Are reading comprehension tests composed of different test formats unidimensional?

Elias Sheybani, Mitra Zeraatpishe

Are reading comprehension tests composed of different test formats unidimensional?

Elias Sheybani

English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran

Mitra Zeraatpishe

English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran

Test method is deemed to affect test scores along with examinee ability (Bachman, 1996). In this research the role of method facet in reading comprehension tests is studied. Bachman divided method facet into five categories, one category is the nature of input and the nature of expected response. This study examined the role of method effect in the Iranian University Entrance Exam (UEE). Research showed the current test method of assessing reading can highly impact test takers' performance and thereby examinees' test scores. Hence, different test formats may extract various traits and therefore, contaminate test scores and threaten validity. The researchers using exploratory factor analysis (EFA) extracted a second method factor in the reading comprehension section of the UEE which is composed of a multiple-choice cloze passage and multiple choice questions based on sustained texts. Results of the analysis showed that cloze test can create construct-irrelevant method variance, contaminating test scores, and led to violation of the unidimensionality assumption of the test.

Keywords: Cloze Test, Exploratory Factor Analysis, Method Bias, Reading Comprehension Ability, Unidimensionality Assumption

1. Introduction

Within an educational setting, students are often evaluated on how much of the assigned materials have been mastered. During the evaluation procedure, they have to cope simultaneously with the content of the test on one hand and the format of the test on the other. A test can take different formats including multiple choice, cloze, matching, true/false, short answer, and essay. Although different in the formats, they share a commonality in that the items have a stem which represents a problem to be solved, but they differ in the subject matter they present (Crook, 1988). For example, in the constructed response questions, examinees must create their own answers which might be a diagram, a short answer, an essay, or an explanation of a procedure (Lukhele, Thissen, & Wainer, 1993). Bachman and Palmer (1996) state that characteristics of test methods and formats affect examinees' scores to some extent. They further suggest that it is impossible to prevent the effects of test methods, so it is necessary to control them, as much as possible. There are so many factors that commonly affect examinees' performance like test formats. Bachman (1996) argues that test score is affected by three major systematic factors, target ability, test method facets (i.e. characteristics of the formats used to elicit knowledge or ability), test takers' personal attributes (i.e. testees' individual or group characteristics). In this research, the aim is to examine the impact of test method facets. Among the most popular test formats in assessing reading are cloze format and reading comprehension questions which are based on sustained passages are the most popular test forms.

1.1. Reading Comprehension Assessments

Reading comprehension is an essential component of learning (Francis, Fletcher, Catts, & Tomblin, 2005). Individuals who experience difficulties reading and understanding information

tend to suffer from problems in school in the workplace and in their communities (National Center for Education Statistics, 2002; National Reading Panel, 2000). Not only formal education is difficult for these individuals, but also is the opportunities to reflect, share ideas, and reason effectively. The salience of gaining reading ability brings to mind the concept of assessing such ability. Among several types of tests to measure reading comprehension ability are text comprehension multiple-choice items based on sustained passages and cloze test items. These two test types are used in international language tests, such as Cambridge English tests including First Certificate in English (FCE), Certificate in Advanced English (CAE), Certificate of Proficiency in English (CPE), Cambridge Preliminary English Test (PET) and also in the Michigan Test of English Language Proficiency (MTELP). These international tests should be reliable and reliability with no unidimensionality is of limited value because we do not know what we are measuring (Heijden, Buuren, Fekkes & Verrips, 2003). So, if a test seeks to be reliable, unidimensionality is inevitable. Unidimensionality in reading assessments composed of multiple-choice text comprehension items and cloze items can be simply expressed as the loading of all items regardless of the format on a single factor (Hambleton, 1991).

