Original Research

SAGE Open
January-March 2019: 1–8
© The Author(s) 2019
DOI: 10.1177/2158244019832706
journals.sagepub.com/home/sgo
SAGE

Method Bias in Cloze Tests as Reading Comprehension Measures

Purya Baghaei[1] and Hamdollah Ravand[2]

Abstract

In many reading comprehension tests, different test formats are employed. Two commonly used test formats to measure reading comprehension are sustained passages followed by some questions and cloze items. Individual differences in handling test format peculiarities could constitute a source of score variance. In this study, a bifactor Rasch model is applied to separate the cloze-specific variance in a reading comprehension test composed of sustained passages (plus questions) and a cloze passage. The results are compared with a unidimensional Rasch model where all items load on a single dimension. The inclusion of the cloze-specific dimension, that is, the method factor, improved the fit and resulted in substantially lower item difficulty estimates for the cloze items. Findings indicate that reading comprehension tests comprising sustained passages and cloze items are not unidimensional and contain a cloze-specific nuisance dimension that contaminates the latent construct variance.

Keywords

cloze test, reading comprehension test, dimensionality, method bias

Introduction

Mixed-format items are gaining increasing popularity in high-stakes testing. The term *mixed format* has commonly been associated with tests that are composed of multiple choice (MC) and constructed response (CR) items. However, there are other combinations of tests that constitute mixed-format assessments. For instance, a combination of MC reading comprehension items and MC cloze tests, which are used in some high-stakes tests (e.g., the Examination for the Certificate of Proficiency in English [ECPE] and the tests prepared by the English Department of Cambridge University, such as Key English Test [KET], Preliminary English Test [PET]), is an instance of mixed format tests. It is believed that tests composed of multimethod items are psychometrically more advantageous (Wang, Drasgow, & Liu, 2016). In tests, where multimethod items are employed, it is believed that items of different formats complement each other. The depth of the knowledge tested by CR items, for example, can complement the breadth of knowledge covered by MC items. However, this advantage might come at a price. Use of mixed format items within the same test intended to be a measure of a single construct might lead to multidimensionality.

A key assumption in educational measurement is unidimensionality. Unidimensionality provides validity evidence for an instrument. A set of items on an instrument intended to measure any given construct should reflect individual difference on the levels of just that construct. Therefore, the dimension should exhaust all the covariation among the indicators and renders them independent conditionally (Baghaei & Tabatabaee-Yazdi, 2016). Otherwise the instrument lacks construct validity, and there are other confounding variables that add construct-irrelevant variance to the target dimension intended to be measured by the instrument. This assumption is referred to as the *local independence* assumption in the latent variable literature and is equivalent to *unidimensionality* (Baghaei, 2007, 2010).

In psychological measurement, unidimensionality is implicit when persons' performances are compared based on a variable, test takers are classified into groups based on an attribute, or individual differences are expressed in terms of a variable. All construct validation attempts, be they through item response theory (IRT), or confirmatory factor analysis (CFA), one way or another, boil down to unidimensionality checks.

There might be serious problems arising from the violation of unidimensionality especially when test takers are allowed to select the subset of items that they answer. When

[1] English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran
[2] Vali-e-Asr University of Rafsanjan, Iran

Corresponding Author:
Purya Baghaei, English Department, Mashhad Branch, Islamic Azad University, Ostad Yusofi St., Mashhad 91871, Iran.
Email: pbaghaei@mshdiau.ac.ir

IRT models are used to score and equate test forms that are comprised of hybrid items, violation of the unidimensionality assumption might lead to thorny problems (Thissen, Wainer, & Wang, 1994). If the subsets represent different dimensions, comparison of the performance of the test takers would be misleading.

Sources of Multidimensionality

Bachman (1990) identified test method as a source of variance in test performance. According to Bachman, if test method is not part of the definition of the construct under study, it is considered a nuisance or, or in terms of Messick (1989), a construct-irrelevant factor. Test methods have been shown to form new dimensions alongside the construct being studied (Baghaei & Aryadoust, 2015). According to Yen (1993), factors such as test speededness, test method, common stimuli (e.g., passage), raters, and test rubrics could lead to multidimensionality and, hence, threaten construct validity of tests.

