ÁP DỤNG LÝ THUYẾT ỨNG ĐÁP CÂU HỎI ĐA CHIỀU VÀO ĐO LƯỜNG VÀ ĐÁNH GIÁ ĐỀ THI ANH VĂN CUỐI KỲ

APPLYING MULTIDIMENSIONAL ITEM RESPONSE THEORY IN VALIDATING AN ENGLISH FINAL TEST

Do Thi Ha
HCMC University of Technology and Education

Received 01/01/2016, Peer reviewed 14/03/2016, Accepted for publication 30/03/2016

ABSTRACT

This paper investigated the application of Multidimensional Item Response Theory (MIRT) in assessing and evaluating an English multiple-choice test. The data was gathered from non-English majors taking the English 2 course at Ho Chi Minh City University of Technology and Education. Firstly, Rasch Testlet model was exploited to determine whether the data were indeed multidimensional. Then factor analyses (FA) were conducted to examine the potential latent dimension(s). Item difficulty and item discrimination were estimated using two-parameter MIRT model. "Mirt" package of the freeware R was used to analyze the data. The findings, therefore, suggest how MIRT can be utilized in the test development process.

Key words: Multidimensional Item Response Theory, Rasch Testlet model, factor analyses, freeware R.

TÓM TẮT

Bài báo nghiên cứu ứng dụng của lý thuyết ứng đáp câu hỏi đa chiều (MIRT) vào đo lường và đánh giá đề thi trắc nghiệm môn Tiếng Anh. Dữ liệu trong bài báo được thu thập từ bài thi cuối kỳ môn Anh Văn 2 dành cho sinh viên không chuyên tại trường Đại học Sư phạm Kỹ thuật Tp. HCM. Trước tiên, mô hình Rasch Testlet được dùng để kiểm tra tính đa chiều của đề thi. Tiếp theo, phân tích nhân tố (FA) được sử dụng để xác định số chiều cần đo. Độ khó và độ phân biệt của mỗi câu hỏi trong đề thi được ước lượng bằng mô hình MIRT 2 tham số. Việc xử lý dữ liệu được thực hiện bằng gói lệnh "mirt" của phần mềm R. Kết quả của bài báo cung cấp thông tin hữu ích cho giáo viên trong việc điều chỉnh phương pháp đánh giá.

Từ khóa: Lý thuyết ứng đáp câu hỏi đa chiều, mô hình Rasch Testlet, phân tích nhân tố, phần mềm R.

I. INTRODUCTION

A test can be studied from different angles and the items in the test can be evaluated according to different theories.

Classical Test Theory (CTT) has been widely used in test development since the 20th century (Bechger et al., 2003) with major focus on total test score. Within a CTT framework, information about the performance depends on the characteristics of the test and the sample.

Meanwhile, **Item Response Theory (IRT)** relates the probability of a particular item response to overall examinee's ability (Camilli & Shepard, 1994). Therefore, in IRT, ability parameters estimated are not test

dependent and item statistics (i.e., item difficulty and item discrimination) are sample independent (Hambleton & Swaminathan, 1985). However, these cannot be achieved without model data fit (Fan, 1998) which involves two basic assumptions: unidimensionality and local item independence (Hambleton & Swaminathan, 1985). The assumption of unidimensionality postulates that items of a test measure only one ability, regardless of individuals' cognitive and personal characteristics, which cannot often be put under control. The other important assumption, local item independence, can be defined as the avoidance of significant association among item responses (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000).

As no real data ever fit these assumptions, **Multidimensional Item Response Theory (MIRT)** is applied to validate test structure and dimensionality. In this paper, the emphasis is on UTE English 2 multiple-choice test administered in June, 2015 (UTE is short for Ho Chi Minh City University of Technology and Education). Based on the procedures illustrated in this case study, any other tests can be evaluated once examinee item response data are collected.

II. LITERATURE REVIEW

1. Test dimensionality

Because validity refers to how well the assessment instrument measures the objectives of the test (Henning, 1987), it is a fundamental consideration in test development. The dimensional structure of a test (i.e. reflection of the intended traits) is used to provide one type of validity evidence. Many IRT models have been applied to analyze language tests and proved to provide construct validity evidence (McNamara, 1991; Embretson & Reise 2000; Alderson & Banerjee, 2002; Walt & Steyn, 2008).

