THE JOURNAL OF ASIA TEFL
Vol. 15, No. 2, Summer 2018, 479-490
http://dx.doi.org/10.18823/asiatefl.2018.15.2.16.479

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Application and Comparison of Multidimensional Latent Class Item Response Theory on Clustering Items in Comprehension Tests

Masoud Geramipour
Kharazmi University, Tehran, Iran

Niloufar Shahmirzadi
Central Tehran Branch, Islamic Azad University, Tehran, Iran

Introduction

Contemporary test development endeavors often incorporate cognitive psychological theory in order to improve the generation of construct-valid items. The use of "mental models in item generation" (Irvine, 2002, pp. 3–34) might be one of the most significant differences between new theories of test development and older technology. Indeed, it is encouraging from a validity perspective that item and instrument development simultaneously test an empirical model. Although the benefits of incorporating cognitive theory into item generation are numerous, the accompanying challenges for item writers and test developers need to be highlighted. Perhaps the most significant new burden is the development and verification of a cognitive model which can be met through having valid measurement.

The desirable goal of measurement is standardized valid measurement, which can rely on the outcome to measure a latent construct. From the 1970s to the 1980s, validity was defined in four levels, among which construct validity was valued since it can measure the latent or psychological traits of a construct. However, according to Cronbach (1971), "The phrase validation of a test is a source of much misunderstanding. One validates, not a test, but an interpretation of data arising from a specified procedure" (p. 447). Taking into account the importance of construct validity, some scholars have carried out research to observe the degree of their validity. However, only a few studies have been done by test practitioners and researchers to apply test dimensionality in English language measurement to a number of examinees' characteristics, especially on the national entrance examination. This study also aims to measure two forms of comprehension tests, namely, reading comprehension and cloze tests based on Multidimensional Item Response Theory (MIRT) and Multidimensional Latent Class Item Response Theory (MultiLC IRT).

IRT and MultiLC IRT

According to Ackerman (1994), Hattie (1985), and Reckase (1990), test dimensionality is defined as the interaction of a set of items and underlying constructs of the examinees. According to Traub and Lam (1985, p. 22) "the assumption of unidimensionality seems inappropriate when it refers to an educational achievement." On the other hand, it should be noted that when different dimensions are measured,

decision making is considered a critical matter (Hattie, 1984); therefore, through the application of MIRT, dimensions can accurately measure the underlying traits (Nandakumar, 1991). Also, some scholars like Reise, Waller, and Comrey (2000) are of the opinion that using a number of dimensions are better than merely measuring unidimensional aspects of a test. Accordingly, Ansley and Forsyth (1985), Drasgow and Parsons (1983), and Reckase (1979) believe that using unidimensional models with multidimensional data can be problematic. Thus, test dimensionality shows the meaningfulness and fairness of the test. It can also depict the students' performance on a set of items. According to Hattie (1985), in the past test dimensionality lacked enough support; however, recent theoretical and empirical studies have yielded sufficient support.

MIRT is a kind of "statistical abstraction of data which can identify cognitive and affective variations in test takers" (Reckase, 2009, p. 79). Moreover, by applying MIRT, specifying a model from structural and incidental perspectives is possible. The former shows the parameters that describe the function of the test item, and the latter describes the vector of the coordinates describing the location of the individual (Hambleton & Swamination, 1983).

Reckase (2009) also believes that in the basic latent class (LC) model, the focus is both on the latent variable in the same class, and one or some characteristics of items, while developed LC models hold for different latent traits (multidimensionality) among different classes, and item characteristics (parametrization). As a result, MIRT allows observation of the following three features, namely, multidimensionality, discreteness of latent traits, and ordinal polytomous responses. These features are different in nature, show test takers a set of ability scores, and represent test items with a set of parameters that are related to that item.