Rauch and Hartig (2010) found that a two-dimensional IRT model fits a reading comprehension test composed of mixed formats of MC and CR better than a unidimensional model. They found the both MC and CR items loaded on a proficiency factor common to the items of both format (which they interpreted as abilities necessary to master basic reading processes that are needed for solving both MC and CR items) whereas CR items also loaded on a latent dimension that represented a proficiency aspect unique to CR items. Although they did not rule out other possible sources of multidimensionality such as test-taking strategies, they suggested that the second dimension, measured only with CR items, tests abilities necessary to master higher reading processes. The same problem might occur with mixed format items composed of MC reading and MC cloze items. The two test items, according to Bachman's (1990) framework of test method facets, are the same in terms of "nature of expected response" but different in terms of the "nature of input." Research (Mizumoto, Ikeda, & Takeuchi, 2016) found that cloze tests require greater cognitive processing than MC items of reading comprehension. Along the same lines, Raymond (1988) argued that cloze tests require greater levels of language awareness for the reconstruction of meaning than normal reading tasks. In the present study, it is hypothesized that the broader range of cognitive resources targeted in cloze tests may be a source of multidimensionality. The greater level of language awareness or greater cognitive processing demanded by cloze tests than what is required in MC reading comprehension test performance is what Marais and Andrich (2008) called trait dependence, which they argued leads to trait multidimensionality. Marais and Andrich argued for another type of dependence in cloze tests, which they called response dependence or item chaining effect, which is a method multidimensionality. Due to these dependencies among cloze items, mixed-format items that include cloze items are hypothesized to be multidimensional.

Reading comprehension is commonly measured through reading comprehension passages followed by a set of questions that check test takers' understanding of the passage. Although there have been long debates over what cloze tests measure (Baghaei & Ravand, 2016), they have been used as measures of reading comprehension. Studies have shown that cloze tests need text-level understanding, and hence, they measure reading comprehension (e.g., Bachman, 1985; Chavez-Oller, Chihara, Weaver, & Oller, 1985; Chihara, Oller, Weaver, & Chavez-Oller, 1977; Jonz, 1990; McKenna & Layton, 1990).

Famous high-stakes language proficiency tests such as the ECPE, which is developed and administered by the University of Michigan, and the tests prepared by the English Department of Cambridge University, such as KET, PET, First Certificate in English, and so on, scale reading comprehension items of hybrid formats, that is, text reading and cloze items together. National tests such as university entrance examinations in Iran mix MC cloze items with passage comprehension items to measure reading comprehension.

The assumption in all the tests that employ mixed-item formats is unidimensionality. Dimensionality of a test has implications for score-reporting strategies for the test. The *structural* aspect of construct validity (Messick, 1989) requires that the score-reporting policy for each test should match the structure of the construct measured by the test. According to Loevinger (1957), *structural fidelity* requires that the scoring model of a test be guided by what is known about the internal structure of the construct measured by the test. If the construct is shown to be multidimensional, multiple scores, rather than a single score, should be reported for the test.

Studies Exploring Dimensionality of Hybrid Tests

The literature on the dimensionality of hybrid reading comprehension items comprising MC and CR forms abounds (e.g., Rauch & Hartig, 2010). Studies comparing trait equivalence of MC and CR are mainly of two types (Barati, Ravand, & Ghasemi, 2013): (a) studies comparing the relative difficulty of the two formats (e.g., In'nami & Koizumi, 2009; Shohamy, 1984) and (b) correlational studies investigating the association between the two formats (e.g., Hancock, 1994; Rodriguez, 2003; Thissen et al., 1994). The results of these studies are equivocal. Rauch and Hartig (2010) found that reading ability measured simultaneously with MC and CR items could be described more adequately with a two-dimensional IRT model than with a unidimensional model. However, Rodriguez (2003) in a meta-analysis of the studies investigating the trait equivalence of MC and CR items found that when stems are equivalent across the two formats, the correlations tend to be significantly higher than when the stems are not equivalent. Hollingworth, Beard, and Proctor (2007), examining the construct equivalence of MC and CR items, found that a two-factor solution for MC and CR items was satisfactory. Thissen et al. (1994) replicating Bennett (1993) found a two-factor solution for the MC and CR items, whereas in the original study, Bennet had found a one-factor solution more parsimonious for the College Board's Advanced Placement test in Computer Science. However, the studies on the dimensionality of a combination of sustained passages and cloze items of reading comprehension are scanty.

