SEAAIR
SOUTH EAST ASIAN ASSOCIATION FOR INSTITUTIONAL RESEARCH

Journal of Institutional Research South East Asia

JIRSEA Issue: Vol. 16 No. 2

September/October 2018

ISSN 1675-6061

JIRSEA is indexed with SCOPUS, the Directory of Open Access Journals and EBSCoHost (Education Research Index/Education Research Complete). As provided in the Ulrich's Directory, all journal in the DOAJ are also indexed in the Ulrich's Directory

The Reliability and Validity of Test Construction from Item Bank based on Item Response Theory

Suwimon Kritkharuehart

Department of Educational Evaluation and Research,
Faculty of Education, Ramkhamhaeng University, Bangkok, Thailand
(suwimon.k@rumail.ru.ac.th)

ABSTRACT

Item banks indicate the advancement of academic administration of educational institutions, in particular the test development that aims to produce more than one parallel test. The purpose of a parallel test is to make the scores of two certain tests equivalent, which renders the scores interchangeable. Furthermore, nowadays computers are able to produce parallel tests automatically. This research endeavoured to construct an Automated Test Assembly (ATA) programme from an item bank capable of producing tests, and to test the accuracy of the test assembly programme in regard to its test parallelism and test overlap rate. The data in this research were obtained from the simulation of test results in relation to the actual test results of 2,800 examinees. The statistical approaches used in this research were the followings: Exploratory Factor Analysis (EFA), 3-parameters Item Response Theory (3 PL-IRT), and Test Overlap Rate. The result of the research suggested that this programme was effectively capable of producing parallel tests, which were used to measure the academic performance of examinees in certain levels of ability. However, when stipulating some more conditions on achievement tests, which matched the various abilities of examinees, the programme became less effective in creating parallel tests. Moreover, some more conditions added on test overlap rate between each test resulted in the inaccuracy with respect to test parallelism. In short, the more added conditions of creating a test paper there were, the less ineffective in creating parallel tests the programme became. The result of this study contributed to the production of parallel tests provided by the ATA programme. The tests created were valid and reliable. Their reliability and validity were measured by the following criteria: their comprehensiveness of the domain of content, their test parallelism, which is one of the important characteristics of achievement tests, their interchangeability in each testing period, and their ability to prevent test leaks effectively.

Keywords: item bank, test assembly, parallel tests, 3-parameters IRT

Introduction

It is generally acknowledged that educational measurement and evaluation are two important processes which determine the success of tertiary education. This helps to create employable graduates who satisfy the requirements of the labor market. Universities should, therefore, acquire a systematic management to affect their educational measurement and evaluation with efficacy.

The three following components are essential for educational measurement and evaluation: items, tests, and examinations. As a result, item bank play a pivotal role in effectively administering educational tests with large scale testing in which each student takes the test more than one time (McAlpine, 2002). In addition, instructors can also select a test which matches the objectives of an examination. By means of using an item bank, which categories items into different groups with respect to their content, to produce more than one parallel test with content validity, tests can be used again with the same student without repeatedly using the same test. This also makes the interchangeability of each different test used at different times possible. According to IRT, the characteristic of the tests mentioned above identifies test parallelism. Moreover, computer programmes nowadays enable instructors to automatically produce parallel tests from item bank via stipulating conditions on a test, such as the amount of contents, the number of questions, the difficulty index and the discrimination index of the tests, which meet the objectives of an examination.

According to the result of numbers of research previously conducted, there were differences in the accuracy of the ATA programme. Lin (2008) investigated the accuracy of the ATA programme, which was developed from weighted deviations model (WDM) heuristic, that aimed to produce parallel tests. It was found that the programme was able to produce six tests with test parallelism, and the tests produced were capable of accurately predicting the scores of examinees in two levels of ability, i.e. being fairly high, and being high. In the light of Chen & Lei (2009), Lin (2010) conducted a further study in which the accuracy of the ATA programme was investigated by using the function that regulated test overlap rate. The study revealed that the ATA programme was capable of producing ten tests with test parallelism, and the tests produced were, with accuracy, capable of predicting the scores of students only in one level of ability– being moderate.

This study further investigated the application of the instruction set lp_Solve Version 5.5 of the R programme which was effective in selecting items due to its algorithm that solves mathematical problems under the two following principles: the principle of decision-making in selecting mathematical variables and the principle of the least aberration. According to van der Linden (2005), the study also stipulated certain conditions according to which items were randomly selected so as to produce tests in relation to the quantitative conditions in the test set level by determining the ratio of the lowest and the highest numbers of items in each content of measurement by using this equation: $\sum x_i \leq n_c^{max}$, $\sum_{i \in V_c} \geq n_c^{min}$. This also offered laypersons an opportunity to use the programme without infringing copyright and to examine the accuracy of the ATA programme developed from tests that have the capability for

estimating the scores of students in four levels of ability; namely, being quite low (θ = -1.50), being moderate (θ = 0.00) , being fairly high (θ = 1.50), and being high

(θ = 2.50). Therefore, the researcher had great interest on developing the ATA programme for producing tests from the item bank, and testing the accuracy of the programme in relation to its test parallelism and test overlap rate. The criterion for assessing the accuracy of the programme was whether the programme achieved the efficiency in producing tests capable of estimating the scores of students in four levels of ability – being quite low, being moderate, being fairly high, and being high. The findings in this research shall be beneficial to the production of parallel tests from the item bank which ensures the reliability of the estimation of the scores of students with the least aberration, and also with low test overlap rate, indicating that items are secured, and the prevention of test leak when each paper is used again.