Based on previous studies, some researchers have provided a number of examples of loglinear multidimensional IRT models (Agresti, 1993; Duncan & Stenbeck, 1987; Kelderman & Rijkes, 1994). For latent traits studies, Lazarsfeld and Henry (1968), and Goodman (1974) proposed research on a homogenous class of individuals with similar latent traits. Also, Christensen, Bjorner, Kreiner, and Petersen (2002) have outlined studies which were conducted based on a simulation in order to propose computational problems encountered during the estimation process of a multidimensional model that is based on normally multivariate distributed ability. Rost (1991), and von Davier and Rost (1995) also studied a mixture of latent classes with a separate Rasch model. Regarding MultiLC IRT models, Bartolucci (2007) simultaneously considered more latent traits in which each item is associated with merely one trait. Zhang (2004) provided more details on these studies. Similarly, Zhang (2005) used a genetic algorithm to estimate complex parameters of a complex MIRT structure.

Strategy Based Instruction

Strategy Based Instruction (SBI) is an approach that incorporates strategy training with a language learning curriculum (Rubin, Chamot, Harris, & Anderson, 2007). Therefore, by taking the importance of SBI into account, the Typology of Learning Strategy Taxonomy (Oxford, 1990, p. 17) highlighted three direct cognitive strategies such as understanding, problem solving, and reasoning as significant factors in education. More specifically, 'understanding' refers to the ability to understand, 'problem solving' is a supporting ability to deal with a problem, and 'reasoning' means providing choices for understanding national university admissions tests.

Hence, due to the need in the literature to take into account the application of MultiLC IRT and MIRT models to special high–stakes university entrance tests from cognitive factors and language content domains in reading comprehension and cloze tests, the purpose of the study is threefold: measuring content domain, measuring ability factors, and measuring and comparing the appropriate methods in the present study, namely, MIRT and MultiLC IRT. In addition, through having a greater understanding of learning strategies, the results will help develop an awareness of language preferences and biases in reading comprehension and cloze tests.

Research Questions

Q1: To what extent do the MIRT models and the MultiLC IRT models fit reading comprehension and cloze tests?
Q2: Among three cognitive factors, namely, understanding, problem solving, and reasoning, which factor(s) is/are more prominent in reading comprehension and cloze tests?

Materials and Method

Participants

The data of the present study were collected from among national university candidates who were at the undergraduate level (B.A. level) majoring in the English language. The total randomly selected sample was 20,000 test takers who were both males and females. This sample was randomly selected over the course of four years. All participants ranged in age from 18 to 20 and were diploma holders in three fields, namely, mathematics, natural sciences, and social sciences.

Data Collection

The corpus of the study was provided by the National Organization for Educational Testing, which undertakes to administer high–stakes tests at both the undergraduate and graduate levels. It is important to stress that conducting measurement studies and decision making are serious undertakings of this organization.

Due to the importance of maintaining consistency, simple random sampling was applied to the data collection. Over four years of administering high–stakes examinations, 487,992 test takers took the test.

Test Administration Task and Procedures

Each year, this high–stakes university admission test was scheduled to run in July. Test takers were required to answer 95 items which tested various aspects of English language proficiency. The allotted time to complete the whole test was 125 minutes. Test confidentiality was also observed by the organization.

The test content was comprised of senior high school English materials consisting of grammar, vocabulary, sentence structure, language functions, cloze test, and reading comprehension.

In what follows, the procedural analyses of the data are provided for the MIRT and MultiLC IRT methods.

Results

In the present study the R package, which bundles together code, data, documentation, and tests, was applied for data processing.

Three Dimensional MultiLC IRT

Having collected the samples over the four years, the researchers measured and compared the MIRT, and MultiLC IRT in different phases. Tables 1 to 4 depict the results of MultiLC IRT in these consecutive years. Figures 1 to 4 also illustrate MultiLC IRT through dendrograms over the four years. Through applying a dendrogram it is possible for researchers to create and compare visually tree diagrams. It may

also provide utility functions for comparing trees to one another, both statistically and visually.

