Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 11, No. 2, October 2021
ISSN 2476-5880

Mokken Scale Analysis of the Reading Comprehension Section of the International English Language Testing System (IELTS)

Fatemeh Firoozi[1]

Received: 23 May 2021 Accepted: 29 June 2021

Abstract

Large-scale standardized ESL tests such as the International English Language Testing System (IELTS) are widely used around the world to measure the language proficiency of test-takers and make different decisions based on their scores. Reading comprehension is an integral part of such tests which requires test-takers to read passages and answer a set of questions. Although IELTS is a popular standardized test and is used for making critical decisions about test-takers, very few attempts have been made to explore the validity of the exam and especially the reading part of the General Training Module. With this in mind, the purpose of the present study was to use a non-parametric item response theory model, called Mokken Scale Analysis (MSA), to examine the validity of the reading part of the General Training module of IELTS. To this end, item responses of 352 test-takers to the reading comprehension test were analyzed. The results of item scalability, total scalability, and item-pair scalability showed that the reading part is a weak unidimensional scale. Using Monotone Homogeneity Model (MHM), the monotonicity results also indicated that there are some items which violate the monotonicity assumption, although their values are insignificant. The analysis of unidimensionality using the AISP revealed that there are two scales and four unscalable items in the reading part. Therefore, the Mokken scale analysis did not support the unidimensional structure of the reading part of the General Training module of IELTS.

Keywords: General Training Module, IELTS, Mokken Scale Analysis, Monotone Homogeneity Model, Reading Comprehension Section, Validity

1. Introduction

1.1 IELTS Reading Moduel

Reading comprehension in a second/foreign language (L2) is regarded as a highly complex cognitive process in which the meaning is constructed through the interaction of the reader with the text (Zhang, 2012). Successful reading comprehension requires several considerable

[1] English Department, Islamic Azad University, Mashhad Branch, Iran.
Email: fatemehfirozi74@gmail.com

ISSN 2476-5880

knowledge sources, including decoding skills (García & Cain, 2014), linguistic resources (Aryadoust & Baghaei, 2016; Grabe, 2009; Perfetti, Landi, & Oakhill, 2005), and (meta)cognitive processes (Pearson, 2009) along with the right psychological attitudes (Baghaei, Hohensinn, Kubinger, 2014) . Many researchers have developed different theories to explain reading comprehension performance and specify different facets of reading behavior. The ability to read efficiently in a second/foreign language is thought to play a significant role in the process of learning English and the success of individuals in school and in the workplace (Alderson, 2000). Consequently, the assessment of reading ability is of critical importance in various educational settings and second language programs. In the field of second language testing and assessment, reading comprehension tests, which require test takers to comprehend a text, are extensively utilized in standardized language proficiency tests. Large-scale standardized English as a second language (ESL) tests are widely used around the world to measure language proficiency of candidates who wish to work or study in English-speaking environments. One well-known language proficiency test is the International English Language Testing System (IELTS), which is widely used in Australia, Canada, New Zealand, the UK, and the USA.

IELTS is an international high-stakes test of English language proficiency which is jointly owned by three test batteries: The British Council, University of Cambridge ESOL Examinations, and the International Development Program of Australian Universities and Colleges (IDP), now known as IDP: IELTS Australia. The test consists of two modules: General Training module and Academic module. Each module includes four parts: Listening, Reading, Writing, and Speaking. Test takers can take the same Listening and Speaking modules, whereas they are administered different reading and writing modules depending on whether they choose to take either general or academic versions of the test. There are a variety of tasks and response types within each part which measure all four language skills. The general module of IELTS measures the language proficiency of those test-takers who want to work in English language environments, migrate to English-speaking countries, intend to study at below degree level, or generally undertake non-academic training activities. The academic module of IELTS, however, measures the degree to which test-takers can study or receive training in English at graduate and undergraduate levels. IELTS provides a score for each module and the test scores of each test component are averaged and rounded to generate an Overall Band Score.

A central part of the IELTS is the reading sub-test. As stated by test specifications of the IELTS (IELTS, 2007), the general reading module, which is the main concern of the current study, measures the ability of test-takers to follow instructions, understand main ideas, skim for general information, understand specific information, identify the relationship between information, and summarize. The reading part consists of three sections with 40 questions and different topics designed to be of general interest. The texts are different in lengths with a total of 2000 to 2750 words and are taken from magazines, books, journals, notices, leaflets, advertisements, newspapers, and instruction manuals. The first section, known as ‘Social Survival’, focuses on texts with detailed factual information. The second section involves texts of more complex

ISSN 2476-5880

language. The third section involves texts with more complex language and with the emphasis on descriptive and instructive texts. Different types of questions or tasks are employed to test candidates' reading comprehension including short-answer, multiple-choice, identifying information (True/False/Not Given), identifying writers' view/claims (Yes/No/Not Given), flow-chart completion, diagram label, sentence completion, summary completion, note completion, table completion, and matching information/ headings. The candidates are supposed to answer the test in 60 minutes. Test takers must enter their answers on an answer sheet during the 60-min test. No extra time is allowed for transferring answers.