Related Literature

Test design

Designing how students ought to learn that matches individual differences has become more and more common in educational institutions. Hence, item banks are considered indispensable for facilitating instructors when it comes to selecting items from an item bank so as to produce a test, which can be used differently in accordance with the aim of an examination. For example, the test can be used to measure and evaluate examinees' academic achievement, or to periodically keep track of examinees' learning progress, or to promote test enhanced learning, when using different tests in relation to learning objectives. It follows, therefore, that designing any test paper must aim to produce a test paper that possesses the ability to estimate the scores of examinees. As regards the discrimination of the type of test design, the two following theories are used: the Classical Test Theory (CTT) and the Item Response Theory (IRT). Designing tests with reference to the IRT is suitable for both measuring and evaluating academic achievement, periodically monitoring examinees' progress, and promoting test enhanced learning.

According to van der Linden (2005), using IRT for test design offers two advantages. First, tests are selected in regard to the probability that examinees would correctly choose the right answers. In other words, it is to select items suitable for the examinees' level of ability. Subsequently, the items will be collected and used as a test in order to measure and evaluate individual achievement. This is regarded as an effective implementation of information technology of item banks. Second, the scores of examinees examined by a parallel test paper, in which each item belonged to the different test papers, can be equated. This ensured the interchangeability of the test result. The process of developing the ATA programme from the item bank in order to produce a test paper comprises three stages:

(1) Designing the item bank. This stage involves determining the content which will be used in constructing the item bank. The number of the items will be enumerated in relation to each

sub-content on one condition: namely, the number of the items must be either equal to or ten times more than the number of the actual items. The following three parameters are used with reference to the objectives of the examination: difficulty, discrimination, and guessing.

(2) Developing the item bank. This stage deals with test writing, test reviewing and improving, and also test trialling and selecting IRT model for assessing the quality of a test. The IRT model consists of 1 PL-IRT, 2 PL-IRT, and 3 PL-IRT. 1 PL-IRT considers the difficulty parameter of items in order to produce a test which is suitable for the ability level of examinees. 2 PL-IRT considers the difficulty and discrimination parameters of items in order to produce a test which is suitable for the ability level of examinees. 3 PL- IRT considers the difficulty, discrimination, and guessing parameters of items in order to produce a test which is suitable for the ability level of examinees. The final step is assessing the quality of items and a test paper with regard to the selected model.

(3) Determining the characteristics of the test which needs to be produced. The characteristics consist of the number of sub-contents used in each item, the number of items enumerated in relation to sub-contents, the parameters of the paper, i.e. difficulty, discrimination, guessing, information function appertaining to the reference test, and test overlap rate.

Test Parallelism

According to McDonald (2009), in Item Response Theory Model, parallel measurements constitute a mathematical requirement that the items in parallel tests themselves matched item parameters – equal item means and loadings in the linear model or equal item parameters. In applications, it is expected that conditions will be stipulated on the substantive content of the items which compose each form and it is a condition for equity. Equity is the requirement that makes the administration of a form of test unknown to examinees. Any two forms will be item-parallel if they share the following features:

1. Equal test characteristic curves – obtained by summing the equated item characteristic curves;
2. Equal test information functions – obtained by summing equated item information function;
3. Equal test-score information functions (and, more generally, equal formula-score information functions).

As a result, the true scores and the error variances of the two forms also matched at every point on the scale of measurement.

Test Information Function

As stated by Baker (2001), the interest, in IRT, was on estimating the value of the ability parameter for an examinee. If the amount of information is adequate, the examinee whose true ability is at that level can be estimated with precision, i.e. all the estimates will be reasonably close to the true value. On the contrary, if the amount of information is inadequate, the ability of the examinee cannot be estimated with precision and the estimates will be widely scattered about the true ability. Furthermore, de Ayala (2009) stated that test information function might be used to design an instrument with specific characteristics.

This capacity takes advantage of the fact that items and persons are located on the same continuum as well as the capacity to assess the amount of information in order to estimate person locations solely based on the item parameter estimates. Success in developing an instrument whose observed total information function is similar to the target information function depends on having an adequate pool of items to work with and on imposing constraint on the item selection so as to ensure that the resultant instrument has validity with respect to the construct of interest. Moreover, any test paper with a high discrimination parameter and a low guessing parameter has a high information function parameter – its predictive ability of the scores of examinees is accurate.

Test overlap rate

Lin (2010) affirmed that test overlap control is important, and might probably be more crucial for automated test assembly (ATA). In the context of assembling equivalent test forms, the test overlap rate could be extremely high because the items selected to fulfill the constraints, for example, target test information function, are likely to be the same across multiple test forms without exposure control. One of the goals of the test assembly process should be the minimization of test overlap rate – the percentage of items shared between any two forms. In automated test assembly, one way to achieve this is to include item usage as another constraint or target in the solution of the assembly problem.

Generally, an item with a high discrimination parameter will have more item overlap rate than that with a lower discrimination parameter because a test with a high discrimination parameter will make the estimation of Item Characteristic Curve (ICC) accurate and suitable for the ability of an examinee (θ_p) which leads to higher Test Information Function (TIF) (Chang & Ying, 1996 as cited in Chang & Zhang, 2002). Way (1998 as cited in Chang & Zhang, 2002) maintained that the control of item exposure rates needs explicit research since the control of item exposure rates should consider both the aspects of rates of the item exposure and the determination of item bank. In addition, it is also suggested that the repetition of using random items should be less than 25%.