1.2. Unidimensionality Assumption

As Hambleton (1992, p.912) explains, "unidimensional scores are what (a) underlying examinee performance on a test is a single ability or trait, and (b) that the relationship between examinee performance on each item and the ability measured by the test can be described by a single factor which is the ability of examinee to answer questions". However, "no measurement is purely unidimensional, there are always some other unwanted factors that contaminate the measurement" (Baghaei, 2009, p. 22). Some findings showed that methods of assessing reading comprehension affect students' scores (Kastner & Stangla, 2010). Validity, however, is defined

based on two issues, what to test and how to test (Bachman, 1995). What refers to the construct and how refers to the test method and if the method affects the construct, the validity of the test is blemished. However, many researchers argued that the methods we use to measure language ability influence performance on language tests (Bachman & Palmer, 1982; Brutsch, 1979; Clifford, 1987; Shohamy, 1984).

1.3. Method Bias

Bachman's (1990) model of language ability (later revised in Bachman & Palmer, 1996) draws our attention to some factors that can have effect on test performance and, therefore, can jeopardize test validity. According to Bachman (1996), one factor is the method of assessment itself. Hence, the problem resides in the fact that uncontrolled method factors can have great effects on a test, as several researchers have demonstrated that method bias can inflate, deflate, or have no effect on the estimates of the relationship between two constructs (e.g., Baumgartner & Steenkamp, 2001; Cote & Buckley, 1988; Podsakoff & MacKenzie, 2003; Siemsen, 2010). In this research, in a combination of cloze and sustained passage comprehension items, we expect to find a single reading comprehension factor and no method factor is expected to emerge.

2. Literature Review

Some researchers (Alderson & Urquhart, 1985; Taillefer & Pugh, 1998) have assumed that cloze test provides valid measurement of reading comprehension skill argue that some test formats, like open-ended questions, can introduce an unwanted nuisance factor into the measurement. They conclude that a unidimensional scaling is not appropriate when different test formats are used in a reading comprehension test (open-ended and multiple-choice), and whenever a multidimensional scaling of reading proficiency test is intended, more open-ended items have to

be included. The validity of the cloze method for assessing reading comprehension ability has been generally accepted for over 50 years due to the correlations between cloze scores and results of other formats of assessment reported by Bormuth (1967, 1968a, 1968b), Tayl. Rauch and Hartig (2010) or (1957), Rankin (1959), and others. Results of some research support the construct validity of cloze tests, by reducing potential sources of bias (e.g. poor wording of questions, ambiguity in recalls, unrepresentativeness of coverage, delayed inference) in other forms of reading assessment (Bormuth, 1968a, 1968b, 1975; Gilliland, 1972; Klare, 1984; Rye, 1982; Ulijn & Strother, 1995). However, several researchers have gained different results and also they challenged the validity of cloze tests. Levenston, Nir and Blum-Kulka (1984) clarify the problem by demonstrating that higher reading skills and global coherence can be clearly assessed only if the placement of deletions is deliberately designed. Given these arguments, it seems reasonable to ask whether cloze tests provide a valid measure of the reading comprehension skill.

The mentioned assumptions reveal that test method can highly affect reading comprehension scores. Keenan (2008) concluded that, although format differences underlie the differences between the tests in their sensitivity to word decoding skills, they are not the format differences that researchers have previously focused on. Previous research has simply suggested that cloze tests differ from other reading comprehension tests (Francis, Fletcher, Catts, & Tomblin, 2005; Nation & Snowling, 1997). Thus Keenan (2008) primarily proved test format is another factor besides reading comprehension, and scores are heavily influenced by this factor. A recent research through factor analysis reported by Weaver and Kingston (1963) showed that comprehension and cloze tests had their largest factor loadings on separate factors and they had

only slight to moderate loadings on the same factors. They concluded that cloze test scores do not appear valid. However, much of the research has shown the exact opposite results, that is, scores on cloze tests are highly correlated with scores on standardized tests of reading comprehension ability. For example, Jenkinson (1957), Ruddell (1963), and Bormuth (1965) all found correlations which ranged from .70 to about .85. Where lower correlations were observed by investigators such as Fletcher (1955) and Rankin (1957), their size could be explained by the reliabilities of the tests administered and the use of their samples.