Current Study

The present study intends to investigate the dimensionality of mixed passage comprehension and cloze items of reading comprehension using a Rasch-model-based bifactor model. In this model, MC comprehension questions based on sustained passages are forced to load on a reading factor, and the cloze items are forced to load simultaneously on the reading factor and a method-specific dimension, which is a cloze factor.

The results of this study would shed light on the justifiability or unjustifiability of the current practice in some high stakes national and international tests in scaling cloze and passage-based items of reading comprehension together and reporting a single score to reflect individual differences on reading comprehension. If findings indicate that the combination of sustained passage items and cloze items is not unidimensional, and there is a cloze method factor, then combining these two test types and reporting a single reading score for the examinees on the combination of both tests is not justified.

Method

Instrument and Participants

Participants of the study were a subsample of the Iranian National University Entrance Examination (INUEE) candidates (N = 1,024, 68% females) who applied for the undergraduate English programs in state universities in 2011. INUEE is a high-stakes test that screens the applicants into English Studies programs at state-run universities in Iran. The INUEE measures general English proficiency at an intermediate level. The test consisted of four sections: the grammar section containing 15 items; the vocabulary section containing 15 items; the language function section with 10 items, where examinees have to read short independent conversations (two to four lines) and fill a few gaps; and the reading comprehension section with 30 items. The reading comprehension section has two subparts. The first subpart contained a cloze passage with 10 multiple-choice items, and the second comprised three sustained passages followed by several questions. The four passages in the reading comprehension section were academic texts, and the questions were four-option MC. Time for completing the whole test with 70 items was 105 min. For the purpose of the present study, only the two subparts of the reading comprehension section were used.

Data Analyses

Unidimensional and multidimensional Rasch models (Rasch, 1960/1980) were employed to analyze the data. The multidimensional Rasch model employed in this study was the multidimensional random coefficient multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997). MRCMLM is an extension of the unidimensional Rasch model and enjoys the measurement properties of these models including separability and existence of a sufficient statistic for parameter estimation (Baghaei, 2012).

More specifically, the bifactor model was employed to analyze the data in this study. Bifactor modeling is an excellent choice to improve model fit and clean the score variance of nuisances. In a bifactor model, variance is decomposed into general and specific variance. In bifactor modeling, there is a general factor on which all items load and some specific factors that represent the unique variance that is not included in the general factor. The general and specific factors are defined to be orthogonal. What is left in the specific factors, after partialing out the effects of the general factor, is uniqueness of the factor. Bifactor modeling is an excellent approach to examine the specific variance that a group of items might share that is not included in the general factor.

A unidimensional Rasch model and a bifactor model were fitted to the data. In the unidimensional model, all the 30 items (the 10 cloze items and the 20 passage comprehension items) were analyzed together and were modeled to load on a single reading comprehension dimension (Figure 1, right). In the bifactor model, all the 30 items were modeled to load on a reading comprehension dimension, with the 10 cloze items simultaneously loading on a cloze-specific dimension (Figure 1, left). Since the cloze-specific dimension is a construct-irrelevant nuisance dimension, it was set to be orthogonal to the target reading ability dimension. By forcing the cloze items to load both on a general reading ability dimension and a specific cloze dimension, we aimed to partition their variance into two components, that is, variance due to the general reading ability dimension and the variance that represents the uniqueness of the cloze items (Baghaei & Aryadoust, 2015). The bifactor model cleans the data from domain-specific peculiarities and provides error-free estimates (Baghaei, 2016). The goal of this study is to decompose a mixed-item reading comprehension test variance into its components, that is, reading comprehension and cloze-specific variance.