Methods

The Objectives

The major objective which this study aimed to achieve was to develop an automated test assembly (ATA) programme from an item bank in order to produce tests that show reliability and validity based on IRT. It also had four minor objectives: namely, first, developing the test assembly programme from an item bank; second, verifying the accuracy of the test assembly programme in respect of its test parallelism which depends on the referenced test; third, verifying the accuracy of the test assembly programme in respect of its low test overlap rate; and fourth, verifying the accuracy of the test assembly programme in respect of its parallelism and test overlap rate.

Participants

The participants in this study were a cohort of 2,800 second-year students who enrolled on the course "Introduction to Statistics and Research in Education" at Ramkhamhaeng University.

Procedure and data analysis

The development of the automated test assembly (ATA) from the item bank in order to produce tests that show reliability and validity based on IRT consists of two stages. The first stage is constructing and verifying the quality of the referenced test based on Classical Test Theory (CTT) and Item Response Theory (IRT). The second stage is constructing, verifying, and selecting items that show the desired qualities according to IRT in order to register them in the item bank.

Stage 1: Constructing and verifying the quality of the referenced test based on CTT and IRT

The referenced test was the test used in the course "Introduction to Statistics and Research in Education". The followings were the procedure of this stage:

(1) The table of specification of the test was constructed by three lecturers. It was found that the test consisted of 100 items with 8 contents. Its numbers of items were 9, 11, 9, 25, 11, 10, 19, 6, respectively;
(2) The test whose items had four multiple choices were produced;
(3) The content validity of the test was examined by five experts and the Index of Item Objective Congruence (IOC) was calculated by using Hambleton's formula (1984). It was found that the test had the index of IOC between 0.80 and 1.00
(4) The test was used by 2,800 students who enrolled on the course. It turned out that there were 64 items that showed the quality based on the CTT with the difficulty index between 0.20 and 0.80, the discrimination index calculated via the biserial correlation from 0.20 upwards, and the KR-20 reliability coefficient of 0.84.
(5) The quality of the items and the test was assessed by 3 PL- IRT. It was found that the referenced test with 64 test items showed the single dominant factor; namely, the ratio between the first eigenvalue and the second eigenvalue was 3.131 which satisfied Gorsuch's (2003) criteria, dictating that the ratio between the first eigenvalue and the second eigenvalue must not be lower than 3.000. Once the test had the single dominant factor, there would be test-answering independence (Hambleton & Swaminathan, 1996). By making model data fit assessment, it was found that the 3PL-IRT test correspondence model corresponded to the empirical data than the 2 PL- IRT test correspondence model with the statistical significance at 0.01; namely, $\chi^2_{\text{II}-\text{III}}$ = 171, 870. 5842 - 171, 273.6639
(6) = 596.9203, and $\chi^2_{0.01,64}$ = 95.6260.
(7) The analysis of the item and test parameters was performed. It was found that the discrimination index was between 0.206 and 3.272, the difficulty index was between (- 2.088) and 2.883, and the guessing index was between 0.001 and 0.315 which met Baker's (2001) and de Ayala's (2009) criteria.

(8) The analysis of the test information function was made. It was found that the referenced test showed a high degree of the assessment accuracy of examinees with moderate ability, followed by those with low ability and fairly high ability, respectively. The referenced test also showed the empirical reliability index of 0.88, suggesting that the referenced test was efficient in assessing the ability of examinees at 77.44%, as shown in Figure 1.

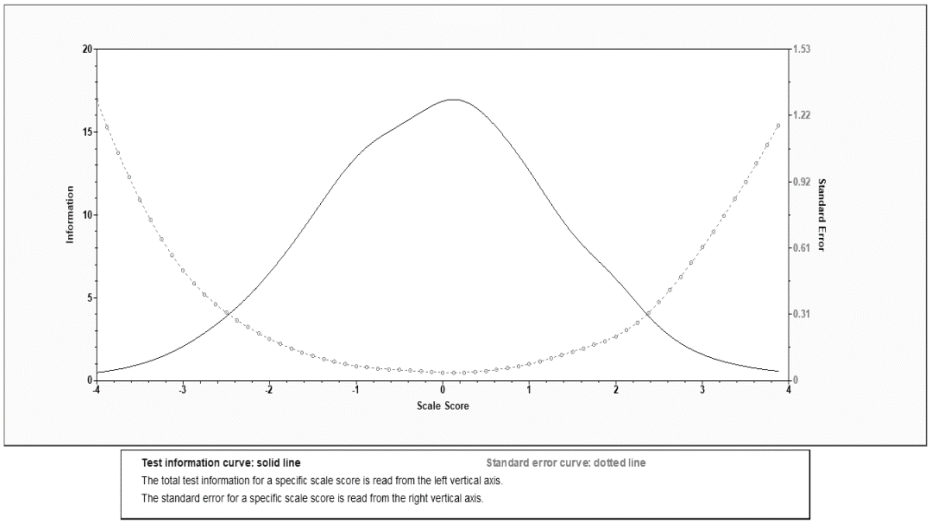


Figure 1 Test information function of referenced test

Stage 2: constructing, verifying, and selecting items which show the desired qualities according to IRT in order to register them in the item bank.