TABLE 1
MultiLC Outputs for 2008

	items		deviance	df	p-value
Step 1	-4	-15	0.128	2	0.938
Step 2	-3	-11	1.468	4	0.832
Step 3	-2	-10	2.761	6	0.838
Step 4	-8	-9	4.375	8	0.822
Step 5	-1	-5	6.933	10	0.732
Step 6	-6	-12	13.296	12	0.348
Step 7	1	4	21.079	14	0.100
Step 8	-13	-14	30.994	16	0.013
Step 9	7	8	44.577	18	0.000
Step 10	-7	2	59.169	20	0.000
Step 11	5	10	78.562	22	0.000
Step 12	9	11	161.412	24	0.000
Step 13	3	6	251.437	26	0.000
Step 14	12	13	548.942	28	0.000

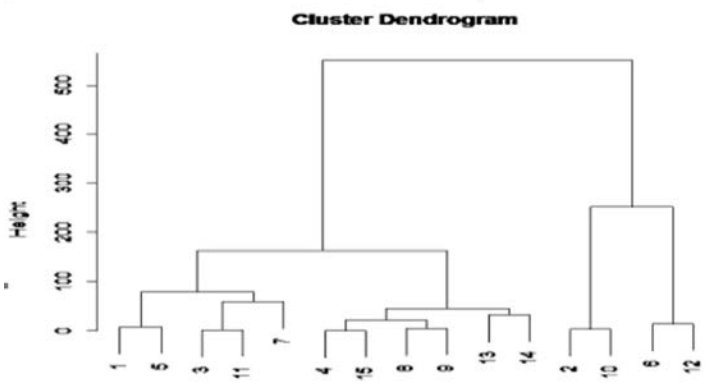

Figure 1. Cluster dendrogram for 2008.

As Figure 1 depicts, items 2, 10, 6, and 12 indicated reasoning, and the rest were devoted to understanding.

TABLE 2
MultiLC Outputs for 2009

	items		deviance	df	p-value
Step 1	-3	-11	0.102	2	0.950
Step 2	-6	-7	0.344	4	0.987
Step 3	-2	-4	1.876	6	0.931
Step 4	-1	-13	3.699	8	0.883
Step 5	-10	3	6.339	10	0.786
Step 6	-5	-14	10.326	12	0.587
Step 7	-12	-15	15.050	14	0.375
Step 8	-8	1	20.442	16	0.201
Step 9	-9	5	26.749	18	0.084
Step 10	2	6	33.632	20	0.029
Step 11	9	10	57.787	22	0.000
Step 12	4	11	116.863	24	0.000
Step 13	7	12	261.213	26	0.000
Step 14	8	13	540.842	28	0.000

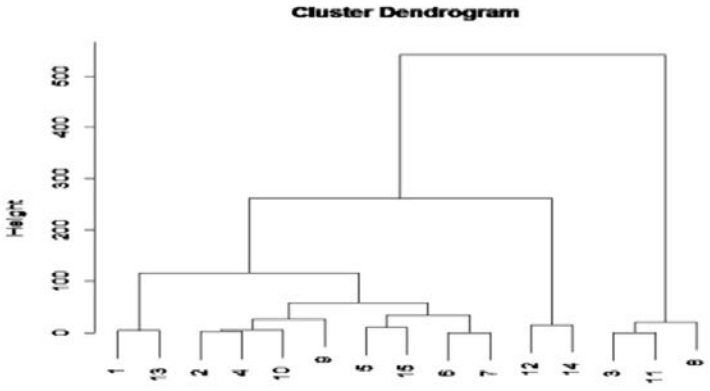

Figure 2. Cluster dendrogram for 2009.

In contrast to Figure 1, in Figure 2 three items including 3, 11, and 8 were categorized under understanding as one of the factors in cognitive complexity, and the majority of items were divided into problem solving and reasoning features.