Multidimensionality does exist to a greater or lesser extent. Previous research has shown that there is high interrelation of skills associated with grammar, vocabulary and reading comprehension in a language test. Even a reading comprehension section may include a number of noticeable subskills or abilities (Schedl et al., 1996; Wilson, 2000).

2. Multidimensional Item Response Theory (MIRT)

When the test assesses more than one underlying ability, MIRT models such as exploratory and confirmatory (Embretson & Reise, 2000) are employed. While exploratory procedures focus on discovering the best fitting model, confirmatory approaches evaluate some hypothesized test structure. Confirmatory MIRT models can be further classified into one of two groups: compensatory and noncompensatory. In compensatory MIRT models, a shortfall in one ability can be evened out by an increase in other abilities. On the contrary, in noncompensatory MIRT models, adequate levels of each measured ability are required, and nothing can make up for the deficiency of any ability.

As regards compensatory models, Reckase (2009) presented the logistic MIRT model in slope-intercept form:

$$P\left(X=1\,|\,\theta,a,d\right)=\frac{e^{a.\theta^{T}+d}}{1+e^{a.\theta^{T}+d}}, \qquad (1)$$

where $\theta=\left(\theta_1,\theta_2,...,\theta_k\right)$ is a vector of person latent traits, $a=\left(a_1,a_2,...,a_k\right)$ is a vector of item slopes and d is the intercept parameter related to difficulty.

Discriminating power of an item for the most discriminating combinations of dimensions can be given as:

$$MDISC = \sqrt{\sum_{j=1}^{k} a_j^2}\ . \quad (2)$$

The difficulty of each item in the test was calculated by the following formula:

$$MDIFF = -\frac{d}{MDISC}. \quad (3)$$

3. Previous research

In Li et al.'s (2012) paper, an empirical K-12 science assessment was investigated for dimensionality validation using MIRT approach. The unidimensional IRT model and testlet model were also included, which provides multiple-dimensional estimates for practitioners. While the procedures for test validation can be cycled back into test design, the findings of the test dimensionality may not be generalized to other assessments.

Heydari et al. (2014) took a closer look at a nationwide large-scale English proficiency test (TOLIMO: The Test of Language by the Iranian Measurement Organization). In this study, 154 participants worked on 50 multiple-choice items of "structure and written expression" section of TOLIMO. Under IRT (Item Response Theory) analysis, the finding that a large number of items (84%) were fitting the IRT model implied the construct validity of the test. However, its principal limitation is the lack of access to the real TOLIMO examinees, and the authors had to use a mock exam instead.

Taking the abovementioned research as guidelines, the researcher adapted some MIRT models for validating a multiple choice test for non-English majors, which so far has not been investigated statistically and appropriately.

III. OBJECTIVE & METHODOLOGY

1. Research objective

The purpose of the study is to determine if the use of a multidimensional analysis is better suited than a unidimensional analysis for the English 2 final test. Therefore, the following questions were examined for the test development:

- How many intended dimensions involve in the test?
- How can the difficulty and discrimination of each item in the test be estimated?

2. Instruments & Methodology

The data for this study was gathered randomly from 138 students taking the English 2 final test of the second term 2014 – 2015 (For further details, raw data can be retrieved from the exam paper archives of UTE Faculty of Foreign Languages). The test consists of three sections aiming at four learning outcomes: Vocabulary (Items 1-8, 25-30), Grammar (Items 9-19, 25-30), Functions of Speech (Items 20-24) and Reading Comprehension (Items 31-60). In this case, 30 multiple-choice items of the two fill-in sections (Items 1-30) were investigated for students' intended abilities.

Firstly, Rasch Testlet Model was exploited to determine whether the data were indeed multidimensional. Then a Principal Component Analysis (PCA) was conducted using freeware R. With some idea about the underlying constructs, Varimax rotation was applied for identifying the most significant evidence. The final stage is an illustration of how item difficulty and discrimination can be appraised using "mirt" package of the freeware R. The main focus of this module is on the two-parameter compensatory MIRT model because it has been extensively developed, studied, and applied to practical

testing problems. This feature makes it possible for an examinee with low ability on one dimension to compensate by having a higher level of ability on other dimensions.