In order to construct, verify, and select items which display the desired qualities according to IRT, this study selected the items that had the difficulty, discrimination, and guessing indexes corresponding to the items in the referenced test by creating the item bank and the result of the test through the simulation programme called WINGEN (Han, 2007). According to van der Linden (2005), the number of the items in the item bank with reference to IRT was equal to or more than ten times of that of the actual items. Regarding the result of the verification and selection of the items in respect of 3PL-IRT, it was found that the simulated item bank had 2,089 test items, all of which had the difficulty index, the discrimination index, and guessing index corresponding to the referenced test, and every test measured the single dominant factor. The result indicated that this item bank had the desired quality with regard to the IRT. The statistical analyses in this stage were the followings: Exploratory Factor Analysis (EFA), Test Parallelism (Luecht, 1998; Chen, Chang, & Wu, 2012), Test Overlap Rate (Chen, Ankenmann & Spray, 2003).

Results

Automated Test Assembly Programme

The automated test assembly (ATA) programme was a modified form of the instruction set called lp_Solve Version 5.5 (Diol & van der Linden, 2013) in R programme. The multiple

assembled tests were created from the collection of the selected test items, as shown in Figure 2.

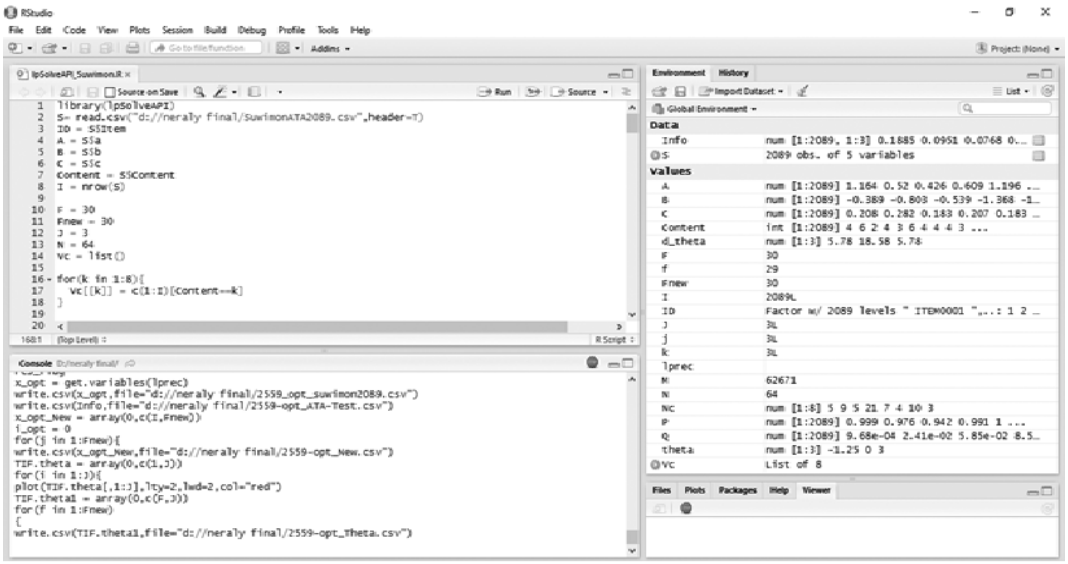


Figure 2 Automated Test assembly programme with R Program

The followings were the details about the ATA programme:

Part 1: Input determination

(1) Determine the specific characteristics of items so as to be collected and used in the test in accord with the referenced test; namely, the assembled test items included 8 contents, the numbers of which were 5, 9, 5, 21, 7, 4, 10, and 3, respectively, the number of tests to be created were 1 to 5 tests, and the number of items per test was 64.

(2) Determine the test information function of different referenced tests based on the ability of examinees ($RTIF_{\theta_k}$), which was further used as a criterion for collecting items so as to produce tests with test information function corresponding to that of the referenced test.

In this research, $RTIF_{\theta_k}$ resolved itself into three different cases: namely, first, the test information function of the referenced test of examinees with moderate ($RTIF_{\theta = 0.00} = 16.80$) and fairly high ability ($RTIF_{\theta = 1.50} = 10.00$); second, the test information function of the referenced test of examinees with quite low ($RTIF_{\theta = (-1.50)} = 10.50$), moderate ($RTIF_{\theta = 0.00} = 16.80$), and fairly high ($RTIF_{\theta = 1.50} = 10.00$); and third, the test information function of the referenced test of examinees with quite low ($RTIF_{\theta = (-1.50)} = 10.50$), moderate ($RTIF_{\theta = 0.00} = 16.80$), fairly high ($RTIF_{\theta = 1.50} = 10.00$), and high ability ($RTIF_{\theta = 2.50} = 5.00$).

Part 2: Procedures and mandatory conditions

(1) Design a decision-making instruction set for selecting items from the item bank which were used in the test which showed binomial distribution:

(2) $x_{it} = \begin{cases} 1 \text{ when the item(i)appeared in the test (t)} \\ 0 \text{ when the item (i) did not appear in the test (t)} \end{cases}$

(3) The instruction set of mandatory conditions of lp_Solve Version 5.5 was created in order to select test items from the item bank and calculate the item information function (IIF), and then calculate the total IIF selected from the programme, which indicated different item information functions in regard to the ability of examinees at k levels, when k = 1, …, K; namely, Test Information Function (TIF_{θ_K}). The test whose items were selected from the item bank ought to have the least different test information function corresponding to the ability of examinees from that of the referenced test ($RTIF_{\theta_K}$).

(4) The instruction set of mandatory conditions for randomly selecting test items was created so as to prevent using the items repeatedly among tests.

(5) The instruction set of how items would be randomly selected was created in order to produce multiple tests with reference to the structure of the referenced test content.

