Applying Multidimensional Three-Parameter Logistic Model (M3PL) In Validating a Multiple-Choice Test

Do Thi Ha

HCMC University of Technology and Education, Ho Chi Minh City, Vietnam

Abstract- This paper investigated the application of Multidimensional Three-Parameter Logistic model (M3PL) in assessing and evaluating an English multiple-choice test. Guessing parameter, when put in the picture, has been proved to strengthen the results of test validation. The data was gathered from 488 non-English majors taking an English final test at a university in Ho Chi Minh City, Vietnam. The findings, therefore, suggest how M3PL can be utilized in the test development process.

Index Terms- Factor analysis,M3PL model, Multidimensional Item Response Theory

I. INTRODUCTION

Item Response Theory (IRT) has long been utilized in test validation. The preliminary ideas of IRT model were presented by Thurston (1925), followed by Lord (1952) with a notion of Item Characteristic Curve (ICC). ICC describes the relationship between the probability of a correct response to item *i* and student *j*'s ability (referred to as θ_j). In 1968, Birnbaum applied logistic models for IRT, which were then built up by Lord and Novick (1968), and then Bock and Aitkin (1981). Meanwhile, they developed some approaches to parameter estimation. One of the salient features of IRT is how it relates each examinee's latent traits to item parameters through his/her response to each question in the test (Wright & Stone, 1979; Camilli& Shepard, 1994; Baker, 2001). Therefore, in IRT, ability parameters estimated are not test dependent and item parameters (i.e. item difficulty and item discrimination) are sample independent (Hambleton&Swaminathan, 1985). However, its two basic assumptions: unidimensionalityand local item independence cannot be satisfied in most cases (Schedl et al., 1996; Wilson, 2000). Schedlet al. (1996) and Wilson (2000) also pointed out that in a language test, especially when it comes to reading comprehension section, multiple skills are required for the best performance.

In the late 1970s and early 1980s, a number of researchers got involved in promoting Multidimensional Item Response Theory (MIRT) (Reckase, 1972). In addition to work by Reckase (1972) on the multidimensional Rasch model, Mulaik (1972), Sympson (1978) and Whitely (1980a, b) proposed multidimensional models for the item/person interaction. According to Resckase (1985), some test items demand more than one ability to deal with (such as arithmetic and algebraic manipulations in a Mathematics test). A study by Kose and Demirtasli (2012) confirmed that these latent traits can be estimated more precisely by MIRT than IRT, i.e. MIRT standard errors are smaller. Kose and Demirtasli added that the more items a test has, the smaller the standard errors of model parameters are, which is a great value to educators in test design. In Li et al.'s (2012) paper, an empirical K-12 science assessment was investigated for dimensionality validation using Multidimensional 2-Parameter Logistic (M2PL) approach. The unidimensional IRT and Testlet models were also included, which provides multiple-dimensional estimates for practitioners. Nevertheless, Li et al. (2012) have yet to group the questions in accordance with factor analysis. In another approach to MIRT, Do (2016) did the classification, but failed to estimate the model fitting level.

In fact, examinees have a tendency to guess answers in a multiple-choice test. Therefore, Birnbaum (1968) took into account the three-parameter model with guessing parameter $c_i \in (0,1)$ to evaluate students' guessing behavior. DeMars (2007) showed that fixing c parameter (guessing parameter) to zero may skew the interpretation of item difficulty. Li and Lissitz (2004) also discussed how poorly estimated c-parameters can lead to large standard errors in assessing the difficulty parameters in the unidimensional 3-parameter logistic model. As a result, the modification of Multidimensional 3-Parameter Logistic model (M3PL) has been a crucial extension of M2PL.

In Vietnam, IRT models have been of interest to recent research with Rasch model by Nguyen (2004), IRT 2PL model by Nguyen (2008) and Nguyen (2014), and IRT 3PL by Le et al. (2016). As MIRT and M3PL have not received proper attention, this paper aims at shedding more light on Multidimensional 3-Parameter Logistic model (M3PL) and its application in validating an English final test.

II. LITERATURE REVIEW

1. Test dimensionality

Validity has been taken into consideration in test development as it refers to how well the assessment instrument measures the objectives of the test (Henning, 1987). One type of validity evidence can be traced through the dimensional structure of a test (i.e. reflection of the intended traits). 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).