ConQuest software package version 2.0 (Wu, Adams, Wilson, & Haldane, 2007) in which the marginal maximum

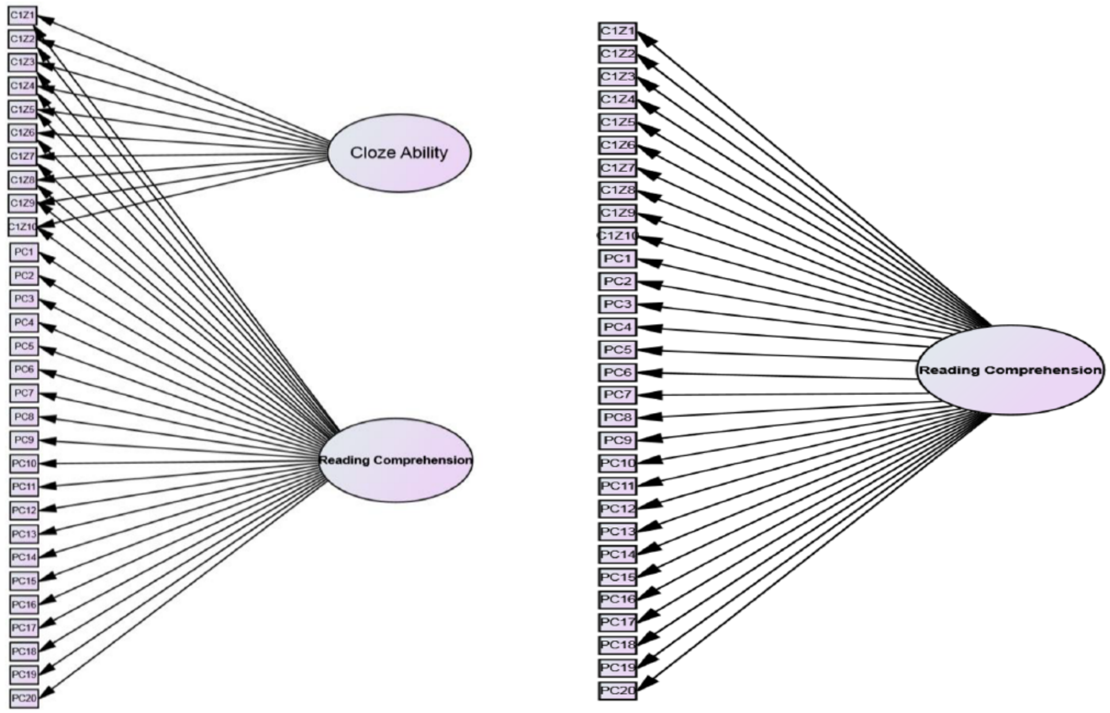

Figure 1. Graphical representation of unidimensional (right) and bifactor models (left).
Note. Clz. = cloze; PC = passage comprehension.

Table 1. Global Model Fit Values in the Two Models.

Model	No. of parameters	G^2	AIC	CAIC	BIC	EAP Rel.	R (PE)	R (IE)
Unidimensioanl	31	22523.82	22585.82	22769.65	22738.65	.85	–2.36-4.46	–.03-3.92
Bifactor	32	22363.86	22427.86	22617.62	22585.62	.81	–2.21-4.62	–.01-3.99

Note. G^2 = deviance; AIC = Akaike's Information Criterion; CAIC = consistent Akaike's Information Criterion; BIC = Bayesian Information Criterion; EAP Rel. = expected a posteriori reliability; R (PE) = person estimates' range; R (IE.) = item estimates' range.

likelihood estimation method (Bock & Aitkin, 1981) is implemented was used to estimate the models. The fits of the unidimensional and the bifactor model were compared to determine the optimal model for the structure of the reading comprehension test. The goodness of fit of the two models were compared with their deviances (–2 log-likelihood) and information criteria. After estimating the item difficulty and person-ability parameters, the log-likelihood of each response pattern was computed and summed across examinees. A high log-likelihood or lower −2 log-likelihood index (deviance) indicates better model fit to the data. In nested models, "where one model is a more constrained version of the other model" (DeMars, 2012, p. 106), the difference between −2 log-likelihoods of the two models should be approximately distributed as chi-square with degrees of freedom equal to the difference in the number of estimated parameters in the two models. Non-nested models are compared with information criteria such as Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). These statistics impose penalties for sample size and the number of parameters estimated.