TABLE 3
MultiLC Outputs for 2010

	items		deviance	df	p-value
Step 1	-9	-10	0.117	2	0.943
Step 2	-3	-6	0.364	4	0.985
Step 3	-2	-14	0.731	6	0.994
Step 4	-13	1	1.451	8	0.993
Step 5	-5	-15	2.281	10	0.994
Step 6	-7	-11	3.431	12	0.992
Step 7	-4	4	4.848	14	0.988
Step 8	-8	2	6.918	16	0.975
Step 9	-1	5	9.315	18	0.952
Step 10	-12	9	17.366	20	0.629
Step 11	3	8	30.524	22	0.106
Step 12	6	10	51.913	24	0.001
Step 13	7	11	79.222	26	0.00
Step 14	12	13	282.192	28	0.00

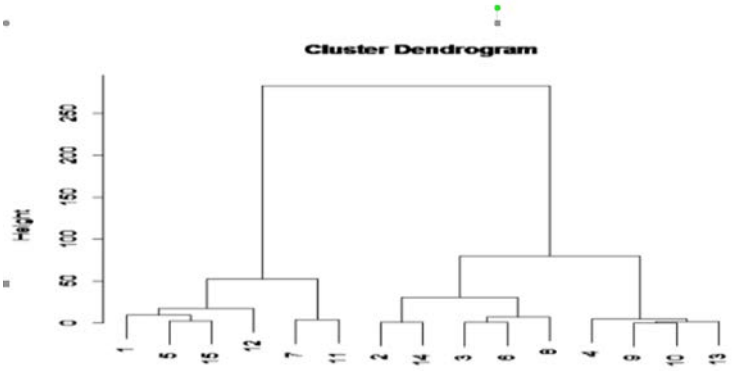

Figure 3. Cluster dendrogram, 2010.

Interestingly, in Figure 3 the results differed, since almost half of the items were related to problem solving, and the rest of the items to reasoning.

TABLE 4
MultiLC Outputs for 2011

	items		deviance	df	p-value
Step 1	-7	-10	0.049	2	0.976
Step 2	-11	-12	0.283	4	0.991
Step 3	-2	-14	1.179	6	0.978
Step 4	-4	-5	2.273	8	0.971
Step 5	-6	-9	3.517	10	0.967
Step 6	-15	2	5.048	12	0.956
Step 7	-3	5	6.764	14	0.943
Step 8	-13	3	11.313	16	0.790
Step 9	-1	1	18.650	18	0.414
Step 10	-8	9	28.397	20	0.100
Step 11	4	7	41.749	22	0.007
Step 12	6	8	73.926	24	0.000
Step 13	11	12	173.638	26	0.000
Step 14	10	13	440.552	28	0.000

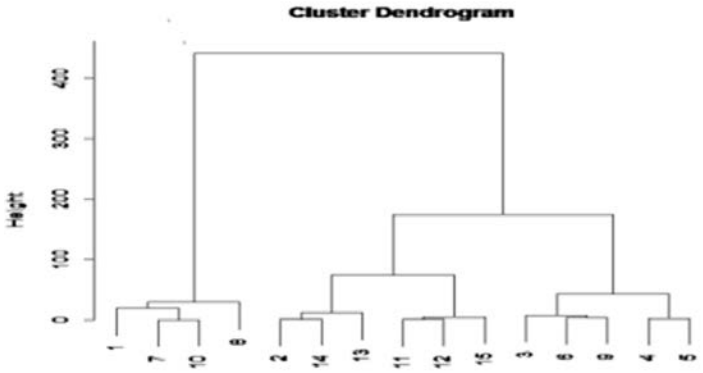

Figure 4. Cluster dendrogram for 2011.

In Figure 4, a few items appeared under understanding, and all other questions appeared under problem solving and reasoning.

As a result, over the four years of administering high–stakes tests, reasoning and problem solving strategies were more frequently tested in questions than understanding.

For the second phase, a comparative model was run in order to observe the fitness of the model over these four years. Table 5 shows the output of MultiLC IRT compared model over four years.