As IELTS is a high-stakes language proficiency test, scores obtained from the test represent test takers' language proficiency and provide appropriate evidence for further decision-making processes. The inferences and decisions made with the scores from the test have tremendous consequences for all stakeholders. Therefore, test developers and test users mainly concern about the validity of IELTS in general and reading comprehension, in particular. It has been well-established that validity is the most important feature of a test. According to Messick (1989), validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13, italics in original). This view states that what needs to be validated is the interpretations and uses of test scores, not the test itself. As Kane (1992, 2013) proposed in his argumentative approach to validity, test scores are used to support test interpretations and uses. In this approach, there are two types of arguments: the interpretive argument and the validity argument. In the first step, researchers should determine what claims, inferences, and uses are supposed to make regarding the intended meaning based on test scores, and in the second step, the interpretive argument is evaluated to find out whether the claims should be supported or refuted. Although the argument-based account has been well received by language testing researchers and practitioners, Borsboom, Mellenberg, and van Heerden (2004) argue that validity is a simple concept as was lucidly defined by Kelley (1927), i.e., a test is valid if it measures what it claims to measure. In fact, validity is only the property of tests, not test scores and interpretations. According to Borsboom et al. (2004, p. 1061), a test is valid for measuring a construct if (a) the construct exists and (b) variations in the level of the construct causally produce variations in test scores. For example, in a reading comprehension test, the reading ability should cause variations in test scores, that is, individuals with higher level of reading ability should get higher scores, and individuals with lower level of reading ability should get lower scores. This view toward validity relies on the causal theory of measurement which is in line with the latent trait model (Borsboom, 2005).

Over the past few decades, a great deal of research has been conducted to examine the validity of the reading part of the IELTS (Bax, 2013; Clapham, 1996; Green & Hawkey, 2007; Moore, Morton, Price, 2007; Pearson, 2019; Weir, Hawkey, Green, & Devi, 2009; Weir, Hawkey, Green, Unaldi, & Devi, 2009). Although these studies provided invaluable information on the validity of the test, they mainly focused on the reading part of the Academic module of IELTS and

ISSN 2476-5880

too little attention has been paid to the General Training module. In addition, the methods used in these studies to establish validity rely heavily on statistical tools which have their roots in classical test theory (CTT). Despite its widespread use in measurement, CTT has received some acknowledged limitations. One problem is the issue of sample-dependent statistics, that is, the person parameters depend on the selection of items and item parameters depend on the sample (Embretson, 1996). Another limitation of CTT is that the data should be continuous and normally distributed. Also, scores derived from CTT are not very informative about the item response patterns, and "any combination of scores on any set of items can give the same score on the latent trait" (Palmgren et al., 2018, p. 3). With respect to the limitations of CTT, item response theory (IRT) was developed as an alternative method to CTT. IRT was developed based on ordinal and nominal data, and models the encounter of an individual with a given ability with a test item. One branch of the IRT approach is Mokken Scale Analysis (MSA; Mokken, 1971). As a non-parametric IRT, MSA is a popular method to evaluate the psychometric quality of data obtained from scales and determine scales from a larger set of items. Similar to the principles of IRT, MSA examines the relationship between the construct and a set of items. The model assumes that items of a scale are hierarchically ordered along the latent trait continuum. However, compared to conventional or parametric IRT, MSA includes less restrictive assumptions about the data, requires smaller sample sizes to produce stable estimation, and its fit to data is better than parametric IRT models (Mokken & Lewis, 1982; Sijtsma & Van der Ark, 2017). These properties allow researchers to retain more useful items from a scale. Given the advantages of the Mokken model, this study seeks to use the model to examine the validity of the reading part of the General Training module of IELTS.