The number of dimensions for a multidimensional analysis of a test data has long been researchers' concern. Holzinger and Harman (1941) figured out the number of variables *n*needed to support the estimation of the factor loading for *m* independent factors in the following expression:

$$n \geq \frac{(2m+1)+\sqrt{8m+1}}{2} \quad (1)$$

This expression was deduced assuming no error in the estimation of correlation. Thurstone (1947) later recommended that the number of variables needed for a convincing analysis with *m* factors should be "two or three times greater" than this figure. Reise et al. (2000) in the research on selecting number of dimensions concluded that it is better to overestimate this figure and that scree plot, parallel analysis, and analysis of the residual correlation matrix can be employed to determine the dimensionality needed to model a matrix of test data.

2. Multidimensional 3-parameter logistic model (M3PL)

Multidimensional Item Response Theory (MIRT), as an extension of unidimensional IRT, allows for the analysis of multiple constructs simultaneously. In order to apply MIRT, the following assumptions need to be met:

- Monotonicity: the probability of each student's correct response will increase when one of his/her abilities increases.
- Local independence (Reckase, 2009): suchpossibility is not affected by other examinees as well as the student's response to other items.

When the test assesses more than one underlying ability, MIRT models such as exploratory and confirmatory (Embretson&Reise, 2000) are adopted. 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. This study focused on the former model because of its popularity in theoretical research (Reckase, 2009).

As mentioned earlier, the addition of guessing parameter should result in improved estimation. Multidimensional 3-Parameter Logistic model (M3PL) was, therefore, implemented:

$$P\left(X_{ij}=1|\theta_j,a_i,c_i,d_i\right)=c_i+(1-c_i).\frac{\exp\left(a_i.\theta_j^{'}+d_i\right)}{1+\exp\left(a_i.\theta_j^{'}+d_i\right)}, \quad (2)$$

in which exp(.) is exponential function with base *e*; $P\left(X_{ij}=1|\theta_j,a_i,c_i,d_i\right)$ is the probability of student *j*'s correct response to item *i*; θ_j is vector of student *j*'s ability; a_i is vector of item *i* slope; $c_i \in (0,1)$ is guessing parameter; and d_i is intercept parameter. Vectors a_i, θ_j have the same elements *m*, which is the number of dimensions.

The M3PL model was designed to account for observed empirical data such as that provided in Lord (1980) which shows that examinees with low capabilities still have a probability of correct response. As a result, this model contains a single lower asymptote or guessing parameter $c_i \in (0,1)$ to specify such probability for examinees with very low value in θ. Theoretically, the interval of c_i is between 0 and 1. In reality, since $c_i \geq 0,35$ is often omitted from the test bank (Baker, 2001), c_i ranges from 0 to 0.35.

Figure 1 (Reckase, 1985, p. 403) illustrates the graph of MIRT characteristic function when *m*=2, or in other words, Item Response Surface (IRS).

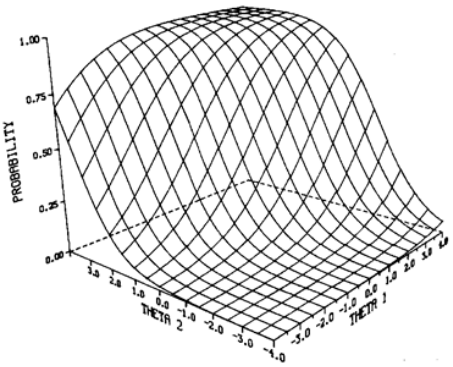

Figure 1.IRS of MIRT model with $a_{i1}=0,75; a_{i2}=1,2; c_i=0; d_i=-1$

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

$$\eta_{MDISC}=\sqrt{\sum_{k=1}^{m} a_{ik}^2} \quad (3)$$

If the exponent in Formula (2) is set to some constant value, *k*, all θ vectors that satisfy the expression $k=a_i\theta_j^{'}+d_i$ fall along the straight line which is called contour line and they (θ) all yield the same probability of correct response for the model. The signed distance from the origin to the 0.5 probability contour line is called the difficulty of a multidimensional item (Reckase, 1985). Its difficulty was calculated as follows:

$$b_{MDIFF}=-\frac{d_i}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}} \quad (4)$$

This formula helps identify the item threshold, i.e. the 0.5 probability of a positive answer. According to Baker (2001) and Hasmy (2014), the discriminating combination and item difficulty can be classified respectively as in Tables 1 and 2:

Table 1. Labels for item discrimination

Very High	$\eta_{MDISC} \geq 1.7$

High	$1.35 \le \eta_{MDISC} < 1.7$
Moderate	$0.65 \le \eta_{MDISC} < 1.35$
Low	$0.35 \le \eta_{MDISC} < 0.65$
Very Low	$\eta_{MDISC} < 0.35$

Table 2. Labels for item difficulty

Very Hard	$b_{MDIFF} \ge 2$
Hard	$0.5 \le b_{MDIFF} < 2$
Medium	$-0.5 \le b_{MDIFF} < 0.5$
Easy	$-2 \le b_{MDIFF} < -0.5$
Very Easy	$b_{MDISC} < -2$

Figure 2 (Reckase, 2009, p. 119) demonstrates the interpretation of η_{MDISC} and b_{MDIFF} in MIRT models.