Results

Table 1 presents the fit statistics for the two estimated models. The likelihood deviance (G^2, −2 log-likelihood) and information criteria, AIC, CAIC (Consistent Akaike's Information Criterion), and BIC indicate that the bifactor model in which the cloze-specific dimension is factored out has a significantly better fit than the standard unidimensional model where all items load on a single dimension.

The difference in the deviances of the two nested models is statistically significant indicating a better fit for the

Table 2. Difficulty Estimates, Standard Errors, and Infit Mean Square Values in the Two Models.

	Unidimensional		Bifactor	
Item	Est. (*SE*)	Infit MNSQ	Est. (*SE*)	Infit MNSQ
Clz. 1	0.04 (.07)	1.18	0.03 (.04)	1.17
Clz. 2	1.09 (.08)	0.98	0.59 (.04)	0.98
Clz. 3	1.28 (.08)	1.10	0.69 (.04)	1.09
Clz. 4	2.20 (.09)	0.92	1.18 (.04)	0.95
Clz. 5	0.61 (.07)	0.96	0.33 (.04)	0.92
Clz. 6	–0.03 (.07)	0.96	–0.01 (.04)	0.97
Clz. 7	1.76 (.08)	1.14	0.94 (.04)	1.15
Clz. 8	3.71 (.31)	1.02	1.97 (.07)	1.10
Clz. 9	2.24 (.09)	0.97	1.20 (.04)	1.04
Clz. 10	1.71 (.08)	1.07	0.92 (.04)	1.13
PC 1	2.63 (.10)	0.91	2.68 (.10)	0.93
PC2	2.37 (.09)	0.92	2.41 (.09)	0.92
PC 3	2.19 (.09)	1.04	2.23 (.09)	1.04
PC 4	3.14 (.11)	0.99	3.19 (.11)	0.99
PC 5	2.73 (.10)	0.99	2.78 (.10)	0.99
PC 6	3.15 (.11)	0.92	3.21 (.11)	0.91
PC 7	2.70 (.10)	0.92	2.75 (.10)	0.90
PC 8	2.46 (.10)	0.95	2.51 (.10)	0.94
PC 9	3.91 (.14)	1.08	3.98 (.15)	1.07
PC 10	3.54 (.13)	0.97	3.60 (.13)	0.94
PC 11	1.57 (.08)	0.97	1.60 (.08)	1.00
PC 12	3.12 (.11)	0.88	3.18 (.11)	0.88
PC 13	1.85 (.08)	1.16	1.88 (.09)	1.19
PC 14	3.67 (.13)	1.03	3.73 (.13)	1.01
PC 15	2.35 (.09)	0.94	2.39 (.09)	0.94
PC 16	2.40 (.09)	1.13	2.45 (.09)	1.11
PC 17	0.97 (.07)	0.96	0.98 (.08)	0.98
PC 18	2.91 (.11)	1.02	2.97 (.11)	1.02
PC 19	3.16 (.11)	0.95	3.22 (.11)	0.93
PC 20	2.98 (.11)	1.07	3.03 (.11)	1.07

Note. Est. = estimate; *SE* = standard error; MNSQ = mean square; Clz. = cloze; PC = passage comprehension.

bifactor model, χ^2 (1) =159.96, $p < .001$. The variance of the reading dimension in the unidimensional model was 2.95 and in the bifactor model was 3.10. The variance and the reliability of the cloze-specific dimension in the bifactor model were .94 and .40, respectively. The higher reliability of the reading dimension in the unidimensional model is due to the local dependence among the cloze items, which has spuriously inflated reliability (Eckes & Baghaei, 2015).