1.2 Mokken Scale Analysis (MSA)

Mokken Scale Analysis (MSA; Mokken, 1971) is a probabilistic non-parametric item response theory (NIRT) model derived from Guttman scaling, as a deterministic model (Baghaei, 2021). Mokken scales are used to explore basic measurement properties and evaluate the dimensionality and scalability of psychometric measures (van Schur, 2003). MSA is considered a non-parametric model because: (1) it is unnecessary to specify a mathematical form for item response function (IRF); and (2) it does not make any assumptions about the distribution of person parameters (Mokken & Lewis, 1982). In MSA, items and persons are hierarchically ordered on an ordinal scale based on individuals' trait level and items' difficulty level (Baghaei, 2020). Mokken (1971) proposed a set of scalability coefficients, that are used to investigate whether individual items, pairs of items, and overall sets of items form a scale, which satisfy the criteria for Mokken scale analysis. There are two model versions of Mokken's nonparametric approach to IRT: the monotone homogeneity model (MHM) and the double monotonicity model (DMM). Mokken was originally developed for dichotomous responses, but polytomous versions have been proposed (Molenaar, 1982, 1997). The monotone homogeneity model (MHM) relies on a number of assumptions: (1) *unidimensionality*: a single common latent trait explains responses on a set of items; (2) *monotonicity*: the relationship between the probability of correct answer and individuals' locations

ISSN 2476-5880

on the latent trait should be monotone non-decreasing; and (3) *local independence*: the responses to items measuring the construct should not depend on each other. When the data fit the MHM assumptions, the IRFs are increasing or at least constant and persons can be invariantly ordered with their raw total scores.

The double monotonicity model (DMM) is the restrictive version of the MHM. In addition to three common assumptions, the DM model requires the assumption of *invariant item ordering (IIO)* or *non-intersecting IRFs*. IIO indicates that item response functions do not intersect or cross over. When the IIO assumption holds for a number of items, the items are ordered from the easiest to the most difficult. In fact, if one item is more difficult than another one for a test taker, the item should be more difficult for all test takers. This provides evidence for invariant ordering of items and test-takers. Similar to MHM, the DMM allows practitioners to order persons on the latent trait based on their raw total scores.

2. Review of Literature

Mokken scale analysis (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002) has successfully been used to evaluate the psychometric quality of scales in different areas including criminology (Santtila et al., 2008), educational measurement (Wind, 2019; Wind & Engelhard, 2016), health sciences (Emons, Sijtsma, & Pedersen, 2012; Palmgren et al., 2018; Watson, Deary, Gow, & Shipley, 2008; Zhang & Li, 2020), marketing (Paas & Sijtsma, 2008), political science (Jacoby, 2008), psychiatry (Bech, Wilson, Wessel, Lunde, & Fava, 2009; Chou, Lee, Liu, & Hung, 2017; Korner et al., 2007), psychology (Myszkowski, 2020; Watson, Roberts, Gow, & Deary, 2008) and sociology (Loner, 2008). However, too little attention has been devoted to MSA in L2 testing research. A review of the literature revealed too few studies in second and foreign language assessment. Tabatabaee-Yazdi, Motallebzadeh, and Baghaei (2021) used MSA on a 20-item reading comprehension test to examine the unidimensionality and scalability of the items. Their results showed that Monotone Homogeneity Model (MHM) has adequate fit to all items as measured by the scalability coefficient because test items could rightly order students on the latent trait with regard to their reading comprehension ability. They also indicated that the ordering of items based on their mean is invariant across examinees. Using the automated item selection procedure (AISP), they concluded that the reading comprehension test is unidimensional which can be considered as evidence of validity, that is, the test measures only a single ability. In a recent study, Baghaei (2021) analyzed the listening part of the IELTS with the monotone homogeneity model of Mokken. The results of AISP showed that the listening part of the IELTS is unidimensional although two items formed the second dimension and four items were unscalable. It is clear that there is a paucity of research on the application of MSA in language testing and assessment and more studies are required to explore the suitability of the MSA on language skills. Therefore, the present study aims to use the MSA on the reading part of the General Training module of IELTS to examine the psychometric quality of test items.

ISSN 2476-5880

3. Method

3.1. Participants and Setting

A total of 352 undergraduate English as a foreign language (EFL) university students participated in this study. There were 115 male and 237 female students. The ages of these participants ranged from 18 to 22 (M= 20; SD=2.33). Participants were selected from the English departments of four universities in Mashhad, Iran. All of the participants were bilingual and their home language background was Persian.

3.2 Instrumentation

Participants were given a version of the General Training IELTS reading test to evaluate test takers' reading comprehension ability. The reading part consists of three passages of different lengths and 40 items. Students were asked to answer the test in 60 minutes. They were reassured that their information would remain confidential and anonymous.

4. Results

4.1. Descriptive Statistics

Table 1 presents descriptive statistics for the test data, computed on SPSS for Windows, Version 23. The total score in the 40-item ranged from 7 and 38 with a mean of 21.61 and a standard deviation of 22.00.