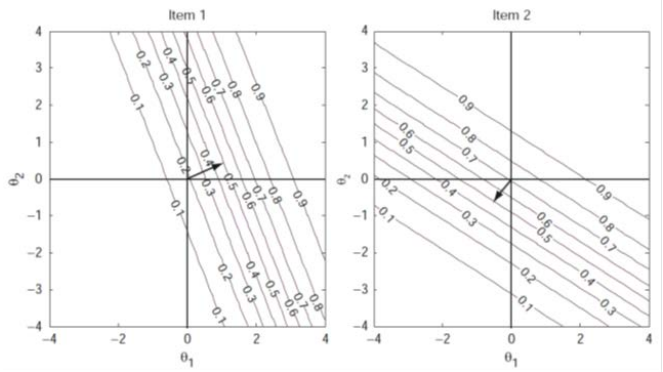

Figure 2. IRS contour lines of MIRT

The IRS contour lines of Items 1 and 2 infer that:

- Closer contour lines in Item 1 mean that the probability of correct response will vary more than that of Item 2 when a student's abilities change. To put it another way, Item 1 has greater discrimination than Item 2.

- Item difficulty was represented by a vector from the origin to 0.5 contour line at IRS steepest slope. If the vector points downward to the left side, the item is supposedly easy, otherwise, difficult. Therefore, it can be concluded from Figure 2 that Item 1 is more difficult than Item 2.

3. Factor analysis

TheKMO Index and Bartlett's Sphericity Test

The KMO Index and Bartlett's Sphericity Test are often applied prior to factor analysis to ensure adequacy. Ranging from 0 to 1, the higher the KMO Index is, the more efficient factor analysis is. Further details about how to interpret this index can be retrieved from Kaiser (1974). For the same purpose, Bartlett's Test is concerned with verifying that variances are equal across groups or samples. If the correlation is perfect, only one factor is counted. Alternatively, the quantity of factors is equivalent to that of observed variables if they are orthogonal (null hypothesis). For factor analysis to be recommended suitable, p-value must be less than 0.05 (Barlett, 1951).

Factor analysis

In addition to establishing underlying dimensions, 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 (Guttman, 1954; Kaiser, 1960). However, Gorsuch (1983) held that the Kaiser–Guttman rule might be subjective and merely applicable to a large sample with less than 40 factors. Cattel (1966) proposed the use of Scree Plot to make the factor selection more convincing. By drawing a straight line through smaller eigenvalues, the point where the factors curve above this line identifies the number of factors (Williams, Brown &Onsman, 2012).

Another important aspect that needs mention is the Rotated Component Matrix.Rotation maximizes high item loadings and minimizes low item loadings, thereby producing a more interpretable and simplified solution (Thurstone, 1947; Cattel, 1978). There are two common rotation techniques - orthogonal rotation and oblique rotation. These days,the most popular rotation method is Varimax rotation developed by Kaiser (1958). With this method, 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 (Krabbe, 2016).

Taking the above-mentioned research as guidelines, the researcher adapted M3PL model for validating a multiple choice test for non-English majors, which so far has not been investigated statistically and appropriately.

III. OBJECTIVE AND 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 final test. Therefore, the following questions were examined for the test development:

- How many intended dimensions involve in the test?
- How can the difficulty, discrimination and guessing parameter of each item in the test be estimated?
- Of all the models IRT, M2PL and M3PL, which is the most suitable for the data? In that case, is the chosen model better-fit than the one suggested in the course learning outcomes?

2. Instruments

The data for this study was gathered randomly from 488 students taking the English final test at a university in Ho Chi Minh city, Vietnam. 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). According to Wainer and Kiely (1987), items of Reading Comprehension (RC) section are better treated as testlet data to control the local dependence. When comparing Testlet 2PL with MIRT models, Min and He (2014) reached the same conclusion that Testlet models provide more appropriate analyses for RC tests. That justifies why in this study, only 30 multiple-choice items of the

two fill-in sections (Items 1-30) were investigated for students' intended abilities. Henceforth, these first 30 items will be referred to as the test for more convenience.