Unfortunately, despite the effect of method on scores that has been expressed, there is little agreement about whether it is really a problem or not. Many authors believe that method bias is an important problem that needs to be controlled (e.g., Baghaei & Ravand, 2016; Campbell & Fiske, 1959; Cote & Buckley, 1987, 1988; Doty & Glick, 1998; Podsakoff et al. 2003; Podsakoff & Organ, 1986; Sharma, 2009), some claim that it is not real (e.g., Chen & Spector, 1991; Spector & Brannick, 2009). This research is going to study the unidimensionality of a reading comprehension test composed of MC cloze tests and MC sustained text comprehensions items. As Baghaei and Aryadoust (2015) argued, if the unwanted dimensions of cognitive and personality factors are as strong as the reading comprehension factor, students' scores are incomparable. A measuring instrument must not be seriously affected in its measuring function by the method of measurement. To the extent that its measuring function is so affected, the validity of the instrument is limited (Baghaei, 2016). EFA analysis is a suitable method to show whether the reading comprehension test composed of MC cloze tests and MC sustained text comprehensions items is unidimensional or there is a method bias in reading assessments. The findings of this research might be contributing to test development, so that test developers could consider influential factors to design a valid test. They should pay attention to factors which

might influence examinees' scores. Method of assessing itself, is one major factor, hence it is crucial to test developers to pick the most appropriate method.

3. Method

3.1. Participants and Setting

The participants of the present study were among high school graduates (year 2011) who desired to pursue their studies for a bachelor degree of English in Iranian state universities. These 1024 students contained 23% male and 77 % female. For this study, the researcher used UEE data by the permission of National Organization of Educational Testing of Iran. In order to solve the problems related to student selection, researcher got the permission of the center of test assessment called Sanjesh Organization. The University Entrance Examination (UEE) is specific for English studies and is held on a yearly basis.

3.2. Materials and Procedure

University Entrance Exam (UEE) is a test developed for BA candidates in English. This test includes 6 sections of grammar, vocabulary, sentence structure, language functions, cloze test, and reading comprehension; containing 10, 15, 5, 10, 10 and 20 items, respectively. The reading comprehension section and the cloze section were selected for this study. This section consisted of three passages included general knowledge in B1 level. The number of items for the text comprehension questions is 20 and for the cloze is 10. All data were gathered from UEE examinees which contained reading comprehension items and cloze test items. All items were analyzed by 'psych' (Revelle, 2011) package in R software. The examinees' data was taken by the permission of Iran's Sanjesh Organization, a subsidiary of Iranian Ministry of Science,

Research and Technology. Exploratory factor analysis was used to study the test dimensionality and its factorial structure.

3.3. Data Analysis

Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) revealed that the sample size is good for this study with a value of 0.943. Garson (2002) argued that KMO is acceptable if it is over 0.60. Preacher and MacCallum (2002) argued that factor-to-variables ratio should be no lower than 5; however, in this study the ratio is 34.13. As Figure 1 shows parallel analysis suggests that the number of factors are two. The eigenvalues of each factor is shown in Table 1. It seems logical that noteworthy factors should have eigenvalues greater than 1.0, as Figure 1 shows the third factor has eigenvalues lower than 1.0. Only for two factors the eigenvalues in actual data are greater than the eigenvalues of the resampled data. So, logically two factors are found to be noteworthy.