Table 1.
Descriptive statistics for the IELTS reading comprehension test

Mean	Median	Mode	SD	Variance	Range	Minimum	Maximum
21.6136	22.0000	23.00	5.96272	35.554	31.00	7.00	38.00

4.2. Mokken Scale Analysis

4.2.1. Scalability Coefficients. Mokken package version 3.0.6 (van der Ark, Koopman, Straat, & van den Bergh, 2021) in R (R Core Team, 2018) was used to run the monotone homogeneity model (MHM). As a first step in working with the Mokken model, item scalability (H_j), item-pair scalability (H_{ij}), and overall scalability *(H)* coefficients were examined. Mokken (1971) classified coefficient values as follows: values smaller than 0.30 indicate a weak scale, 0.40 ≤ H < 0.50 a medium scale, and H ≥ 0.50 a strong scale. As Table 3 shows, none of the item scalability coefficients was smaller than zero, but there are many values below 0.30.

Table 3.
Item scalability coefficients for the 40 items and their standard errors

Item	Scalability Coefficients	SE
V 121	0.225	(0.030)
V 122	0.123	(0.033)

ISSN 2476-5880

V 123	0.214	(0.032)
V 124	0.319	(0.031)
V 125	0.189	(0.033)
V 126	0.275	(0.056)
V 127	0.291	(0.028)
V 128	0.291	(0.038)
V 129	0.419	(0.064)
V 130	0.192	(0.034)
V 131	0.270	(0.041)
V 132	0.430	(0.055)
V 133	0.425	(0.050)
V 134	0.397	(0.038)
V 135	0.224	(0.032)
V 136	0.258	(0.030)
V 137	0.276	(0.032)
V 138	0.271	(0.045)
V 139	0.268	(0.039)
V 140	0.254	(0.063)
V 141	0.216	(0.031)
V 142	0.240	(0.045)
V 143	0.209	(0.035)
V 144	0.282	(0.035)
V 145	0.463	(0.066)
V 146	0.311	(0.056)
V 147	0.179	(0.060)
V 148	0.170	(0.035)
V 149	0.187	(0.034)
V 150	0.254	(0.037)
V 151	0.212	(0.052)
V 152	0.106	(0.037)
V 153	0.245	(0.049)
V 154	0.491	(0.056)
V 155	0.276	(0.039)
V 156	0.243	(0.039)
V 157	0.302	(0.033)
V 158	0.280	(0.045)
V 159	0.334	(0.026)
V 160	0.237	(0.031)

Note: SE: Standard Errors

ISSN 2476-5880

The overall scalability coefficient was 0.25 which indicates a weak scale. Table 4 shows the number of negative item pair scalability coefficients (H_{ij}) that each item is involved in. As one can see, most of the items have at least one negative item pair scalability coefficient. However, these negative values are very small and near zero with extremely large standard errors and therefore can be ignored as random fluctuations in the data.

Table 4.
Number of negative item pair scalability coefficients (H_{ij}) for each item

Item	No. Negative H_{ij}
1	0
2	1
3	1
4	1
5	0
6	0
7	2
8	1
9	3
10	3
11	2
12	1
13	2
14	1
15	0
16	1
17	2
18	0
19	1
20	2
21	1
22	1
23	0
24	1
25	3
26	5
27	7
28	1
29	4
30	1

ISSN 2476-5880

31	1
32	1
33	0
34	1
35	1
36	1
37	0
38	1
39	0
40	0

4.2.2. Monotonicity. Table 4 shows the number of violations of monotonicity for each item (#vi). As can be seen, items 10, 12, and 28 violated the assumption of monotonicity. However, these violations are not significant (#zsig). Crit values in the last column also show that item 28 extremely violates the monotonicity assumption. As recommended by Molenaar and Sijtsma (2000), values smaller than 0.40 show that items do not extremely violate the monotonicity assumption whereas values larger than 0.40 indicate a violation of monotonicity. Therefore, item 28 is the most serious item. As the z-value for the item is insignificant, we can keep it for further analysis. Thus, the monotone homogeneity model holds for the reading comprehension part of the IELTS and the raw scores can be used to locate examinees on an ordinal scale.

Table 5.
Number and statistical significance of monotonicity violation for the items

Item	H	#ac	#vi	#zsig	crit
1	0.22	6	0	0	0
2	0.12	6	0	0	0
3	0.21	6	0	0	0
4	0.32	6	0	0	0
5	0.19	6	0	0	0
6	0.27	6	0	0	0
7	0.29	6	0	0	0
8	0.29	6	0	0	0
9	0.42	2	0	0	0
10	0.19	6	1	0	36
11	0.27	6	0	0	0
12	0.43	6	1	0	23
13	0.42	6	0	0	0
14	0.40	1	0	0	0
15	0.22	6	0	0	0