R is a free software used for statistical computing in recent research because of its flexibility (Kelley, Lai & Wu, 2008; Vance, 2009). The package distributed by R can be easily downloaded free of charge at http://CRAN.R-project.org. For factor analysis, FactorMineR, Psych and REdaS were utilized. The mirt package was applied when it came to M3PL and ANOVA.

3. Methodology

Previous research has confirmed the multidimensionality inherent in most tests. In this study, M3PL was adapted to investigate item difficulty, discrimination and guessing parameter of the first 30 questions as in Li et al. (2012). Firstly, Bartlett's Test and KMO Index were exploited to determine whether the data was indeed multidimensional. Then Principal Component Analysis (PCA) was conducted to bring out strong patterns in the dataset. With some idea about the underlying constructs, Varimax rotation was applied for identifying the most significant evidence. Parallel analysis played the role of Confirmatory Factor Analysis (CFA) to confirm the factor structure. In this analysis, actual eigenvalues are compared with random order eigenvalues. Factors are retained when actual eigenvalues surpass random ordered eigenvalues (Williams, Brown &Onsman, 2012).

After that, to determine the appropriateness of MIRT models, ANOVA was carried out. The model with smaller AIC, BIC or $-\text{Loglikelihood}$ is supposed to be more suitable (Rijmen, 2010). The final stage is an illustration of how item difficulty and discrimination can be appraised by estimating MIRT parameters such as slope a_i, intercept d_i and guessing parameter c_i and applying Formulas (3) and (4).

IV. DATA ANALYSIS

1. Test dimensionality

The packages REdaS and psych were utilized for Bartlett's Test and KMO Index.

Table3. The results of Bartlett's Test and KMO Index

Keiser – Meyer – Olkin Statistics	KMO criterion: 0.786
Bartlett test	$\chi^2 = 1434.766$ $df = 435$ p-value < 0.0001

According to Kaiser (1974), with KMO = 0,786 and p-value smaller than 0,0001, it can be concluded that factor analysis is applicable to the data.

Then came Principal Component Analysis (PCA) with the use of FactoMineR. The number of factors can be determined by selecting those for which the Eigenvalue 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).

Table4. Principal Component Analysis

	Eigenvalue	Percentage of VAR	Cumulative percentage of VAR
Component 1	4.0232774	13.410925	13.41092
Component 2	1.4190181	4.730060	18.14099
Component 3	1.3688881	4.562960	22.70395
Component 4	1.3006708	4.445569	27.03951
Component 5	1.2396500	4.132167	31.17168
Component 6	1.1893436	3.964479	35.13616
Component 7	1.1464178	3.821393	38.95755
Component 8	1.1132705	3.710902	42.66845
Component 9	1.0845349	3.615116	46.28357
Component 10	1.0379547	3.459849	49.74342
Component 11	1.0174491	3.391497	53.13492
Component 12	0.9876690	3.292230	56.42715
Component 13	0.9621180	3.207060	59.63421
…			

The eigenvalues are reported in Table 4. Among the 11 components (i.e. factors) meeting the rule, the first five components have eigenvalues much greater than 1 (i.e. 4.023, 1.419, 1.368, 1.300, 1.240), which strongly proves multidimensionality. The percentage of variance illustrates the variance proportion of observed variables. For example, 13.41% of variance of the first factor indicates that 13.41% of the variation can be explained. The cumulative percentage of variance of the 5 chosen factors accounts for 31.17%, which is equivalent to the results in Li et al. (2012). Moreover, the selection of these 5 factors satisfies the rule of Holzinger and Harman (1941). A corresponding Scree plot of the PCA is shown in Figure 3 for the pattern. Furthermore, Williams, Onsman and Brown's (2012) rules when drawing a Scree Plot indicated that components 1, 2, 3, 4, 5 are meaningful to MIRT models.

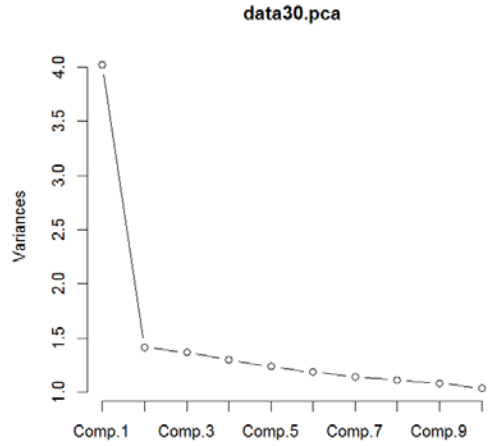

Figure 3. Scree plot của PCA

Parallel analysis acted as CFA shows the following results:

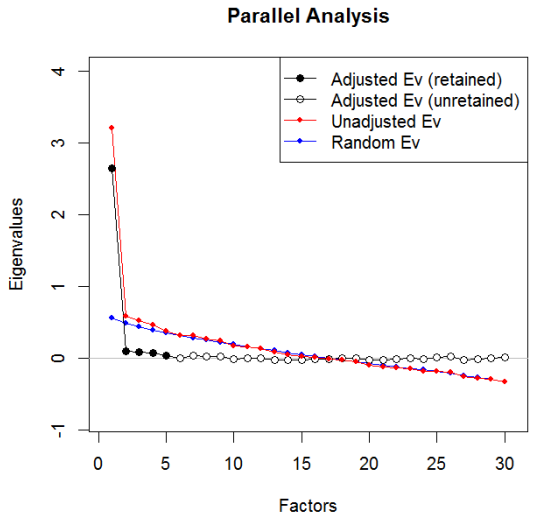

Figure 4. Parallel analysis

Subsequently, Varimax rotation once again proved that the 5 factors examined in PCA are eligible.

Table5.Varimax rotation

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5
Item1	0.248	0.380	0.159	0.105	
Item2	0.262	0.107		0.359	0.132
Item3	0.177		0.213	0.187	0.106
Item4	0.190		0.138	0.279	
Item5	0.223	0.134		0.293	
Item6	0.269	0.278			
Item7	0.302	0.238	0.170		
Item8		0.308			
Item9	0.185	0.155			0.315
Item10	0.163	0.279	0.164	0.148	0.286
Item11		0.130		0.423	
Item12				0.126	
Item13	0.442			0.124	
Item14	0.161			0.770	
Item15	0.308	0.247			
Item16	0.153	0.350		0.122	0.112
Item17		0.247			
Item18	0.217	0.231			
Item19		0.112		0.399	-0.117
Item20	0.402	0.209		0.115	0.151
Item21	0.248	0.178			
Item22	0.486			0.157	
Item23		0.444		0.134	0.239
Item24	0.139	0.327			
Item25	0.202			0.100	0.261
Item26	0.111			0.258	0.659
Item27				-0.178	-0.161
Item28					-0.148
Item29		0.159			0.185
Item30	0.191	0.255			

The values presented in columns Factor1, Factor2, etc. are item loadings for each factor. As Gorsuch (1983) put it, factors after orthogonal rotation are often uncorrelated. Therefore, χ^2 test was conducted for models with 2, 3, 4, 5 factors to investigate their correlation. The null hypothesis assumed no relationship among these variables. The results can be found in Table 6.

Table 6. χ^2 test

No. of factors	χ^2	Degree of freedom	p-value
2	$\chi^2 = 468.55$	$df = 376$	$p = 0.0008$
3	$\chi^2 = 405.50$	$df = 348$	$p = 0.0181$
4	$\chi^2 = 351.07$	$df = 321$	$p = 0.1190$
5	$\chi^2 = 308.33$	$df = 295$	$p = 0.285$

In this model, 4 factors are expected to account for the 4 content areas of the language assessment (i.e. Vocabulary, Grammar, Functions of Speech and Reading Comprehension, if any). However, Table 6 shows that the models may have more than 4 factors. As Riese et al. (2000) asserted the need for factor overestimation, 5 factors were chosen for M3PL model and classified as follows:

- Factor 1: Items 1,2,3,4,5,6,7,9,10,13,14,15,16,18,20,21,22,24,25,26,30.
- Factor 2: Items 1,2,5,6,7,8,9,10,11,15,16,17,18,19,20,21,23,24,29,30.
- Factor 3: Items 1,3,4,7,10,13,14,16,19,23,25,27.
- Factor 4: Items 1,2,3,4,5,10,11,12,20,22,23,24,25,26,27.
- Factor 5: Items 2,3,9,10,16,18,19,20,26,28,29.

2. Model suitability

7 models including IRT 2PL, IRT 3PL, M2PL (4 factors), M3PL (4 factors), M2PL (5 factors), M3PL (5 factors) and LO-3PL to evaluate data fit, in which LO-3PL represents MIRT

model with 3 factors designated for the first 30 questions (as stated in the Learning Outcomes, i.e. Vocabulary, Grammar and Functions of Speech).

Table 7. Goodness of fit

	Estimation Model						
	IRT-2PL	IRT-3PL	M2PL4	M3PL4	M2PL5	M3PL5	LO-3PL
–loglikelihood	8793.661	8758.690	8718.171	8697.089	8718.774	8694.642	8944.445
AIC	17707.32	17697.38	17644.34	17662.18	17655.49	17667.28	18078.89
BIC	17958.74	18074.51	18080.14	18223.68	18112.23	18249.74	18476.97

Goodness of fit indices are reported in Table 7 for the four pairs of models. M3PL5 model with smaller AIC, BIC and –loglikelihood tends to fit significantly better than unidimensional models, models without guessing parameter and the LO model.