TABLE 5
MultiLCIRT Comparative Model over Four Years

K1, 2, 3irt, mirtlc	log–lik (2008)	log–lik (2009)	log–lik (2010)	log–lik (2011)
out = class_item(S,yv,k=3,link=2)	-29460.91	-25743.94	-21456.53	-31316.91

According to Table 5, the log–likehood in 2010 led to the best-fitting model, because this year showed the lowest value compared to the other years. Hence, the test was designed properly. Table 6 also shows the results for MultiLC IRT.

TABLE 6
MultiLCIRT Outputs for 3 Dimensions

	3-Dimensional		
	LL	AIC	BIC
2008	-29460.91	-4533.83	-59788.82
2009	-25743.94	-3961.91	-52354.88
2010	-21456.53	-33.1.004	-43780.06
2011	-31316.91	-4817.986	-63500.82

According to Table 6, the lowest AIC and BIC values along with the best-fitting model for MultiLC IRT were obtained for 2010. Thus, the structure of the reading comprehension tests and cloze tests were more valid than in other years. Akaike (1974) proposed that when comparing various models, the best-fitting model was the one with the lowest AIC and BIC values. For example, in Table 3 the AIC and BIC values were lowest values for the two dimensional models in 2010, with AIC=42,832.67 and BIC=43,119.43. Moreover, a similar AIC value (42,821.61) and BIC value (43,193.09) were obtained for two consecutive years, 2009 and 2010.

The two and three dimensional IRT results are provided below.

Two-Dimensional and Three-Dimensional IRT

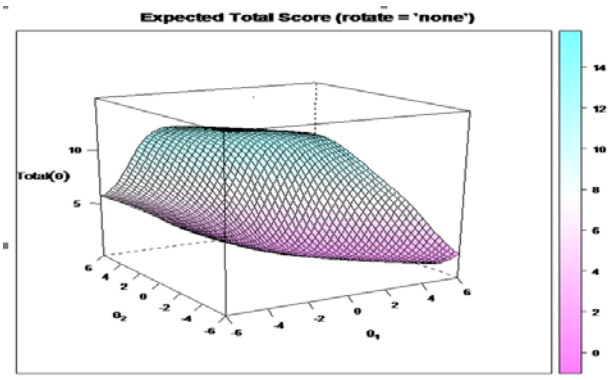

Figure 5. 2-dimensional IRT model for 2008.

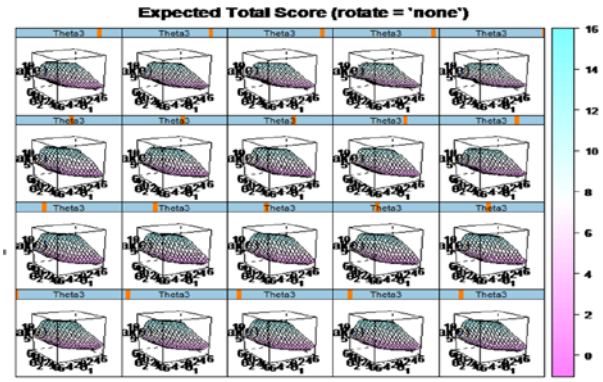

Figure 6. 3-dimensional IRT model for 2008.

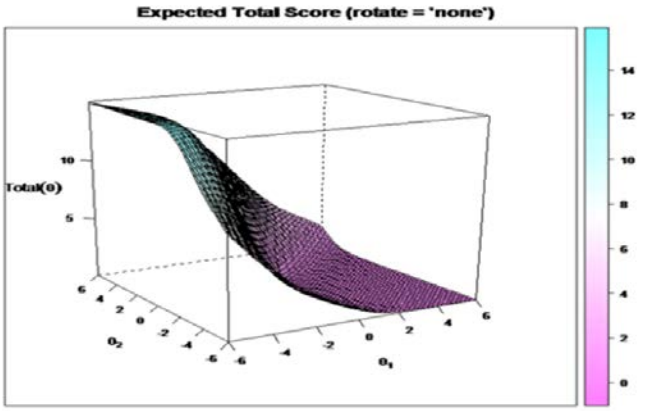

Figure 7. 2-dimensional IRT model for 2009.