Part 3: The results of procedures and mandatory conditions

(1) The results of test items selected from the item bank to be used in multiple tests. x_{it} = 0 means the item was not selected from the item bank in regard to the mandatory conditions, whilst x_{it}= 1 means the item was selected from the item bank in regard to the mandatory conditions, as shown in Figure 3.

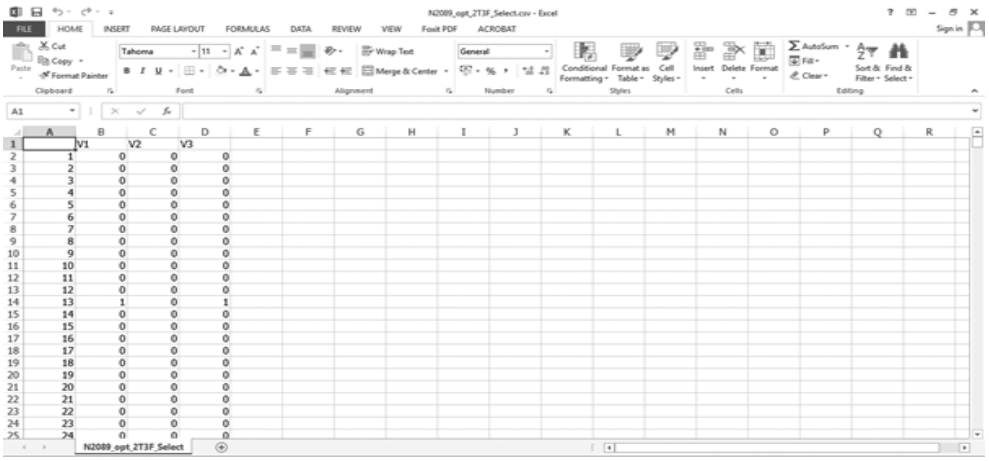
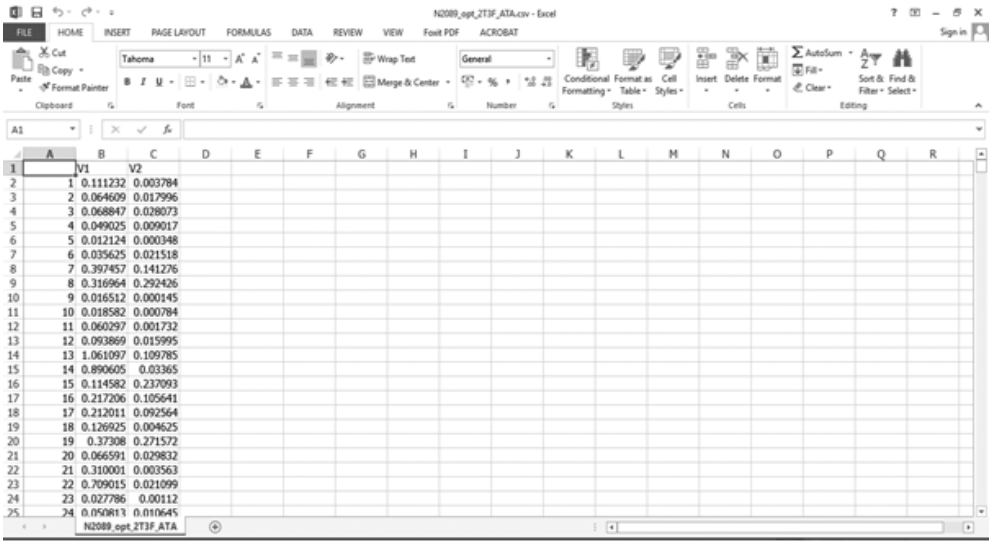


Figure 3 Output of item selection program of item bank

(2) The results of the assembled tests had the Assembled Test Information Function ($ATIF_{\theta_K}$) which had the least difference from the Referenced Test Information Function ($RTIF_{\theta_K}$), as shown in Figure 4.



	V1	V2
1	0.111232	0.003784
2	0.064609	0.017996
3	0.068847	0.028073
4	0.049025	0.009017
5	0.012124	0.000348
6	0.035625	0.021518
7	0.397457	0.141276
8	0.316964	0.292426
9	0.016512	0.000145
10	0.018582	0.000784
11	0.060297	0.001732
12	0.093869	0.015995
13	1.061097	0.109785
14	0.890605	0.03365
15	0.114582	0.237093
16	0.217206	0.105641
17	0.212011	0.092564
18	0.126925	0.004625
19	0.37308	0.271572
20	0.066591	0.029832
21	0.310001	0.003563
22	0.709015	0.021099
23	0.027786	0.00112
24	0.050813	0.010645

Figure 4 The results of the test information function according to the ability of examinees

The result of the accuracy verification of the automated test assembly programme regarding parallelism

McDonald (2009) stated that test parallelism should be considered from the correspondence between the information function of the referenced test ($RTIF_{\theta_K}$) and the automated test information function ($ATIF_{\theta_K}$).

This study determined the test information function according to the ability of examinees (θ) in order to be used in evaluating the accuracy of parallelism between the referenced test and the tests created from the ATA programme. There were three different cases as follows:

- Case 1 - the test information function of the referenced test of examinees with moderate ($RTIF_{\theta = 0.00} = 16.80$) and fairly high ability ($RTIF_{\theta = 1.50} = 10.00$).
- Case 2- the test information function of the referenced test of examinees with quite low ($RTIF_{\theta = (-1.50)} = 10.50$), moderate ($RTIF_{\theta = 0.00} = 16.80$), and fairly high ($RTIF_{\theta = 1.50} = 10.00$).
- Case 3 - the test information function of the referenced test of examinees with quite low ($RTIF_{\theta = (-1.50)} = 10.50$), moderate ($RTIF_{\theta = 0.00} = 16.80$), fairly high ($RTIF_{\theta = 1.50} = 10.00$), and high ability ($RTIF_{\theta = 2.50} = 5.00$).