3. M3PL parameter estimation

The package mirt was then employed for M3PL model with 5factors. The outcomes are summarized in Table 8.

Table 8.Parameter estimation of M3PL model with 5 factors

	a1	a2	a3	a4	a5	d	g	u
$Item1								
par	1.183	1.426	0.536	0.665	0	0.423	0.238	1
$Item2								
par	0.793	0.235	0	1.761	0.202	1.242	0	1
$Item3								
par	0.439	0	0.88	0.557	0.198	0.873	0.001	1
$Item4								
par	0.6	0	0.332	1.005	0	0.645	0.038	1
$Item5								
par	0.507	0.339	0	0.912	0	0.104	0	1
$Item6								
par	1.029	0.639	0	0	0	0.199	0	1
$Item7								
par	6.142	0.666	1.651	0	0	-0.734	0.342	1
$Item8								
par	0	4.563	0	0	0	-0.115	0.233	1
$Item9								
par	0.911	0.16	0	0	1.349	0.402	0.173	1
$Item10								
par	4.566	4.736	1.959	0.469	4.752	0.442	0.409	1
…								

The values in columns a_1, a_2, a_3, a_4, a_5 indicate the slopes of each item, d intercept and g guessing parameter. 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 (3) and (4) were calculated.

Table 9. Discrimination and Difficulty

	a_1	a_2	a_3	a_4	a_5	d	g	η_{MDISC}	b_{MDIFF}
Item 1	1.183	1.426	0.536	0.665		0.423	0.238	2.04	-0.21
Item 2	0.793	0.235		1.761	0.202	1.242	0	1.96	-0.63
Item 3	0.439		0.880	0.557	0.198	0.873	0.001	1.15	-0.76
Item 4	0.600		0.332	1.005		0.645	0.038	1.22	-0.53
Item 5	0.507	0.339		0.912		0.104	0	1.1	-0.09
Item 6	1.029	0.639				0.199	0	1.21	-0.16
Item 7	6.142	0.666	1.651			-0.734	0.342	6.39	0.11
Item 8		4.563				-0.115	0.233	4.56	0.03
Item 9	0.911	0.160			1.349	0.402	0.173	1.64	-0.25
Item 10	4.566	4.736	1.959	0.469	4.752	0.442	0.409	8.36	-0.05
Item 11		0.645		0.854		1.637	0.002	1.07	-1.53
Item 12				3.993		-8.349	0.243	3.99	2.09
Item 13	1.916		0.461			0.307	0.08	1.97	-0.16
Item 14	0.546		4.897			-0.462	0	4.93	0.09
Item 15	1.904	0.695				-1.535	0.225	2.03	0.76
Item 16	0.647	0.902	0.091		0.198	1.381	0.002	1.13	-1.22
Item 17		1.079				-0.644	0.385	1.08	0.6
Item 18	1.101	0.567			0.714	-1.261	0.118	1.43	0.88

Item 19		0.276	2.782		-0.312	-0.801	0.003	2.81	0.28
Item 20	0.988	0.875		0.572	0.66	-0.079	0.194	1.58	0.05
Item 21	0.667	0.615				-0.939	0.13	0.91	1.03
Item 22	1.474			0.894		-0.856	0.1	1.72	0.5
Item 23		2.035	0.396	0.565		0.73	0.02	2.15	-0.34
Item 24	0.629	1.397		0.968		0.888	0.203	1.81	-0.49
Item 25	0.455		0.308	0.599		0.831	0.005	0.81	-1.02
Item 26	0.663			0.988	1.691	2.216	0	2.07	-1.07
Item 27			-2.048	-3.775		-8.591	0.135	4.29	2
Item 28					-3	-6.603	0.252	3	2.2
Item 29		0.401			0.567	0.166	0.002	0.69	-0.24
Item 30	0.543	0.543				0.371	0.001	0.77	-0.48

Based on Baker (2001) and Hasmy (2014)'s labels for item difficulty and discrimination (as shown in Tables 1 and 2), it can be deduced that:

- Item discrimination ($^{\eta}MDISC$): over 50% (16 items) have Very High discrimination. 3 items (10%) are at High level. About 35% (11 items) fits in Moderate level. There are no questions at Low and Very Low levels. With a majority of items (60%) at good discrimination levels, the test is supposed to differentiate well among students.