Table 2 shows the item difficulty estimates, their standard errors, and their infit values across the two models. Item difficulty estimates obtained from the bifactor model (M = 2.09, SD = 1.15), where cloze-specific variance is modeled and conditioned out, correlated at 0.91 with those obtained from the standard unidimensional model (M = 2.28, SD = 1.03), indicating that the two models yielded similar item difficulty parameter estimates. Absolute differences between the item parameters from the two analyses ranged from 0.01 to 1.73 logit with a mean of 0.26.

Person-ability parameters obtained from the bifactor model (M = 0.00, SD = 1.59) correlated at 0.987 with the ability parameters obtained from the standard unidimensional model (M = 0.00, SD = 1.58). Absolute differences between the person parameters from the two analyses ranged between 0 and 0.76 logit (M = 0.20).

Discussion

It is commonly known that individual differences in handling test format peculiarities could constitute a source of construct irrelevant variance (Bachman, 1990). In this study, we aimed to separate the cloze-specific irrelevant variance in a reading comprehension test using a multidimensional Rasch model. We separately fitted a unidimensional Rasch model and a bifactor Rasch model, in which the cloze items loaded on a method-specific dimension while simultaneously loading on the general reading dimension, to a reading comprehension test composed of MC questions based on sustained passages and MC cloze items. Findings showed that the bifactor model fitted the data significantly better than the unidimensional model. This was an indication that the cloze items share variance above and beyond the general factor. In other words, the cloze items produce variance that is not explained by the target reading comprehension dimension.

In bifactor models, if a specific factor does not contain something unique and only reflects the general factor, it ceases to exist after controlling for the general factor. This is reflected by low variance and reliability estimates for the specific factor (Baghaei, 2016). In this study, the variance and reliability of the cloze-specific dimension were big enough to make it a nonignorable dimension. The better fit of the bifactor model and the relatively large variance and reliability of the cloze-specific dimension indicate that the structure of the reading comprehension test, composed of sustained passages and cloze passages, is not unidimensional, as not all the variance can be explained by a single reading dimension. That is, the cloze-specific dimension has unique variance beyond and above the target reading comprehension construct and forms a dimension separate from the main reading dimension. In other words, the cloze items share something above and beyond the primary reading comprehension construct, and the variance in the reading comprehension test cannot be explained by a single dimension.

There are two possible reasons for the large variance and reliability of the cloze-specific dimension: (a) method bias and (b) the level of cognitive processing targeted by cloze test items. The variance of the specific factor shows the size of the method bias. In the context of this research, the cloze-specific dimension could indicate individuals' facility in handling cloze items. Cloze and its variations have traditionally been used as measures of intelligence (Binet & Simon, 1905; Ebbinghaus, 1897; Spearman, 1927). Recently, other researchers have suggested cloze and its variations as measures of crystallized intelligence (Ackerman, Beier, & Bowen, 2000; Baghaei & Tabatabaee-Yazdi, 2015; Schipolowski, Wilhelm, & Schroeders, 2014). Therefore, reading comprehension tests composed of cloze tests could measure intelligence as well, which is a nuisance dimension in a reading comprehension test. Through bifactor modeling, the irrelevant specific variance of the cloze items can be separated and improve the model fit and by implication improve the validity of the reading comprehension test. This construct-irrelevant uniqueness is ignored in the unidimensional model and becomes part of the reliable variance-inflating reliability. The uniqueness of the cloze items could be anything like intelligence, closure speed, or closure flexibility, which are irrelevant to the reading comprehension construct.

Another possibility is that the unique cloze-specific dimension might reflect greater cognitive processing or language awareness compared with what is demanded by MC reading comprehension items. Although test-taking strategies and testwiseness cannot be ruled out as a possible source of multidimensionality, as Rauch and Hartig (2010) argued, comparing MC and CR reading comprehension items, the second dimension, which is uniquely measured by cloze items, may test abilities necessary to master higher order reading processes, whereas the first dimension, which is common to both types of items, measures basic reading processes.