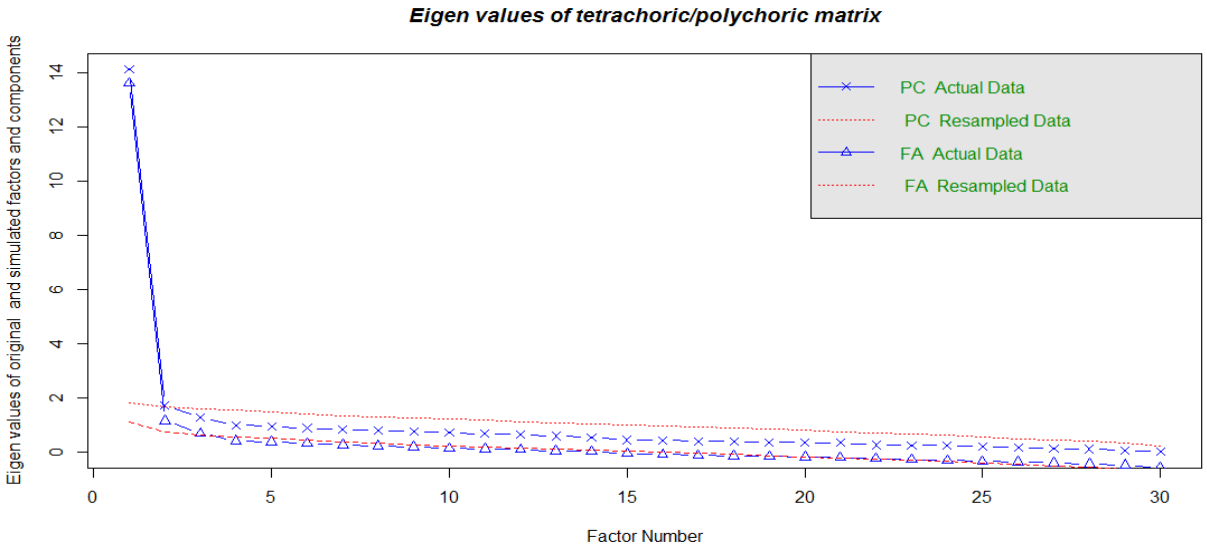


Figure 1: Eigenvalues of tetrachoric matrix

Table 1

Factors' Eigenvalues in parallel analysis

Factors	Eigenvalues
Factor 1	13.61
Factor 2	1.16
Factor 3	0.71

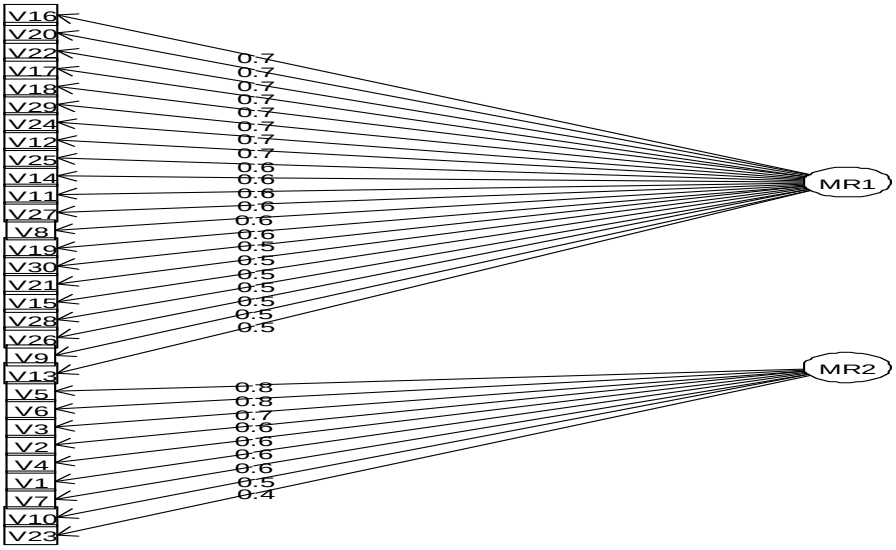
The number of factors identified by the parallel analyses was two for our data (N=1024)

.