The criterion adopted for evaluating the accuracy of parallelism was the Mean Square Deviation of Test Information Function (MSD of TIF), which had an index of less than 0.05 (Luecht, 1998; Chen, Chang & Wu, 2012). The followings were the details of each case:

Case 1 – When the automated test assembly programme stipulated the conditions of the test information function of the referenced test according to the two levels of ability of examinees, i.e. 0.00 and 1.50, it was found that the programme was able to produce four tests, two of which had the accuracy of parallelism. The MSD of TIF of the first parallel test and the second test were 0.02 and 0.00, respectively, as shown in table 1.

Table1: The test information function of the referenced test and the assembled test corresponding to the two levels of ability of examinees and the mean square deviation of test information functions of the tests

Levels of ability of examinees	TIF				
	$RTIF_{\theta_K}$	$ATIF_{\theta_k}$			
		Test 1	Test 2	Test 3	Test 4
0.00	16.80	16.85	16.80	17.00	17.03
1.50	10.00	9.80	10.00	10.30	7.50
MSD of TIF	-	0.02	0.00	0.06	3.37
The accuracy of parallelism	-	Being parallel to the referenced test	Being parallel to the referenced test	Not being parallel to the referenced test	Not being parallel to the referenced test

Case 2 - When the automated test assembly programme stipulated the conditions of the test information function of the referenced test according to the three levels of ability of examinees, i.e. (-1.50), 0.00 and 1.50, it was found that the programme was able to produce two tests, none of which showed the accuracy of parallelism, as shown in table 2.

Table 2: The test information function of the referenced test and the two assembled test according to the three levels of ability of examinees and the mean square deviation of test information functions of the tests

Levels of ability of examinees	TIF		
	$RTIF_{\theta_K}$	$ATIF_{\theta_K}$	
		Test 1	Test 2
(- 1.50)	10.50	6.00	6.00
0.00	16.80	16.85	16.80
1.50	10.00	12.80	13.20
MSD of TIF	-	9.36	10.16
Parallelism	-	Not being parallel to the referenced test	Not being parallel to the referenced test

Case 3 - When the automated test assembly programme stipulated the conditions of the test information function of the referenced test according to the four levels of ability of examinees, i.e. (-1.50), 0.00, 1.50, and 2.50, it was found that the programme was able to produce one test which did not show the accuracy of parallelism, as shown in table 3.

Table 3: the test information function of the referenced test and the two assembled test according to the four levels of ability of examinees and the mean square deviation of test information functions of the tests

Levels of ability of examinees	TIF	
	$RTIF_{\theta_K}$	$ATIF_{\theta_K}$
(- 1.50)	10.50	6.00
0.00	16.80	16.85
1.50	10.00	12.20
2.50	5.00	8.00
MSD of TIF	-	.837
Parallelism	-	Not being parallel to the referenced test

The results of the verification of the test assembly programme in terms of test overlap according to the levels of ability of examinees of the referenced test

Table 4: item exposure mean, item exposure variance, and average test overlap rate

Levels of ability of examinees	Test	Item exposure mean	Item exposure variance	$\hat{\bar{T}}$	Fulfillment of test overlap criterion
2	1	0.050	0.001	0.070	Fulfilled
	2	0.250	0.020	0.320	Not Fulfilled
	3	0.280	0.025	0.369	Not Fulfilled
	4	0.330	0.027	0.411	Not Fulfilled
3	1	0.050	0.001	0.070	Fulfilled
	2	0.350	0.030	0.430	Not Fulfilled
4	1	0.050	0.001	0.070	Fulfilled

Case 1 –When determining the test information function in respect of the two levels of ability of examinees, i.e. 0.00 and 1.50, it was found that the test assembly programme could produce one test paper whose average test overlap rate ($\hat{\bar{T}}$) was 0.070, which was lower than the agreed criterion (0.250). This suggested that one test produced from the test assembly programme had the ratio of test overlap lower than the agreed criterion, whereas the average test overlap rate of the other three tests was higher than the agreed criterion with the average test overlap rate of 0.320, 0.369, and 0.411, respectively, as shown in table 4.

Case 2 – When determining the test information function in respect of the three levels of ability of examinees, i.e. (-1.50), 0.00, and 1.50, it was found that the test assembly programme could produce one test whose average test overlap rate ($\hat{\bar{T}}$) was 0.070 which was lower than the agreed criterion (0.250). This suggested that one test produced from the test assembly programme had the ratio of test overlap lower than the agreed criterion, while the average test overlap rate of

the other test was higher than the agreed criterion with the average overlap of 0.430, as shown in table 4.

Case 3 –When determining the test information function in respect of four levels of ability of examinees, i.e. (-1.50), 0.00, 1.50, 2.50, it was found that the test assembly programme could produce one test paper whose average test overlap rate ($\hat{\bar{T}}$) was 0.070, which was lower than the agreed criterion (0.250). This suggested that one test produced from the test assembly programme had the ratio of test overlap lower than the agreed criterion, as shown in table 4.