ISSN 2476-5880

16	0.26	6	0	0	0
17	0.28	6	0	0	0
18	0.27	6	0	0	0
19	0.27	6	0	0	0
20	0.25	6	0	0	0
21	0.22	6	0	0	0
22	0.24	6	0	0	0
23	0.21	6	0	0	0
24	0.28	6	0	0	0
25	0.46	1	0	0	0
26	0.31	1	0	0	0
27	0.18	6	0	0	0
28	0.17	6	1	0	57
29	0.19	6	0	0	0
30	0.25	6	0	0	0
31	0.21	6	0	0	0
32	0.11	6	0	0	0
33	0.24	6	0	0	0
34	0.49	3	0	0	0
35	0.28	6	0	0	0
36	0.24	6	0	0	0
37	0.30	6	0	0	0
38	0.28	6	0	0	0
39	0.33	6	0	0	0
40	0.24	6	0	0	0

Figures 1, 2, and 3 show the item response functions (IRF) for items 10, 12, and 28, respectively. These are the items that violated the monotone homogeneity assumption. As the IRFs for items 10, 12, and 28 show, they do not monotonically increase across the latent ability continuum, and thus, they violated the monotonicity assumption of the MSA. However, as Table 4 showed, the violations are small and statistically non-significant. Figure 4 shows the IRF for item 21 which does not violate the monotonicity assumption. As can be seen, the IRF is monotonically increasing with no break across the trait scale.