Exploratory factor analysis with two factors was run on the tetrachoric correlation matrix because the items were dichotomous. Factors loading on the two factors are shown in Table 1. Tucker Lewis Index (TLI) of factoring reliability with 30 items was 0.93. Absolute fit indices including the root mean square of the residuals (RMSEA) indicated the result is reliable. Root mean square of the residuals (RMSR) was 0.05. The RMSEA ranges from 0 up to 1.00, with smaller values indicating better model fit and here 0.05 shows the model is well fit, also a value of .06 or less is indicative of acceptable model fit, so the model is well accepted too.

As Table 1 shows, cloze items mostly loaded on the second factor. Items from 1 up to 10 are cloze test items and items 11 up to 30 are reading comprehension items based on sustained passages. The study indicated that, a reading comprehension test, composed of cloze and sustained passages followed by some questions, is not unidimensional; therefore, this type of test which is currently used in FCE, CAE, CPE, PET, and so many local and global exams may not be valid. Results show that loadings of the cloze items on the second factor are higher than their loadings on the first factor which is the reading comprehension ability factor. Since the results

were confirmed by high reliability of raw alpha, the researcher names this second factor as cloze



method factor.

Figure 1 Items' loading

Table 1

Items Loading on each Factor

Items	Factor1	Factor2	Items	Factor1	Factor2
V1	0.17	0.59	V16	0.73	0.28
V2	0.39	0.65	V17	0.68	0.38
V3	0.23	0.67	V18	0.58	0.34
V4	0.46	0.62	V19	0.55	0.16
V5	0.31	0.77	V20	0.71	0.24
V6	0.29	0.76	V21	0.55	0.44
V7	0.25	0.58	V22	0.69	0.39
V8	0.57	0.25	V23	0.36	0.37
V9	0.52	0.49	V24	0.66	0.18
V10	0.42	0.47	V25	0.63	0.38
V11	0.61	0.45	V26	0.52	0.22

V12	0.65	0.40	V27	0.57	0.45
V13	0.51	0.41	V28	0.52	0.39
V14	0.63	0.30	V29	0.67	0.34
V15	0.54	0.43	V30	0.55	0.25

The first ten items have higher loading on the second factor than the other items, only two items load on the first factor; however, the differences between loadings are not high.

A question arises itself, that whether the second factors influencing the test results could be due to the difficulty or easiness of the cloze items. To answer this question, Figure 2 displays the item difficulty of the whole exam. Figure 2 shows that in X axis the difficulty of all items\questions, both part of reading comprehension test, the first tenth questions and the next 20 questions are almost equally difficult. Y axis says the amount of information each question

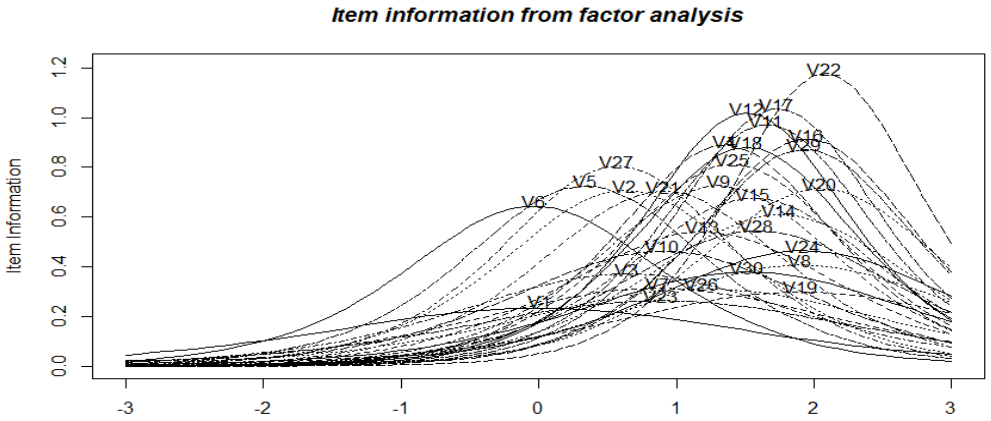


Figure 2 *Items Difficulty*

contains; the more information it has, the harder it become; however, they are almost at the same

Level of information. Matrix for this analysis is based on how much these 1024 participants answered questions correctly.