IV. DATA ANALYSIS

1. Rasch Testlet Model

Originally suggested by Wang and Wilson (2005), Rasch Testlet Model was extended by Wainer et al. (2007). Each testlet effect was, therefore, treated as a different dimension together with one general factor underlying each testlet. In this model, the probability of a correct response to an item i nested in testlet $d(i)$ for a person j with ability θ_j is given by:

$$P\left(X_{ij}=1\middle|\theta_j\right)=\frac{1}{1+e^{-\left(a_i\theta_j-b_i+\gamma_{jd(i)}\right)}}, \quad (4)$$

where a_i and b_i are the item discrimination and difficulty parameters, respectively, and $\gamma_{jd(i)}$ is the testlet effect parameter for person j on testlet $d(i)$. When there is no testlet effect (i.e. $\gamma_{jd(i)}=0$), the model reduces to the standard two-parameter IRT model where local item independence is assumed to hold. Testlet-based local item dependence manifests itself through the testlet effect variance $\sigma^2_{jd(i)}$. That is, the greater the testlet effect variance of a testlet $d(i)$, the higher is the degree of associated local item dependence; if the testlet effect variance is zero, there is no indication of local dependence within the testlet (Wainer & Wang, 2000).

The EAP (Expected A Posteriori) reliability of Rasch Testlet Model is 0.668, whereas the EAP reliability of unidimensional IRT is 0.544. This means that the MIRT model was a better-fitting model.

2. Principal Component Analysis

Principal Component Analysis (PCA) minimizes the number of observed variables to a smaller number of principal components that make up most of the variance of the observed variables. The number of factors can be determined by selecting those for which the Eigenvalues are greater than 1. This value means that these factors account for more than the mean of the total variance in the items, which is known as the Kaiser–Guttman rule (Guttman, 1954; Kaiser, 1960).

Table 1. Principal Component Analysis Eigenvalue and variance explained

	Eigenvalue	Percentage of VAR	Cumulative percentage of VAR
Component 1	3.2914018	10.9713393	10.97134
Component 2	2.6159060	8.7196865	19.69103
Component 3	1.7105208	5.7017359	25.39276
Component 4	1.5755281	5.2517602	30.64452
Component 5	1.4819369	4.9397898	35.58431
…			

The Eigenvalues are reported in Table 1. Among the ten components (i.e. factors) meeting the rule, the first three components had Eigenvalues much greater than 1 (i.e. 3.2914018, 2.6159060 and 1.7105208), which strongly proves multidimensionality.

The following seven components had Eigenvalues only slightly over 1. A corresponding scree plot of the PCA is shown in Figure 1 for the pattern. The magnitude of the Eigenvalues can lead to a conclusion that at least three components exist in the 30 multiple-choice items of the test. Meanwhile, the percentage of VAR illustrates the variational proportion of observed variables. For example, 10.97% of VAR of the first factor indicate that 10.97% of the variation can be explained.

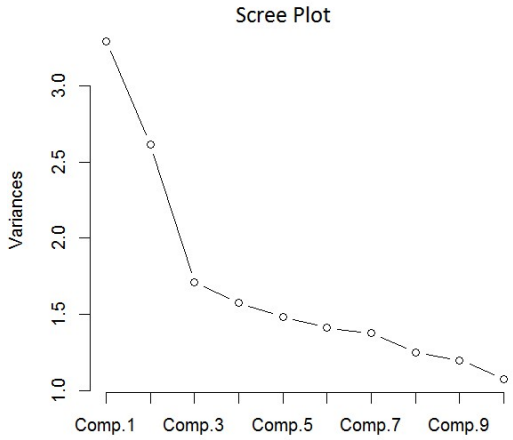

Figure 1. *Scree plot of Principal Component Analysis*

3. Confirmatory Factor Analysis

With Varimax rotation method (Kaiser, 1958), each original variable tends to be associated with one (or a small number) of factors, and each factor represents only a small number of variables. In addition, the factors can often be interpreted from the opposition of few variables with positive loadings to few variables with negative loadings. Factor loading numbers which are greater than 0.3 help categorize the items. For instance, according to Table 2, Item 1 can be designated for Factor 1 at the strongest rate of 0.568.