ISSN 2476-5880

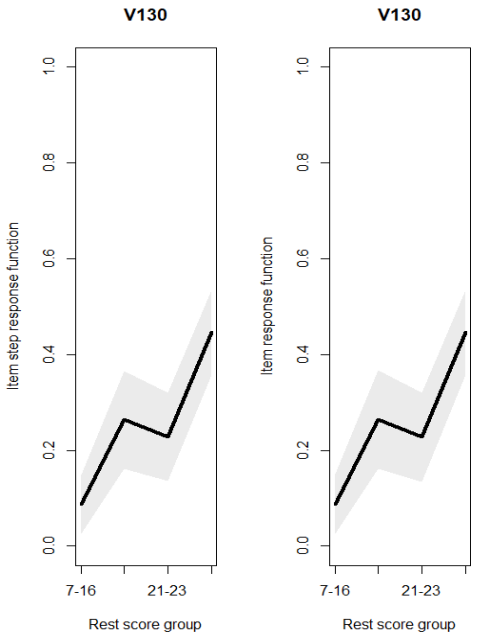

Figure 1. Item response function for Item 10

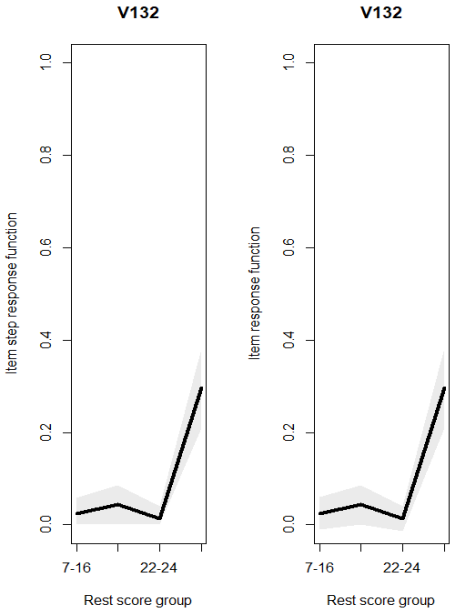

Figure 2. Item response function for Item 12

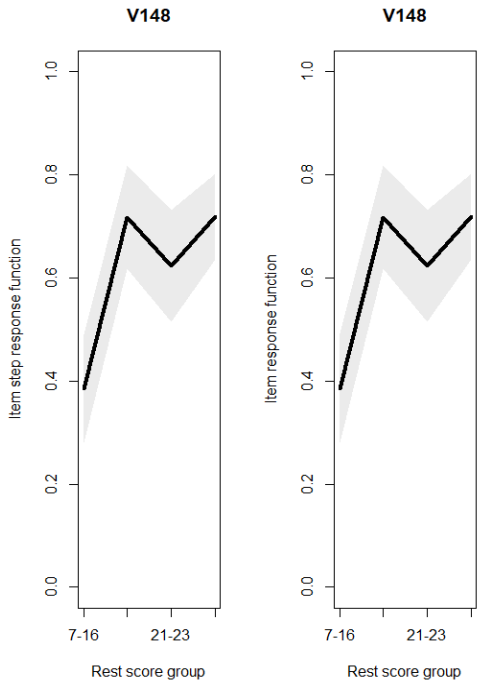

Figure 3. Item response function for Item 28

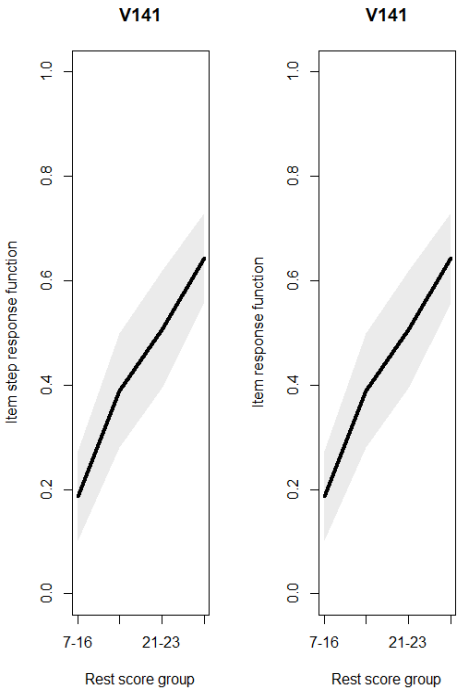

Figure 4. Item response function for Item 21

ISSN 2476-5880

4.2.3. Automatic Item Selection Procedure (AISP)

To identify the dimensionality of the IELTS reading comprehension test, automatic item selection procedure (AISP) was employed. AISP can specify and eliminate non- or low-discriminating items from the scale (Sijtsma, & van der Ark, 2017). Considering Sijtsma and van der Ark's (2017) suggestion for the AISP with a cut-off value of c=0.3 ($HJ \geq c \geq 0$), AISP identified two scales and four unscalable items. Table 5 shows the items of each scale.

Table 5.
AISP results for the IELTS reading module

	Scale 1	Scale 2	Unscalable
Items	Rest	10,11,22,24,26,29,30,31,34,38	2,20,27,32

The researcher computed four different reliability coefficients: Mokken scale (MS) reliability ρ, (Mokken, 1971), Lambad-2 (Guttman, 1945), Cronbach's alpha (Cronbach, 1951), and the latent class reliability coefficient (van der Ark, van der Palm, & Sijtsma, 2011). Table 6 shows the different reliability coefficients for the 40-item IELTS reading part. As can be seen, all the values are above 0.80, indicating acceptable reliability.

Table 6.
Reliability indices for the IELTS reading comprehension section

Reliability Index	MS ρ	Lambad-2	Alpha	LCRC
Value	.860	.844	.841	.885

5. Discussion

The purpose of the present study was to examine the psychometric properties of the reading comprehension part of the International English Language Testing System (IELTS) using the monotone homogeneity model (MHM) of the Mokken Scale Analysis. The results of the scalability coefficient showed that all items have positive item scalability coefficient although most of them were below the cut-off value (e,g., 0.30). The overall scalability coefficient was 0.24, indicating that the scale is weak. The values of item pair scalability coefficients also showed that all coefficients are positively related with very large standard errors, indicating a non-negative relation between the construct and the items.

The analysis of monotonicity using the restscore method and graphical checks showed that only three items out of 40 violate the monotonicity assumption of the MHM with item 28 as the

ISSN 2476-5880

most serious item. However, the violations were not statistically significant. That is, as the latent trait level increases, the probability of a correct response increases too. This is a major assumption of the MHM and other item response theory models and a key condition for validity (Baghaei & Shoahosseini, 2019; Baghaei & Tabatabaee-Yazdi, 2016). Therefore, one can order the examinees on an ordinal scale using the total raw scores.

Automatic item selection procedure (AISP) was also used to evaluate the dimensionality of the reading comprehension part of the IELTS. Results of the AISP showed that the items do not cluster around a single dimension, that is, 26 items formed the first scale, 10 items formed the second scale, and four items were unscalable. Content analysis of the items which formed the second scale and those which were unscalable revealed the multidimensionality nature of the IELTS reading part. The possible reason for the multidimensionality of the reading test would be the use of multiple methods for measuring reading comprehension. Alderson (2000) argued that an interesting feature of the IELTS reading test is that different techniques are used to assess understanding of any one passage. The use of multiple methods is a strength for tests since

> "it is now generally accepted that it is inadequate to measure the understanding of text by only one method, and that objective methods can usefully be supplemented by more subjectively evaluated techniques. Good reading tests are likely to employ a number of different techniques, possibly even one the same text, but certainly across the ranges of texts tested. This makes good sense, since in real-life reading, readers typically respond to texts in a variety of different ways." (p. 206)

Although the use of multiple methods could be more interesting and authentic, they are considered as potential threats to validity (Baghaei & Kubinger, 2015; Baghaei & Ravand, 2019). As Bachman (1990) noted, "When test performance is unduly affected by factors other than the ability being measured, the meaningfulness or validity of score interpretations is lessened." (p.156). Therefore, it seems that test methods are the main source of construct-irrelevance variance in the data (Messick, 1989).