The results of the verification of the test assembly programme in terms of parallelism and test overlap in respect of the levels of ability of examinees of the referenced test

Table 5: The result of the accuracy of test parallelism and the low average test overlap rate between test papers of the test assembly programme

Levels of ability of examinees	The accuracy of the test assembly programme			Numbers of test papers created by the programme
	Parallelism	Test overlap rate lower than 25%		
		valid	invalid	
2	valid	1	1	**2**
	invalid	-	2	**2**
total		**1**	**3**	**4**
3	valid	-	-	**-**
	invalid	1	1	**2**
total		**1**	**1**	**2**
4	valid	-	-	**-**
	invalid	1	-	**1**
total		**1**	**-**	**1**

Case 1 - When determining the test information function in respect of the two levels of ability of examinees, i.e. 0.00 and 1.50, it was found that the test assembly programme could produce one test that showed parallelism and had a low average test overlap rate ($\hat{\bar{T}}$) was 0.070 which was lower than the agreed criterion 0.250), as shown in table 5.

Case 2 - When determining the test information function in respect of the three levels of ability of examinees, i.e. (-1.50), 0.00, and 1.50, it was found that the test assembly programme could not produce any test that showed parallelism and had a low average test overlap rate ($\hat{\bar{T}}$), as shown in table 5.

Case 3 - When determining the test information function in respect of four levels of ability of examinees, i.e. (-1.50), 0.00, 1.50, 2.50, it was found that the test assembly programme could not produce any test that showed parallelism and had a low average test overlap rate ($\hat{\bar{T}}$), as shown in table 5.

Discussion

This study aimed to create the test assembly programme from the item bank in order to produce tests and to test the accuracy of the programme in terms of test parallelism of the test and test overlap rate. The result of the study indicated that the programme could produce two tests parallel to the referenced test, when considering the accuracy of test parallelism. The tests produced were suitable for predicting the scores of the two levels of ability of examinees: being moderate and fairly high. When considering both test parallelism and test overlap rate, the programme was able to produce only one test; this suggested that the programme could produce only one test with the accuracy of predicting the scores of the ability of examinees corresponding to the accuracy of the referenced test – namely, being moderate and fairly high. Furthermore, the item bank ensured the safety of item leak rate between tests that the rate was lower than 25%. When considering only test overlap rate, the programme could produce three tests with test overlap rate lower than 25%. The first paper could be used with examinees whose abilities were moderate and fairly high; but it could not estimate the scores of examinees corresponding to the referenced test. The second paper could be used with examinees whose abilities were quite low, moderate, and fairly high; but it could not estimate the scores of examinees corresponding to the referenced test. The third paper could be used with examinees whose abilities were quite low, moderate, fairly high, and high; but it could not estimate the scores of examinees corresponding to the reference, as shown in table 5.

From this study, the test assembly programme could produce two tests parallel to the referenced test. The tests produced had the accuracy of predicting the scores of the two levels of ability of examinees. When stipulating conditions of three and four levels of ability, the test assembly programme could produce one test that satisfied each condition of the levels of ability. The test produced, however, lacked of the accuracy of parallelism; the test could not accurately estimate the scores of examinees at three and four levels of ability. This study was supported by the principles established by de Ayala (2009) and McDonald (2009), according to which there is a correlation between a parallel test whose discrimination index is high (over 1.00) and a test information function (TIF) which indicates the accuracy of estimating the scores of ability of examinees in certain levels. Nevertheless, the item bank used in this study had a limitation as to the number of items that had high discrimination index. Consequently, items selected and used in one test could not adequately be used in the subsequent tests. It followed that the test assembly programme capable of producing parallel tests suitable for examinees with moderate and fairly high level of ability could not produce parallel tests suitable for examinees with quite low, moderate, and fairly high level of ability, and also not suitable for those with quite low, moderate, fairly high, and high level of ability. Furthermore, the result of this study also corresponded to that of Lin's (2008). Lin (2008) tested the accuracy of the test assembly programme created by WDM heuristic. It was found that the programme could produce parallel tests. The tests produced showed the accuracy of predicting the scores of the two levels of ability of students. The programme lacked, however, the accuracy of parallelism when more conditions of the levels of ability of examinees were stipulated; namely three and four.

Regarding the accuracy of the test assembly programme in terms of test parallelism of the test and test overlap rate, the programme could produce only one test with the accuracy of estimating the scores of the ability of examinees corresponding to the accuracy of the referenced test – namely, being moderate and fairly high. The safety of item leak rate between tests was lower than 25%. The test assembly programme could not produce any test that met the conditions of ability levels stipulated at the three and four levels of ability. This study corresponded to the result of a study conducted by Lin (2010) in which the accuracy of the automated assembly programme was tested in order to produce parallel tests through weighted deviations model (WDM) heuristic. The study also used a function that controlled test overlap rated in accord with the concept formulated by Chen and Lei (2009).

It was found that the test assembly programme was able to produce ten tests with test parallelism, and the tests produced could accurately estimate the scores of students only in one degree of ability – being moderate. This might result from the objectives of the production of parallel tests through the use of the test assembly programme: the accuracy of estimating the scores of ability of examinees and of controlling the test overlap rate between tests. The automated assembly programme stipulated four conditions on the item bank: first, the number of the items must be at least ten times more than the number of the items used in the actual test; second, the number of items with the difficulty index corresponding to the ability levels of examinees whose scores need to be estimated must be high; third, the number of items with high discrimination index must be high; and fourth, the stipulation of the conditions on the accuracy of estimating the scores of ability which corresponded to every level in the programme would affect the test overlap rate which did not satisfy the agreed criterion.