- Item difficulty ($^{b}MDIFF$): over 20% (7 items) are categorized as Easy; half of the items as Medium. 20% of the items are ranked as Hard and Very Hard. With 80% items at Medium and below levels, the test is not so challenging, but the distribution of item difficulties can be considered adequate for student level assessment.

Further analysis navigated our concern to some test items. Items 12, 27 and 28 can be regarded as well-designed with high levels of difficulty and discrimination. Having the highest discrimination among 30 items, Items 7, 8, 10, 14 are supposed to be reusably good. However, Items 17 and 21 need revision as they are at high difficulty level but moderate discrimination.

Most of the questions involve decent amount of guessing behavior, except for Items 7, 10, 17 (with guessing parameters of 0.409, 0.385 and 0.342, respectively). That these items are at fairly high difficulty level may lead to the fact that guessing behavior (rather than ability) can promote the possibility of positive answers. Furthermore,other factors rather than linguistic knowledge (General Belief, Contextual Knowledge or Logical Thinking) may influence a student's response. Taken Item 21 as an example, its intended ability of Speech Functions is inferior to Grammar knowledge, thereby causing students confusion over choices.

V. DISCUSSION AND IMPLICATION

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 M3PL is the best fitting model. The results reflect overlapping trait issues inherent in the test as in any kind of assessment, which reinforces Shedl et al. (1996) and Wilson (2000)'s idea of multidimensionality. Unlike Li et al. (2012), the above-mentioned selection of factors went beyond the 4 content areas of the language assessment for the following reasons:

- Besides the 4 intended abilities listed in the learning outcomes, a student's response may be manipulated by other non-linguistic factors such as General Belief (Item 8), Contextual Knowledge (Item 17) and Logical Thinking (Item 12).

- Factor analysis and ANOVA revealed that the model with more than 4 factors is better fitting than the Learning Outcome model. Moreover, according to Riese et al. (2000), it is better to overestimate the number of dimensions.

χ^2 test was employed to evaluate the goodness of fit of IRT, M2PL and M3PL models, which demonstrated that M3PL model with guessing parameter has the best data fit. This model with the emergent factors has been proved to measure students' real abilities. In addition, the right factor classification acts as a premise for the next steps of estimating item difficulty and discrimination.

More often than not, high-challenge questions tend to distinguish well among students. Nonetheless, there are cases in which difficult items have mediocre discrimination (for example, Item 17), which can be justified by guessing parameter. High values of guessing parameter acknowledge that students' guesswork, rather than knowledge, may engage in figuring out the answers.

All the 5 chosen factors turn out to involve a combination of different skills and abilities, which makes it hard to get them labeled. Especially, Grammar has emerged among the skills as a prerequisite for students' best performance. In addition, the knowledge of functions of speech is also an indispensable source for their comprehending the test questions.

VI. CONCLUSION

To sum up, more qualitative item analyses will be needed in future research to determine 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 multiple dimensions using Bartlett's Test and KMO Index.

- 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.

- Then ANOVA is done to confirm the best fitting model.

- 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, J. (2002). Language testing and assessment.*Language Testing, 35*(2), 79-113.

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

[3] Bartlett, M.S. (1951). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika, 38*(3), 337-344.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’sability. In F.M. Lord & M.R. Novick (Eds), Statistical theories of mental test scores (pp. 395-479). Reading, MA: Addison-Wesley.

[4] Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters:Application of an EM algorithm. Psychometrika, 46(4), 443-459.

[5]

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

[7] Cattel, R.B. (1966). The scree test for the number of factors.*Multivariate Behavioral Research,1*(2), 245-276.

[8] Cattel, R.B. (1978). Matched determiners vs. factor invariance: A reply to Korth.*Multivariate Behavioral Research, 13*(4), 431-448.

[9] DeMars, C.E. (2007). “Guessing” parameter estimates for multidimensional item response theory models. *Educational and Psychological*

Measurement, 67(3), 433-446.

[10] Do, T. H. (2016). Applying multidimensional item response theory in validating an English final test.*Journal of Technical Education*

Science, 36, 103-110.

[11] [Embretson, S.E., &Reise, S.P. (2000). *Item reponse theory for psychologists.* Mahwah, NJ: Lawrance Erlbaum Associates.

[12] Gorsuch, R.L. (1983). *Factor analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[13] Guttman, L. (1954). Some necessary conditions for common-factor analysis.*Psychometrika, 19*(2), 149-161.

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

[15] 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.

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

[17] Holzinger, K.J., & Harman, H.H. (1941). *Factor analysis: A synthesis of factorial methods.* The University of Chicago Press, Chicago.