Table 2

Reliability of Factor 1

Raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	SD
0.84	0.84	0.84	0.27	5.4	0.01	0.13	0.19

The raw alpha reveals that the factor reliability and as shown for the first factor is 0.83 which is upper than 0.70. The results of the second factor analyzing are displayed in Table 3:

Table 3

Reliability of Factor 2

Raw_alpha	Std.alpha	G6(smc)	Average_r	S\N	ase	mean	SD
0.78	0.78	0.77	0.29	3.6	0.015	0.33	0.28

Based on obtained results in Table 3, the raw alpha for the second factor is 0.84.

Findings showed that the combination of cloze test items and reading comprehension items is not unidimensional. The first reason that comes to mind is that it might be an artifact of difficulty, i.e., the reading and cloze item may have different difficulty. However, results

confirmed that both formats have the same difficulty and the cloze is not easier than the other format, so the extracted factors should not be attributed to difficulty. Hambleton's (1991) explanation about unidimensional scores is already mentioned that a test should measure examinee performance on a test as a single ability or trait, it means that the best result from 1024 participants should not have more than one factor on which all items load The second characteristic of unidimensionality is the relationship between examinee performance on each item and the ability measured by the test can be described by a single factor which is the ability of examinee to answer questions. However, in this research, the results clearly confirmed another ability or trait which was measured by reading comprehension test. Cloze items had only two questions loading on the reading comprehension factor, i.e., measuring reading ability, the eight other items measured cloze factor. Therefore, format has an impact on unidimensionality of a test either RC items or cloze test items.

Based on Haladyna (2004), sources of construct irrelevant variance include test wiseness, poorly constructed questions, guessing, item bias, indefensible passing score, testing irregularities. When a test contains flawed items, 'noise' is represented in shape of misleading, and confusing questions that make it harder for the student to answer correctly, even if the student has already mastered the questions' content domain. Flawed questions or items are likely to produce 'false negatives' or they are indication of students who failed the examination but should not have failed. The test wise student can use flaws in the structure of items to get the correct answer without even knowing anything about the content of a questions. These types of questions lead to guessing, and may reduce the chance of picking the correct answer.

Discussion

The main purpose of this study was to investigate whether a reading comprehension test composed of cloze item and MC questions based on sustained passages is unidimensional or not. Findings revealed that the cloze test created irrelevant variance which was different from the reading variance created by the sustained passages and cloze items combined. The nature of this auxiliary construct is not known. It might be called the ability to handle cloze format, a sort of verbal intelligence (Baghaei, & Tabatabaee, 2015) or simply local dependence (Eckes & Baghaei, 2015; Baghaei, 2010). Future studies should investigate the nature of this nuisance dimension using confirmatory factor analysis or multidimensional item response theory models (Baghaei, 2012; Baghaei, 2013).

Construct-irrelevant variance of cloze test is extremely high, UEE language exam of 2011 reading section was used, the test design was the same as the other tests administered in the previous years. Based on Manhattan reviews, Educational Testing Service (ETS) has revised the TOEFL a numerous of time since the exam was first created in 1964. These changings meant to make the TOEFL more valid and also minimize the construct-irrelevant variance. One of the recent changes is deleting ineffective sentence completion exercises and replacing them with assessments of written English that required understanding of context. These changes make TOEFL (IBT) to be considered as one of the most standardized tests, and international students benefit from an exam that much more effectively demonstrates their ability in English-speaking world. To see all changes in reading section of TOEFL you may need to consult the Manhattan review for TOEFL changes.

