v3i19.4896

Hall, C., Vaughn, S., Barnes, M. A., Stewart, A. A., Austin, C. R., & Roberts, G. (2019). *The Effects of Inference Instruction on the Reading Comprehension of English Learners with Reading Comprehension Difficulties. Remedial and Special Education, 074193251882498.* doi:10.1177/0741932518824983

Hemati, S., Baghaei, P. (2020). A cognitive diagnostic modeling analysis of the english reading comprehension section of the Iranian national university entrance examination. *International Journal of Language Testing*, *10*(1), 11-32.

ISSN 2476-5880

Jamilah, N. (2021). The effectiveness of using metacognitive strategy to teach online reading comprehension for English Young Learners. *Journal of Research on English and Language Learning (J-REaLL)*, *2*(1), 209-215.

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211–220. doi:10.1177/01466210122032028

Li, M., Geva, E., D'Angelo, N., Koh, P. W., Chen, X., & Gottardo, A. (2021). Exploring sources of poor reading comprehension in English language learners. *Annals of Dyslexia*, *2021*, 1-23. https://doi.org/10.1007/s11881-021-00214-4

Ligtvoet, R., Van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*, 578-595. doi:10.1177/0013164409355697

McGrew, K. & Flanagan, D. (1998). *The Intelligence Test Desk Reference: Gf-Gc cross-battery assessment*. Allyn & Bacon.

Mikulecky, P. S. (2008). Teaching reading in a second language. *Pearson Education*, 0-13-503146-X.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*, 354–368. https://doi.org/10.1037/1082-989X.9.3.354

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. Revick (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 85–110). New York, NY: Routledge.

Mokken, R. J. (1971). A theory and procedure of scale analysis. The Hague, The Netherlands: Mouton.

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden*, *3*(8), 145–164. Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitative Methoden*, *37*(12), 97–117.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York, NY: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows: A program for Mokken scale analysis for polytomous items (Version 5.0)* [Software manual]. Groningen, Netherlands: University of Groningen, ProGAMMA.

Motallebzadeh, K., & Tabatabaee-Yazdi, M. (2016). The relationship between EFL learners' reading comprehension ability and their fluid intelligence, crystallized intelligence, and processing speed. *Cogent Education*, *3*(1), 1228733.

Osikomaiya, O. M. (2021). Secondary school students' perception to reading comprehension in English language in Ijebu Ode local government of Ogun State. *Dutsinma Journal of English and Literature*, *2*(1), 273-289.

Palmgren, P. J., Brodin, U., Nilsson, G. H., Watson, R., & Stenfors, T. (2018). Investigating psychometric properties and dimensional structure of an educational environment measure (DREEM) using Mokken scale analysis–a pragmatic approach. *BMC medical education*, *18*(1), 1-16.

Roohani Tonekaboni, F., Ravand, H., Rezvani, R. (2021). The construction and validation of a q-matrix for a high-stakes reading comprehension test: A G-DINA study. *International Journal of Language Testing*, *11*(1), 58-87.

Setiyadi, R., Kuswendi, U., & Ristiyana, G. M. (2019). Learning of reading comprehension through reading workshop in the Industry 4.0. *Mimbar Sekolah Dasar*, *6*, 160–173. https://doi.org/10.17509/mimbar-sd.v6i2.17397.

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137-158.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (1st ed.). Thousand Oaks, CA: SAGE.

Van der Ark, L. A. (2021). R package Mokken V 3.0.6**.** Retrieved from https://cran.r-project.org/web/packages/mokken

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1-27.

Van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. Thousand Oaks, CA: SAGE.

Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*(2), 139-163.

Tabatabaee-Yazdi, M. (2020). Hierarchical Diagnostic Classification Modeling of reading comprehension. *SAGE Open*, *10*(2), 2158244020931068.

Tabatabaee-Yazdi, M., & Baghaei, P. (2018). Reading comprehension in English as a foreign language and some Cattell-Horn-Carroll cognitive ability factors. *The Reading Matrix: An International Online Journal*, *18*(1), 19-26.

Watson, R., van der Ark, L. A., Lin, L. C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in clinical practice. *Journal of clinical nursing*, *21*(19-20), 2736-2746.

Wind, S. A. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice*, *36*(2), 50-66.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Psychoeducational Battery-Third Edition*. Itasca, IL: Riverside Publishing.