This study has limitations: namely, first, the test assembly programme could produce parallel tests with very low number and could not keep control of test overlap rate; second, the number of items that had the difficulty index corresponding to the ability levels of various examinees and with high discrimination index (over 1.00) was inadequate; and finally, the item bank used in this study was suitable for producing parallel tests for measuring academic performance only due to the fact that the parallel tests produced showed a high degree of the predictive accuracy of examinees with moderate ability, followed by those with low ability and fairly high ability. Future studies should take the followings into consideration. First, they should put emphasis on increasing the number of items in the item bank, so that the number of the items is adequate for producing parallel test effectively.

In particular, the number of items classified by content must be adequate for producing tests and the number of items with high discrimination index (over 1.00) ought to be increased.

Further research should investigate the influences of the size of an item bank and the range of discrimination index on the effectiveness of the automated test assembly programme so as to produce parallel tests whose test overlap rate are below the agreed criterion. The accuracy of the automated test assembly programme from the item bank should also be tested. Furthermore, so as to be able to perform one of the vital roles in examinations in educational institutes, a test assembly programmed should be constructed to produce parallel tests.

Implication

The test assembly programme constructed in this study can be used for organising examinations in order to make an effective measurement and evaluation, in particular for developing tests that show test parallelism and are capable of effectively controlling test leaks. These characteristics make the interchangeability possible. In addition, it can be beneficial to instructors when using this programme to produce parallel tests for test enhanced learning through repeated tests. Similarly, Butler and Roediger (2007) suggested that the long-term retention in learning among students could be supported by test-enhanced learning which allowed learners to participate in the repeated tests or simulated situations. Therefore, this opportunity is considered as being capitalized upon and testing is a vital process of students' learning process to improve their long-term memory in learning if students are engaging in assessment. It can be said that focusing on the assessment by utilizing parallel tests will ensure that learners are demonstrating the achieved or intended learning outcomes (Biggs & Tang, 2007).

Acknowledgments

This work was supported by the Ramkhamhaeng University.

References

Baker, F.B. (2001). *The Basics of Item Response Theory*. United States of America: ERIC Clearinghouse on Assessment and Evaluation.

Biggs, J.B. & Tang, C. (2007). *Teaching for Quality Learning at University*. (3rd Ed.).

Maidenhead: McGraw Hill Education & Open University Press

Butler, A.C. & Roediger, H.L. (2007). Testing improves long-term retention in a simulated classroom

setting. *European Journal of Cognitive Psychology*, 19, 514-527.

Chang, H-.H., & Zhang, J. (2002). *Hypergeometric family and item overlap rates in computerized adaptive Testing*. Retrieved August 15, 2015, from http//:library.lib.ru.ac.th/

Chen, P-.H., Chang, H-H., & Wu, H. (2012). *Item Selection for the Development of Parallel Forms from an IRT-Based Seed Test Using a Sampling and Classification Approach.* Retrieved April 30, 2016, from http//:epm.sagepub.com. search~S0*thx?/Xjournal +of+educational+measurement.

Chen, S. & Lei, P., 2009, *Investigating the relationship between item exposure and test overlap :Item sharing And item pooling***.** Retrieved September 30, 2016, from http://www.bpsjournals.Co.uk.

Chen, S., Ankenmann, R.D., & Spray, J.A. (2003). *The relationship between item exposure and test overlap in computerized adaptive testing*. Retrieved April 30, 2016, from http://link.springer.com/article/10.1007%2FBF0202294991.

de Ayala, R .J. (2009). *The Theory and Practice of Item Response Theory.* New York: The Guilford Press.

Diol, Q., & van der Linden. (2013). *Automated Test Assembly Using lp_Solve Version 5.5 in R.* Retrieved June 9, 2016, from https//:apm.sagepub.com.

Gorsuch, R.L (2003). Factor Analysis. In J.A. Schinka & W.F.Velicer (Vol.Eds). *Handbook of psychology : Vol. 2. Research methods in psychology. (pp.143-164)*. New York: John Wiley & Sons.

Hambleton, R. K. (1984). *Validating the test scores. In A Guide to Criterion-Referenced Test Construction*. Berk, R.A. (Ed). p. 199-230. Baltimore: The Johns Hopkins University Press.

Hambleton, R. K., & Swaminathan, H. (1996). *ITEM RESPONSE THEORY: Principles and Applications*. Norwel: Kluwer Academic Publishers.

Han, K .T., & Hambleton, R .K. (2007). *User's Manual :WinGen Center for Educational Assessment Report No .642*. Amherst, MA: University of Massachusetts, School of Education.

Lin, C-J. (2008). *Comparisons between Classical Test Theory and Item response Theory in Automated Test Assembly of Parallel Forms.* Journal of Technology, Learning, and Assessment .6 (8), 1–42.

Lin, C-J. (2010). *Controlling Test Overlap Rate in Automated Test Assembly of Multiple Equivalent Test Forms. Journal of Technology*. Retrieved July 9, 2016, from Available at: http://www.jtla.org.

Luecht, R .M. (1998). *Computer-assisted test assembly using optimization heuristics.* Applied Psychological Measurement, 22, 224-236

McAlpine, M. (2002). *Design Requirements of a Databank*. England: Computer Assisted.

McDonald, R. (2009). *Test theory :a unified treatment* .New Jersey: Lawrence Erlbaum Associates.

van der Linden, W.J. (2005). *Linear Models for Optimal Test Design*. New York: Springer Science +Business Media.