Table 2. Rotation Component Matrix

	Factor 1	**Factor 2**	**Factor 3**
Item1	0.568	0.208	0.228
Item2	0.338	0.061	0.411
Item3	0.348	0.081	0.037
Item4	0.602	0.118	0.072
…			

The Varimax rotation procedure applied to the table of loadings gives a new set of rotated factors for the 30 test items:

Factor 1: Items 1-7, 15, 20, 23, 24, 26, 27, 29

Factor 2: Items 10-13, 17-19, 22, 23, 25, 29

Factor 3: Items 8, 14, 16, 21, 28, 30

With the exclusion of the test Reading section (Items 31-60), the three new-found factors are not really compatible with the learning outcomes of the test:

Vocabulary: Items 1-8, 25-30

Grammar: Items 9-19, 25-30

Functions of Speech: Items 20-24

The above-mentioned mismatch indicates that only MIRT model with the emergent factors can measure students' real abilities. In addition, the right factor classification acts as a premise for the next steps of estimating item difficulty and discrimination.

4. Multidimensional item difficulty and discrimination

The table below shows the figures of slopes and intercept when "mirt" package of the freeware R is applied. The values in the first column (a1) reflect the item slopes for Factor 1, (a2) for Factor 2 and (a3) for Factor 3 while the values in the fourth column (d) correspond to the item intercept:

Table 3. Parameter slopes and intercepts

Item1	a1	a2	a3	d	g	u
par	1.509	0	0	1.058	0	1
Item2	a1	a2	a3	d	g	u
par	0.684	0	0	1.184	0	1
Item3	a1	a2	a3	d	g	u
par	0.592	0	0	0.746	0	1
Item4	a1	a2	a3	d	g	u
par	1.686	0	0	0.507	0	1
…						

The discrimination of items is characterized by their slopes. The positive slopes show that the probability of a correct response of a good student is higher than that of a bad student, while the negative slopes depict the opposite trend. For further analysis, the discriminating combination and item difficulty mentioned in formulas (2) and (3) should be calculated.

Table 4. Discrimination and Item difficulty

	a1	a2	a3	d	MDISC	MDIFF
Item1	1.51	0	0	1.06	1.51	–0.70
Item2	0.68	0	0	1.18	0.68	–1.73
Item3	0.59	0	0	0.75	0.59	–1.26
Item4	1.69	0	0	0.51	1.69	–0.30
…						

In Table 4, MDISC stands for the discriminating combination, and MDIFF represents the item difficulty. According to Baker (2001) and Hasmy (2014), the discriminating combination and item difficulty can be classified respectively as follows:

Table 5. Labels for item discrimination

Very high	$MDISC \geq 1.7$
High	$1.35 \leq MDISC < 1.7$
Moderator	$0.65 \leq MDISC < 1.35$
Low	$0.35 \leq MDISC < 0.65$
Very low	$MDISC < 0.35$

Table 6. Labels for item difficulty

Very hard	$MDIFF \geq 2$
Hard	$0.5 \leq MDIFF < 2$
Medium	$-0.5 \leq MDIFF < 0.5$
Easy	$-2 \leq MDIFF < -0.5$
Very easy	$MDIFF < -2$

From Tables 4, 5 and 6, it can be deduced that:

- A majority of the items have fairly good discriminations (18 items are at moderator level and above). Meanwhile just 4 items need to be improved as their MDISC are less than 0.35.
- Regarding item difficulty, it can be seen that more than 80% of items can be ranked at medium and below.

The following bar charts give an overview of item discrimination and item difficulty of 30 items in the target test.