Currently, it is not possible to extricate the effect of the span of cognitive processings required by different tests methods from testwiseness and test-taking strategies. Most of the studies exploring dimensionality of mixed-item tests have attributed emergence of new dimensions to test-taking strategies rather than to the differences in the span of cognitive processes targeted by the test methods. It should be emphasized that test-taking strategies, testwiseness, and differences in the span of the cognitive processes contribute to construct irrelevant variance. The irrelevant cloze-specific variance might alter the correlation between reading comprehension and external criteria, affecting the validity of the reading comprehension measure (Danner, Aichholzer, & Rammstedt, 2015). Therefore, it is crucial to model method-specific irrelevant variance in reading comprehension tests containing cloze items.

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.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2000). Explorations of crystallized intelligence: Completion tests, cloze tests, and knowledge. *Learning and Individual Differences*, *12*, 105-121.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*, 535-556. doi:10.2307/3586277

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

Baghaei, P. (2007). Local dependency and Rasch measures. *Rasch Measurement Transactions*, *21*, 1105-1106.

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

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. (2016). Modeling multidimensionality in foreign language comprehension tests: An Iranian example. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 47-66). Newcastle upon Tyne, UK: Cambridge Scholars.

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

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.

Barati, H., Ravand, H., & Ghasemi, V. (2013). Investigating relationships among test takers' characteristics and response formats in a reading comprehension test: A Structural Equation Modeling Approach. *Iranian Journal of Language Testing*, *3*, 38-59.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum.

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux [New methods for the diagnosis of the intellectual level of the abnormal]. *L'Annee Psychologique*, *11*, 199-244.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W., Jr (1985). When are cloze items sensitive to constraints across sentences? *Language Learning*, *35*, 181-206. doi:10.1111/j.1467-1770.1985.tb01024.x

Chihara, T., Oller, J. W., Jr., Weaver, K. A., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning*, *27*, 63-70. doi:10.1111/j.1467-1770.1977.tb00292.x

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119-130. doi:10.1016/j.jrp.2015.05.004

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36,* 104-121.

Ebbinghaus, H. (1897). Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihrer Anwendung bei Schulkinder [On a new method for testing mental abilities and their application to school children]. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, *13*, 401-459.

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

Hancock, C. R. (Ed.). (1994). *Teaching, testing, and assessing: Making the connection* (Northeast Conference Reports). Lincolnwood, IL: National Textbook.

Hollingworth, L., Beard, J. J., & Proctor, T. P. (2007). An investigation of item type in a standards-based assessment. *Practical Assessment Research and Evaluation*, *12*, 1-13.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*, 219-244.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly*, *24*, 61-83. doi:10.2307/3586852

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635-694.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*, 200-215.

McKenna, M. C., & Layton, K. (1990). Concurrent validity of cloze as a measure of intersentential comprehension. *Journal of Educational Psychology*, *82*, 372-377. doi:10.1037//0022-0663.82.2.372

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Mizumoto, A., Ikeda, M., & Takeuchi, O. (2016). A comparison of cognitive processing during cloze and multiple-choice reading tests using brain activation. *ARELE: Annual Review of English Language Education in Japan, 27*, 65-80.

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

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*, 354-379.

Raymond, P. (1988). Cloze procedure in the teaching of reading. *TESL Canada Journal*, *6*, 91-97.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*, 163-184.

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156-168.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

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

Spearman, C. (1927). *The nature of "intelligence" and the principles of cognition*. New York, NY: MacMillan.

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *13*, 113-123.

Wang, W., Drasgow, F., & Liu, L. (2016). Classification accuracy of mixed format tests: A bi-factor item response theory approach. *Frontiers in Psychology*, *7*, 1-11.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2: Generalized item response modeling software*. Camberwell: Australian Council for Educational Research.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213. doi:10.1111/j.1745-3984.1993.tb00423.x

Author Biographies

Purya Baghaei is an associate professor in the English Department of Islamic Azad University, Mashhad, Iran. His major research interest is foreign language proficiency testing with a focus on the applications of item response theory models in test validation and scaling. He has also conducted research on the role of cognition in second language acquisition.

Hamdollah Ravand is an assistant professor at Vali-e-Asr University of Rafsanjan. His major areas of interest are Language Testing Assessment, Cognitive Diagnostic Modeling, Structural Equation Modeling, Multilevel Modeling, and Item Response Theory.