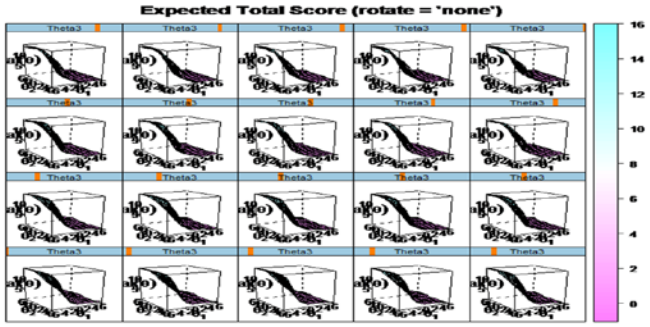

Figure 8. 3-dimensional IRT model for 2009.

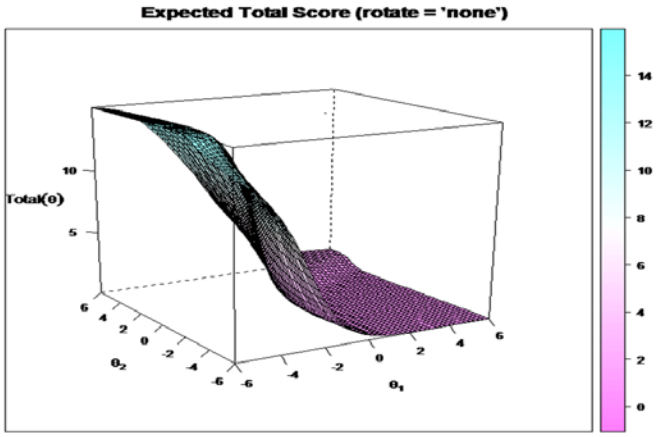

Figure 9. 2-dimensional IRT model for 2010.

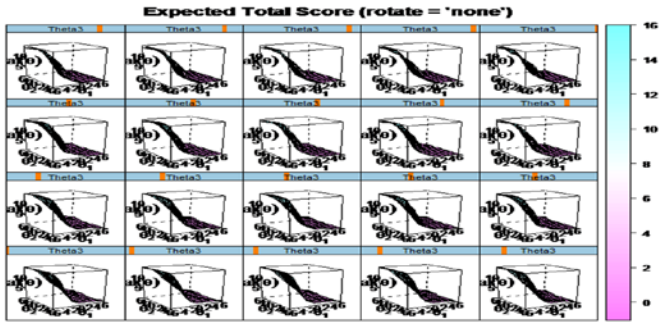

Figure 10. 3-dimensional IRT model for 2010.

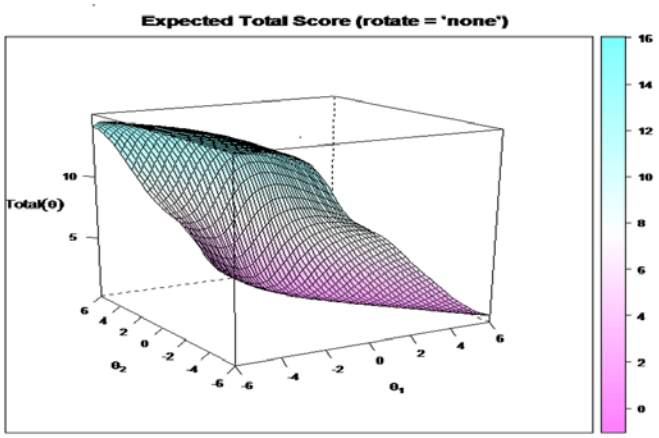

Figure 11. 2-dimensional IRT model for 2011.