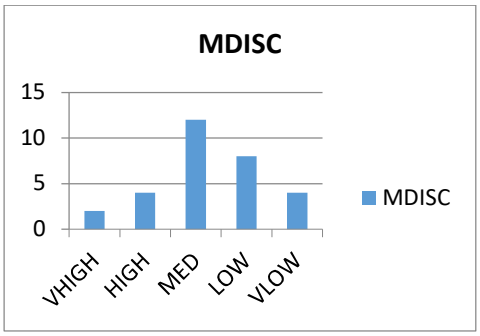

***Figure 2**. Discrimination of 30 items*

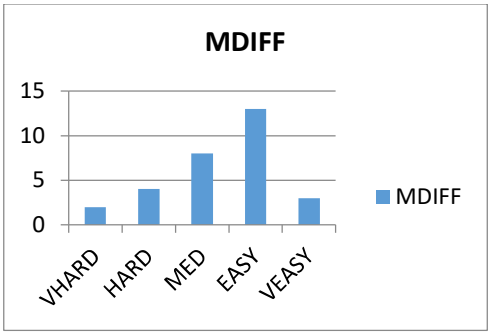

***Figure 3**. Item difficulty of 30 items*

V. DISCUSSION AND IMPLICATIONS

This study investigated the application of factor analyses to validate test dimensionality. A 30-item excerpt of an English multiple-choice test was used as an example when MIRT is a better-fitting model. The results reflect overlapping trait issues inherent in the test as in any kind of assessment. One item does not measure only one ability, and in some cases the real measurement goes beyond the intended outcome. For example, Item 15 was aimed at Grammar knowledge (Factor 2), but it turned out to be testing Vocabulary (Factor 1) instead. Therefore, the use of MIRT model suggests more accurate ideas for evaluating the test and examinees' competence.

When the discrimination and difficulty levels are taken into consideration, items 9, 14 and 21 should be altered or removed from the test bank because of their very low rates. Further insight into item analyses is especially valuable to distinguish among students according to how well they meet the learning goals. Once the quality of each item (i.e. the discrimination and difficulty) and of the whole test is assessed, educators and stakeholders can decide what changes to make for a good test bank construction.

The procedures illustrated in this real example can be utilized to validate the test dimensionality as follows:

- First, one should identify the test's intended dimensions of ability using Rasch Testlet Model.
- Second, exploratory approaches (e.g., PCA) should be implemented to determine the potential latent dimension(s).
- Third, confirmatory analysis can then be conducted by Varimax rotation to simplify the interpretation and categorize the items.
- And finally, "mirt" package of the freeware R is employed to shed light on the multidimensional difficulty and discrimination of each item in the test.

REFERENCES

[1] Alderson, J. C., & Banerjee, E. (2002). Language testing and assessment. *Language Testing, 35*, 79-113.

[2] Baker, F. (2001). *The basic of item response theory*. USA: ERIC Clearinghouse on Assessment and Evaluation.

[3] Bechger, T.M., Maris, G., Verstralen, H.H.F.M., & Beguin, A.A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27*(5), 319-334.

[4] Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

[5] Embretson, S. E., &Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

[6] Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19*, 149–161.

[7] Hambleton, R. K., &Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. USA: Kluwer-Nijhoff Publishing.

[8] Hasmy, A. (2014). Compare unidimensional & multidimensional Rasch model for test with multidimensional construct and items local dependence. *Journal of Education and Learning*, *8*(3), 187-194.

[9] Henning, G. (1987). *A guide to language testing*. Cambridge, Mass.: Newbury House.

[10] Heydari, P., Bagheri, M. S., Zamanian, M., Sadighi, F., & Yarmohammadi, L. (2014). Investigating the construct validity of "Structure and Written Expression" section of TOLIMO through IRT. *International Journal of Language Learning and Applied Linguistics World, 5*(2), 105-123.

[11] Kaiser, H. F. (1958). The Varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187-200.

[12] Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychology Measurement, 34*, 111–117.

[13] Li, Y., Jiao, H., &Lissitz, R. W. (2012). Applying multidimensional item response theory models in validating test dimensionality: An example of K-12 large-scale science assessment. *Journal of Applied Testing Technology, 13*(2), 1-27.

[14] McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8*(2), 139-159.

[15] Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

[16] Schedl, M., Gordon, A., Carey, P. A., & Tang, K. L. (1996). *An analysis of the dimensionality of TOEFL reading comprehension items (TOEFL Research Report No. 53)*. Princeton, NJ: ETS.

[17] Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

[18] Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203–220.

[19] Walt, J., &Steyn, F. (2008). The validation of language tests. *Linguistics, 38*, 191-204.

[20] Wang, W. C., & Wilson, M. R. (2005). The RaschTestlet model. *Applied Psychological Measurement, 29*, 126–149.

[21] Wilson, K. M. (2000). *An exploratory dimensionality assessment of the TOEIC test (Research Report No. 14)*. Princeton, NJ: ETS.