The theoretical framework for this study was twofold, first of all it was a research in the areas of language testing and reading. In the zone of language testing, Bachman's (1990) language ability model (later revised in Bachman and Palmer, 1996) was the main focus. The researchers mainly

focused on Bachman discussions, 'method facets' as well as 'trait facets', also they tried to draw attentions to a range of factors that can affect test performance and as a result, jeopardize validity of the test. According to Bachman (1996), method facets can be divided into five main categories; namely test rubrics, testing environment, the nature of input, the nature of the expected response and, the interaction between input and response. This research focused on the nature of input and the nature of expected response by manipulating test format and organization that plays a significant role. Test methods are unwanted variance associated with the way observations were made rather than with the construct of interest (Maul, 2013). Therefore, method variance is an instance of construct irrelevant variance (Messick, 1989) and a source of test invalidity. Bachman (1990, p. 156) stated that "When test performance is unduly affected by factors other than the ability being measured, the meaningfulness or validity of score interpretations is lessened." This study suggests that examination of the structure of reading comprehension test batteries composed of different test methods, especially cloze format, using a latent trait model is an important steps in validation (Baghaei & Tabatabee Yazdi, 2016).

The result of this research contains an important contribution to testing practice because the number of studies which have examined the effect of factors on examinees seems to be limited, especially in the second language. The findings suggest that examiners of second language ought not to choose formats of comprehension tests simply because of assuming that all formats measure the same trait or ability, they must be aware of the potential effects of test formats on examinees' performance. In this way, test scores will be more reliable and meaningful. It is important for teachers and examiners to identify the exact nature of different test formats. Each format may measure an ability not totally related to language ability. Furthermore, the result of this research stated that different test formats seem to measure different aspects of reading

comprehension. This has been already stated in the literature and also confirmed by the present study. The findings are contributing to the test development process because it clarified the complex nature of reading comprehension questions. Syllabus designers should also make a revision of reading exercises they have put in syllabi. It is highly recommended for language testers to conduct in-depth qualitative analysis of test items and questions for expert judgment in examinees' scores. This is particularly very urgent in contexts where examination results have very strong impact on test takers' lives for example, university entrance exams or even CPE. Test results are used to make major decisions for educational purposes, they play an important role in many applied linguistics research projects. The level of examinees' proficiency should be discussed here, since our data has good variety in proficiency level, the researcher can conclude the results are more reliable than other tests. Because examinees in other tests, like CAE, are almost upper-intermediate, so the result may not be valid enough to generalize. This research has demonstrated that there is another factor in students' test performance. It is, therefore, very important for language testers and teachers, or anyone involved in the assessment, to pay attention to the test methods they use.

References

Alderson, J. C., & Urquhart, A. H. (1985). Thy: test is unfair: I'm not an economist. *Second language performance testing.* Ottawa: University of Ottawa Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Baghaei, P. (2009). *Understanding the Rasch model*. Mashhad, Islamic Azad University.

Baghaei, P., & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional Rasch model. *International Journal of Testing*, *15*(1), 71-87.

Baghaei, P. (2016). Modeling multidimensionality in foreign language comprehension tests: An Iranian example. In Aryadoust, V., & Fox, J. (Eds.), *Trends in Language Assessment Research and Practice: The View from the Middle East and the Pacific Rim* (pp.47-66). New Castle, England: Cambridge Scholars.

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. & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional Rasch model. *International Journal of Testing, 15,* 71-87. DOI: 10.1080/15305058.2014.941108.

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology, 10,* 233-252.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling, 52,* 313-322.

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica, 37,* 85-104.

Baghaei, P., & Tabatabaee, M. (2015). The C-Test: An integrative measure of crystallized intelligence. *Journal of Intelligence, 3,* 46-58. Available: http://www.mdpi.com/2079-3200/3/2/46

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of marketing research*, *38*(2), 143-156.

Balow, I. H., & Ruddell, A. K. (1963). The effects of three types of grouping on achievement. *California Journal of Educational Research*, *14*, 108-117.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, *10*(5), 291-299.