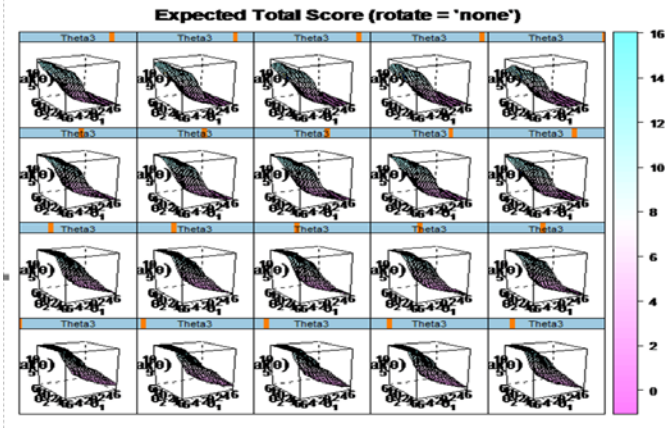

Figure 12. 3-dimensional IRT model for 2011.

In Figures 5, 6, 7, 8, 9 and 10, 11, and 12, the results showed an item characteristic surface, instead of a curve. That is, items were not at the same ability level for test takers.

For the MultiIRT, the output is provided over the four years in Table 7.

TABLE 7
IRT Outputs, 2-3 Dimensions

	2 Dimensional			3 Dimensional		
	LL	AIC	BIC	LL	AIC	BIC
2008	-29433.49	58954.97	59241.73	-29381.63	58877.27	59248.75
2009	-25861.71	51811.43	52098.18	-21353.8	42821.61	43193.09
2010	-21372.34	42832.67	43119.43	-21353.8	42821.61	43193.09
2011	-31037.7	62163.41	62450.16	-31010.61	62135.22	62506.7

Discussion and Conclusion

The purpose of the present study was to ascertain find the importance of cognitive factors, namely, understanding, problem solving and reasoning, in reading comprehension and cloze tests. The reason for choosing these two tests was that cognitive factors play crucial roles in reading comprehension and cloze tests. Danili and Reid (2006) emphasized how these two factors affect performance and learning. It was also aimed at observing the best-fitting model based on IRT and MultiLC IRT, and it was of paramount importance to scrutinize this high–stakes test to strengthen its validity.

For the first research question, the fittest models in both MIRT and MultiLC IRT were in 2010, which

meant the majority of items focused on reasoning and problem solving. Interestingly, Traub and Lam (1985, p. 22) believed that the assumption of dimensionality and the best-fitting model might be evident in MIRT; however, the results showed multidimensionality in MultiLC IRT rather than just in MIRT. Regarding the provided dendrograms in the present study, the distribution of items mostly lay in reasoning and problem solving, under the subcategory of the cognitive domain.

Practically, the consequences of these findings overshadowed the newly developed national textbooks called "English for Schools." Although they were developed based on the communicative approach, students were not adequately trained and prepared for this. In addition, hiring experts is necessary for standardizing high–stakes tests.

Considering the second research question, the findings did not show the real abilities of test takers, since their scores distribution were surf-shaped rather than curved, especially in 2008. According to Bartolini Bussi, Boni, Ferri, and Garuti (1999), Douek (2006), and Schoenfeld (1992), these cognitive factors are the main complex constructs; hence, there is a need both to provide test takers with standardized high–stakes tests (Tout & Spithill, 2014), and to take into account the cognitive factors in teaching to test takers. These results also call for serious attention to designing, teaching, and testing, because the future of examinees is at stake. Therefore, it is suggested that materials developers, teachers, test designers, and the National Measurement Organization consider and apply the findings of the study.

The Authors

Masoud Geramipour has a Ph.D. in Assessment and Measurement from the Faculty of Psychology and Education at Allameh Tabatabie University, Tehran. Currently, he is an Assistant Professor in the Department of Curriculum Studies, Kharazmi University. He has taught research methodology and psychometrics courses to educational research students since 2010.