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

[19] Kaiser, H.F. (1960). The application of electronic computers to factor analysis.*Educational andPsychology Measurement, 20*(1), 141-151.

[20] Kaiser, H.F. (1974). An index of factorial simplicity.*Psychometrika, 39*(1), 31-36.

[21] Kelley, K., Lai, K., & Wu, P.J. (2008). Using R for data analysis: A best practice for research. In J.W. Osborne (Ed.), *Best advanced practices in quantitative methods* (pp. 535-572). Thousand Oaks, CA: Sage.

[22] Kose, I.A., &Demirtasli, N.C. (2012). Comparison of unidimensional and multidimensional models based on item reponse theory in term of both variables of test length and sample size. *Procedia-Social and Behavior Sciences, 46*, 135-140.

[23] Krabbe, P. (2016). *The measurement of health and health status.* San Diego: Elservier.

[24] Le, A.V., Doan, H.C., & Pham, H.U. (2016). Applying 3-parameter logistic model in validating the difficulty, discrimination and guessing

of items in a multiple-choice test. *Journal ofScience, University of Pedagogy, 85*(7), 174-184.

[25] Li, Y.H., &Lissitz, R.W. (2004). Application of the analytically derived asymptotic standard error of item response theory item parameter

estimate. Journal of Educational Measurement, 41(2), 85-117.

[26] Li, Y., Jiao, H., &Lissitz, R.W. (2012). Applying multidimentsional item response theory in validating test dimensionality: An example of

K-12 large-scale science assessment.*Journal of Applied Testing Technology, 13*(2), 1-27.

[27] Lord, F.M. (1952). A theory of test scores.*Psychometric Monographs, 7*, Richmond, VA: Psychometric Coporation. Retrieved from

http://www.psychometrika.org/journal/online/MN07.pdf

[28] Lord, F.M., &Novick, M.R. (1968). *Statistical theory of mental test scores*. Reading, MA: Addition-Wesley.

[29] Lord, F.M. (1980). *Application of item response theory to practicle testing problems*.Lawrence Erlbaum Associates, Hillsdale, NJ.

[30] Min, S. and He, L. (2014). Applying unidimentional and multidimentional item reponsetheory models in testlet-based reading assessment.

Language Testing, 31(4), 453-477.

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

[32] Mulaik, S.A. (1972). A mathematical investigation of some multidimensional Rasch model for psychological tests. Paper presented at the annual meeting of the Psychometric Society, Princeton, New York.

[33] Nguyen, T.H.M., & Nguyen, D.T. (2004). Đolường và đánh giá trong thi trắc nghiệm khách quan: Độ khó câu hỏi và khả năng của thí sinh. Journal of Science, VNU Ha Noi, 197-214.

[34] Nguyen, B.H.T. (2008). Sử dụng phần mềm Quest để phân tích câu hỏi trắc nghiệm khách quan.*Journal of Science and Technology, 2*, 119-126.

[35] Nguyen, T.N.X. (2014). Sử dụng phần mềm Quest/ConQuest để phân tích câu hỏi trắc nghiệm khách quan.*Journal of Science, TraVinh* University, 12, 24-27.

[36] Reckase, M.D. (1972). Development and application of a multivariate logistic latent trait model (Unpublished doctoral dissertation). Syracuse University, Syracuse NY.

[37] Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

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

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

[40] 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.

[41] Sympson, J.B. (1978). A model for testing with multidimensional items.In Weiss D.J. (Ed). Proceeding of the 1977 Computerized

Adaptive Testing Conference, University of Minnesota, Minneapolis.

[42] Thurstone, L.L. (1925). A method of scaling psychological and education test. *Journal of Education Psychology, 16*, 433-451.

[43] Thurstone, L.L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

[44] Vance, A. (2009). *Data analysis captivated by R’s power*.New York Times. Retrieved from

http://www.nytimes.com/2009/01/07/technology/busine%20ss-computing/07program.html/?pagewanted=all

[45] Wainer, H., &Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185–202.

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

[47] Whitely, S.E. (1980a). Measuring aptitude processes with multicomponent latent trait models. (Technical Report No.NIE -80-5). University of Kansas, Lawrence, KS.

[48] Whitely, S.E. (1980b). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

[49] Williams, B., Brown, T., &Onsman, A. (2012). Exploratory factor analysis: A five step guide for novices. *Australasian Journal of* Paramedicine, 8(3), 1-13.

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

[51] Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

AUTHORS

First Author – Do Thi Ha, MA, Institutional Affiliation: Ho Chi Minh City University of Technology and Education,
Ho Chi Minh City, Vietnam, Email address: dothiha1985@gmail.com