Furthermore, reliability analysis using Mokken scale reliability, Lambda-2, coefficient alpha, and latent class reliability coefficient showed that the test is highly reliable with indices above .84. This finding indicates that the order of the test-takers would be the same if the test was repeated.

Our findings on the reading part of the General Training module of IELTS converge with Baghaei (2021) who employed the monotone homogeneity model of Mokken to investigate the validity of the listening part of the IELTS. Similar to the current study, Baghaei (2021) showed that the listening part of the IELTS is unidimensional even though two items formed the second dimension, four items were unscalable, and the rest of the items created a single scale. All item scalability coefficients were positive and the overall scalability coefficient was low (H=.36). However, he reported higher reliability for the listening part. Baghaei (2021) concluded that, although test methods can affect the dimensionality of the test, the listening part of the IELTS is unidimensional and reliable, and thus the test can correctly order test-takers across the latent trait continuum. The current study, however, is in disagreement with Tabatabaee-Yazdi et al.'s (2021)

study which applied the Mokken model to assess the dimensionality and scalability of a 20-item English reading comprehension test. They showed that the monotone homogeneity model has adequate fit to all items as measured by the scalability coefficient. Using automated item selection procedure (AISP), they indicated that the reading comprehension test is unidimensional and thus is considered as evidence of validity. They concluded that students can be properly ordered on the latent trait continuum based on their reading comprehension ability.

6. Conclusion

This study was aimed to investigate the dimensionality and scalability of the reading part of the General Training module of IELTS. The findings indicated that due to the use of different test methods, the reading test is multidimensional. However, the test can correctly order test-takers on the latent trait continuum, that is, with the increase of latent trait, the probability of correct answers increases too.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Aryadoust, V., & Baghaei, P. (2016). Does EFL readers' lexical and grammatical knowledge predict their reading ability? Insights from a Perceptron Artificial Neural Network study. *Educational Assessment, 21,* 135-156.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: OUP.

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

Baghaei, P., & Shoahosseini, R. (2019). A note on the Rasch model and the instrument-based account of validity. *Rasch Measurement Transactions, 32,* 1705–1708.

Baghaei, P., & Tabatabaee-Yazdi, M. (2016). The logic of latent variable analysis as validity evidence in psychological measurement. *The Open Psychology Journal, 9,* 168-175.

Baghaei, P., Hohensinn, C., & Kubinger, K.D. (2014). The Persian adaptation of the foreign language reading anxiety scale: A psychometric analysis. *Psychological Reports, 114,* 315-325.

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20,* 1-11.

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), 1-8. doi:10.1177/2158244019832706

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441–465. doi:10.1177/0265532212473244

Bech, P., Wilson, P., Wessel, T., Lunde, M., & Fava, M. (2009) A validation analysis of two selfreported HAM-D-6 versions. *Acta Psychiatrica Scandinavica*, *119*(4), 298–303. doi:10.1111/j.1600-0447.2008.01289.x

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. doi:10.1037/0033-295x.111.4.1061

Chou, Y.H., Lee, C.P., Liu, C.Y., Hung, C.I. (2017). Construct validity of the Depression and Somatic Symptoms Scale: Evaluation by Mokken Scale Analysis. *Neuropsychiatr Disease and Treat*. *13*, 205-211. doi:10.2147/NDT.S118825

Clapham, C. (1996). *The Development of IELTS: a study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. doi:10.1007/BF02310555

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment*, *8*, 341-349. doi:10.1037/1040-3590.8.4.341

Emons, W.H.M., Sijtsma, K, Pedersen, S.S. (2012). Dimensionality of the Hospital Anxiety and Depression Scale (HADS) in Cardiac Patients: Comparison of Mokken Scale Analysis and Factor Analysis. *Assessment*, *19*(3):337-353. doi:10.1177/1073191110384951

García, J. R., & Cain, K. (2014). Decoding and reading comprehension: a meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research, 84*(1), 74-111. doi:10.3102/0034654313499616

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Green, A., & Hawkey, R. (2007). An empirical investigation of the process of writing Academic Reading test items for the International English Language Testing System, *IELTS Research Reports, 11*(5), 273-374.
URL: https://www.ielts.org/for-researchers/research-reports/volume-11-report-5

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255-282. doi: 10.1007/BF02288892

IELTS (2007). *The IELTS handbook.* UCLES/British Council, IDP Education Australia, Cambridge. Retrieved from: www.ielts.org

Jacoby, W.G. (2008). Comment: The dimensionality of public attitudes toward government spending. *Political Research Quarterly*, *61*(1), 158–161. doi:10.1177/1065912907309860

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1): 1-73. doi:10.1111/jedm.12000

Korner, A., Lauritzen, L., Abelskov, K., Gulmann, N. C., Brodersen, A. M., Wedervang-Jensen, T., & Kjeldgaard, K. M. (2007). Rating scales for depression in the elderly: External and internal validity. *Journal of Clinical Psychiatry*, *68*(3), 384–389. doi:10.4088/jcp.v68n0305

Loner, E. (2008). The importance of having a different opinion: Europeans and GM foods. *European Journal of Sociology*, *49*(1), 31–63. doi: 10.1017/S0003975608000027

Messick, S. (1989). Validity. In: Linn RL (ed.) *Educational Measurement*. New York: Macmillan, pp. 13-103.