Email: mgramipour@yahoo.com

Niloufar Shahmirzadi (corresponding author) is a Ph.D. candidate in Applied Linguistics at Islamic Azad University, Central Tehran Branch. She is a part-time lecturer, and has published some articles and books. She has also attended some national and international conferences. She is a member of the Young Researchers and Elite Club. Her areas of interest lie in Applied Linguistics, Language Testing, Assessment, and Educational Measurement.

Email: niloufar_shahmirzadi83@yahoo.com

Acknowledgements

The authors would like to acknowledge the National Organization for Educational Testing for a grant for the data, without which the present study could not have been conducted.

Funding

This work was supported by Kharazmi University.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255–278.

Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters. *Scandinavian Journal of Statistics, 20*, 63–71.

Akaike, H. (1974). IEEE Transactions on Automatic *Control, 19*, 716.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*(1), 37–48.

Bartolini Bussi, M. G., Boni, M., Ferri, F., & Garuti, R. (1999). Early approach to theoretical thinking: Gears in primary school. *Educational Studies in Mathematics, 39*, 67–87.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72*, 141–157.

Christensen, K., Bjorner, J., Kreiner, S., & Petersen, J. (2002). Testing unidimensionality in polytomous Rasch models. *Psychometrika, 67*, 563–574.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd edition) (pp. 443–507). Washington, DC: American Council on Education.

Danili, E., & Reid, N. (2006). Cognitive factors that can potentially affect pupils' test performance. *Chemistry Education Research and Practice, 7*, 64–83.

Douek, N. (2006). Some remarks about argumentation and proof. In P. Boero (Ed.), *Theorems in school: From history, epistemology and cognition to classroom practice* (pp. 137-161). Rotterdam: Sense Publishers.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189–199.

Duncan, O., & Stenbeck, M. (1987). Are Likert scales unidimensional? *Social Science Research, 16*, 245–259.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231.

Hambleton, R., & Swaminathan, H. (1983). *Item response theory: Principles and applications*. Kluwer, Boston.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49–78.

Irvine, S. H. (2002). Item generation for test development: An introduction. In S. H. Irvine, & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3-34). Mahwah, NJ: Erlbaum.

Kelderman, H., & Rijkes, J. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*(2), 149–176.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99–117.

Oxford, R. (1990). *Language learning strategies: What every teacher should know*. Boston: Heinle and Heinle Publishers.

Reckase, M. D. (2009). *Mutli–dimensional item response theory: Statistics for social and behavioral sciences*. New York: NY, Springer.

Reckase, M. D. (1990). Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests. Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16–20).

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications.

Journal of Educational Statistics, *4*, 207–230.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*(3), 287–297.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *The British Journal of Mathematical and Statistical Psychology*, *44*, 75–92.

Rubin, J., Chamot, A. U., Harris, V., & Anderson, N. J. (2007). Intervening in the use of strategies. In A. D. Cohen, & E. Macaro (Eds.), *Learning strategies in foreign and second language classrooms* (pp. 117–139). London: Continuum.

Schoenfeld, A. (1992), Learning to think mathematically: problem solving, metacognition, and sense making in mathematics. In D. Grows (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334-370). Macmillan, New York.

Tout, D., & Spithill, J. (2014). The challenges and complexities of writing items to test mathematical literacy. In R. Turner, & K. Stacey (Eds.), *Assessing mathematical literacy: The PISA experience* (pp. 145-172). New York: Springer.

Traub, R. E., & Lam, Y. R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology*, *36*, 19–48.

von Davier, M., & Rost, J. (1995). *Polytomous mixed Rasch models*. In G. Fischer, & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). Springer–Verlag: New York.

Zhang, J. (2004). Comparison of unidimensional and multidimensional approaches to IRT parameter estimation. Technical report, ETS Research Rep. No. RR–04–44, Princeton, NJ: ETS.

Zhang, J. (2005). *Estimating multidimensional item response models with mixed structure* (Research Report No. RR–05–04). Princeton, NJ: Educational Testing Service. https://www.r-project.org/