Bormuth, J. R. (1968). The cloze readability procedure. *Elementary English*, *45*(4), 429-436.

Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of educational measurement*, *5*(3), 189-196. Page, W. D. (1975). Help for the Reading Teacher: New Directions in Research.

Brütsch, S. M. (1979). Convergent/discriminant validation of prospective teacher proficiency in oral and written French by means of the MLA cooperative language proficiency tests, French direct proficiency tests for teachers (TOP and TWP), and self-ratings. *Unpublished Ph. D. dissertation, University of Minnesota*.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Chen, P. Y., & Spector, P. E. (1991). Negative affectivity as the underlying cause of correlations between stressors and strains. *Journal of applied psychology*, *76*(3), 398.

Clifford, G. J. (1987). A Sisyphean Task: Historical Perspectives on the Relationship between Writing and Reading Instruction. Technical Report No. 7.

Cote, J. A., & Buckley, M. R. (1988). Measurement error and theory testing in consumer research: An illustration of the importance of construct validation. *Journal of consumer research*, *14*(4), 579-582.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, *3*(3), 265-286.

Daane, M. C. (2005). *The Nation's Report Card: Fourth-grade students reading aloud: NAEP 2002 special study of oral reading*. National Center for Education Statistics.

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-Tests. *Applied Measurement in Education, 28,* 1-14. DOI: 10.1080/08957347.2014.1002919.

Doty, D. H., & Glick, W. H. (1998). Common methods bias: does common methods variance really bias results?. *Organizational research methods*, *1*(4), 374-406.

Gilliland, J. (1972). *Readability*. University of London Press for the United Kingdom Reading Association.

Greene, B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading*, *24*(1), 82-98.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*(1), 19.

Messick, D. M., & Mackie, D. M. (1989). Intergroup relations. *Annual review of psychology*, *40*(1), 45-81.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*(3), 359-370.

Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. *Children's reading comprehension and assessment*, 369-394.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education*, *5*(1), 1-16.

Jenkinson, M. D. (1957). *Selected processes and difficulties of reading comprehension*. Department of Photoduplication, University of Chicago Library.

Kastner, M., & Stangla, B. (2010). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia-Social and Behavioral Sciences*, *12*, 263-273.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281-300.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879.

Rankin, E. F. (1977). Sequence strategies for teaching reading comprehension with the cloze procedure. In *Reading: Theory, Research and Practice. Twenty-sixth Yearbook of the National Reading Conference. Clemson, SC: The National Reading Conference* (pp. 185-88).

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*(4), 354-379.

Siemsen, E., Roth, A., & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational research methods*, *13*(3), 456-476.

Sharma, M., Purdy, S. C., & Kelly, A. S. (2009). Comorbidity of auditory processing, language, and reading disorders. *Journal of Speech, Language, and Hearing Research*, *52*(3), 706-722.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language testing*, *1*(2), 147-170.

Spector, P. E., & Brannick, M. T. (2009). Common method variance or measurement bias? The problem and possible solutions. *The Sage handbook of organizational research methods*, 346-362.

Taillefer, G., & Pugh, T. (1998). Strategies for professional reading in L1 and L2. *Journal of Research in Reading*, *21*(2), 96-108.

Ulijn, J. M., & Strother, J. B. (1995). *Communicating in business and technology: From psycholinguistic theory to international practice*. Peter Lang Pub Incorporated.

Van der Heijden, P. G. M., Van Buuren, S., Fekkes, M., Radder, J., & Verrips, E. (2003). Unidimensionality and reliability under Mokken scaling of the Dutch language version of the SF-36. *Quality of Life Research*, *12*(2), 189-198.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2),

Weaver, W. W., & Kingston, A. J. (1963). A factor analysis of the cloze procedure and other measures of reading and language ability. *Journal of Communication*, *13*(4), 252-261.