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

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*(4), 417– 430. doi:10.1177/014662168200600404

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden*, *3*(8), 145–164.

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.

Moore, T., Morton, J., Price, S. (2007). Construct validity in the IELTS academic reading test: a comparison of reading requirements in IELTS test items and in university study. *IELTS Research Reports, 11*(4), *1-89.* URL:https://www.ielts.org/for-researchers/research-reports/volume-11-report-4

Myszkowski N. (2020). A Mokken Scale Analysis of the last Series of the Standard Progressive Matrices (SPM-LS). *Journal of Intelligence*, *8*(2), 22. doi:10.3390/jintelligence8020022

Paas, L.J., & Sijtsma, K. (2008). Nonparametric item response theory for investigating dimensionality of marketing scales: A SERVQUAL application. *Marketing Letters*, *19*, 157–170. doi:10.1007/s11002-007-9031-0

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), 235. doi:10.1186/s12909-018-1334-8

Pearson, W. S. (2019). 'Remark or retake'? A study of candidate performance in IELTS and perceptions towards test failure. *Language Testing in Asia*, *9*(17), 1-20. doi:10.1186/s40468-019-0093-8

Pearson, P. D. (2009). The roots of reading comprehension instruction. In S. E. Isreal & G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 3–31). New York: Routledge.

Perfetti, C.A., Landi, N., Oakhill, J.V. (2005). The acquisition of reading comprehension skill. In Snowling, M. J., Hulme, C. (Eds.), *The science of reading: A handbook* (pp. 227- 247). Oxford, UK: Blackwell. doi:10.1002/9780470757642.ch13

Santtila, P., Pakkanen, T., Zappalà, A., Bosco, D., Valkama, M., & Mokros, A. (2008).

ISSN 2476-5880

Behavioural crime linking in serial homicide. *Psychology, Crime & Law*, *14*(3), 245-265. doi: 10.1080/10683160701739679

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Thousand Oaks, CA: Sage.

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. doi:10.1111/bmsp.12078

Tabatabaee-Yazdi, M., Motallebzadeh, Kh., & Baghaei, P. (2021). A Mokken Scale Analysis of an English reading comprehension test. *International Journal of Language Testing*, *11*(1), 131-143. URL: http://www.ijlt.ir/article_130373.html

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*(5), 380–392. doi:10.1177/0146621610392911

van der Ark, L.A., Koopman, L., Straat, J.H., & van den Bergh, D. (2021). R package Mokken V 3.0.6**.** Retrieved from https://cran.r-project.org/web/packages/mokken

van Schur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*(2), 139–163. doi:10.1093/pan/mpg002

Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine*, *38*(4), 575–579. doi:10.1017/S003329170800281X

Watson, R., Roberts, B., Gow, A., & Deary, I. J. (2008). A hierarchy of items within Eysenck's EPI. *Personality and Individual Differences*, *45*(4), 333–335. doi:10.1016/j.paid.2008.04.022

Weir, C., Hawkey, R., Green, T., & Devi, S. (2009). The cognitive processes underlying the academic reading construct as measured by IELTS. *British Council/IDP Australia IELTS Research Reports*, *9*(4), 157–189. URL: https://www.ielts.org/for-researchers/research-reports/volume-09-report-4

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. *IELTS Research Reports*, *9*, 97–156 URL: https://www.ielts.org/for-researchers/research-reports/volume-09-report-3

Wind, S.A. (2019). Nonparametric evidence of validity, reliability, and fairness for rater-mediated assessments: an illustration using Mokken Scale Analysis. *Journal of Educational Measurement*, *56*(3), 478-504. doi:10.1111/jedm.12222

Wind, S.A, & Engelhard, G. (2016). Exploring rating quality in rater-mediated assessments using Mokken Scale Analysis. *Educational and Psychological Measurement*, *76*(4), 685-706. doi:10.1177/0013164415604704

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*(4), 558–575. doi:10.1111/j.1540-4781.2012.01398.x

ISSN 2476-5880

Zhang, L., & Li, Z. (2020). A Mokken scale analysis of the Kessler-6 screening measure among Chinese older population: Findings from a national survey. *BMC Geriaitrics*, *20*, 1-11. doi:10.1186/s12877-020-01771-w