Determination the Balance of Contents in the Academic Potential Test

Rigella Ribka Dwiwanih[1], Istiani[1]
[1]*Psychology Department, Faculty of Humanities, Bina Nusantara University, Jakarta, Indonesia 11480*

Keyword: Content Balancing, Academic Potential Test, Computerized Adaptive Tests

Abstract : This study aims to determine Content Balancing on Academic Potential Test based on adaptive. Content balancing is one of the topics from computerized adaptive testing (CAT) that is still very under-researched in Indonesia. This study uses quantitative method with 2041 sample data from prospective students of University X in 2016 who have followed the Academic Aptitude Test (TPA). Analysis of data processing of research sample is to change the sample result from Classical Test Theory (CTT) to Item Response Theory (IRT), and then processed by using Pearson Correlation statistic analysis to find out the most appropriate number of items for each subtest. Based on the results of data processing, obtained the number of items that best match the value of each subtest, that is the number of 24 items in the abstraction subtest, and the number of 21 items for the logical subtest. This study determines blueprint which is the earliest stage in determining content balancing.

1 INTRODUCTION

Psychological tests are often used for individual or group counseling, work selection, work placement, and much more. According to Gregory (Gregory, 2011) understanding psychological testing is a standard procedure for taking behavioral samples and describing them in categories or scores. In carrying out the administration there are two ways, the first is a test in the process of using pencils/pens and paper or referred to as Paper and Pencil Tests (PPT), and the second is a test in the process using a computer or referred to as Computer Based Tests (CBT).

In general, and especially in Indonesia, until now people are still more familiar with paper and pencil tests because of the way these tests are used more often. It's just that, along with the development of the times, computers are increasingly developing so that the tests conducted using CBT are also increasing (Hardcastle, Errmann-Abell, & DeBoer, 2017).

One example of the application of CBT in Indonesia which is a topic of discussion in the community is the National Examination (UN), the way it is given to change from PPT to CBT. In principle, CBT is the same as PPT in terms of assessment, where both use the Classical Test Theory (CTT) approach. CTT approach is used to predict the results of a test by considering the ability parameters of participants and the level of difficulty of the questions (Sumintono, 2015). However, this approach to CTT depends on the results of group tests that have been obtained, thus preventing the generalization of assessments of other groups (Johnson, 2006). In addition, the approach generalizes the administration of questions to all test takers, resulting in suboptimal and inaccurate results.

In its development, there are various computer-based test innovations, one of which is Computerized Adaptive Testing (CAT), which answers the shortcomings of PPT and CBT. CAT is a type of computer-based test that adapts to the ability of participants who take the test. CAT is also increasingly being applied in educational and psychological tests, because it adjusts the test to the test taker's performance thereby reducing the length of the test, increasing the motivation of test-takers to respond according to their cognitive level, and avoiding giving test kits with questions that are too easy or too difficult (Wainer, Dorans, Flaugher, Green, & Mislevy, 2000 in (Veldkamp, 2014)).

The CAT can adjust individual abilities and question items expressed on the same scale because it uses the IRT approach. IRT is a family of mathematical models that explains how people interact with items from a test. IRT also answers the

Dwiwanih, R. and Istiani, .
Determination the Balance of Contents in the Academic Potential Test.
DOI: 10.5220/0010009100002917
In *Proceedings of the 3rd International Conference on Social Sciences, Laws, Arts and Humanities (BINUS-JIC 2018)*, pages 412-417
ISBN: 978-989-758-515-9

weaknesses that exist in CTT, where the CAT using the IRT approach makes test takers receive the items selected optimally to measure their potential and each test taker may not receive the same question items. The principle of IRT is to be involved in the selection of items that are most suitable for test takers and equalize the scores of all the different items (Embretson & Reise, 2000).

In Indonesia, the use of CAT is still very small and only exists in several government agencies. The new CAT was first present in 2014 at the Army Psychology Service (Dispsiad) under the name CAT 5 and is the result of cooperation between the Dispsiad and the German Armed Forces Psychology Service (Psychologische Dienst Der Bundeswehr) (Dispenad, 2014). Then other government agencies also started using the CAT.

In Indonesia, Computerized Adaptive Testing (CAT) is also often mistaken for ordinary computer-based tests. Many agencies or institutions claim to use CAT but are actually Computer-Based Tests (CBT) that do not provide question items that are tailored to individual abilities. The difference in understanding can occur because the knowledge of Computerized Adaptive Testing (CAT) is still little known by the public and many who misinterpret it with Computer Based Test (CBT). Therefore, research is needed on CAT that can help the development of a better CAT in Indonesia.

In supporting the development of CAT in Indonesia, there are many research topics that need to be developed, one of which is content balancing. By the nature of adaptive tests, examinees who take the same test will be given different items, but each must receive the same item distribution based on the area of the field being tested. For example, for a mathematics test consisting of 28 items, it would not be valid if 28 arithmetic items were given to one participant and 28 geometry items to another participant. There must be a balance in all content or domains measured (Leung, Chang, & Hau, 2003). In addition, the content balance is not only limited in the implementation of CAT administration but also in administration in the form of PPT and CBT. This is because the balance of content is needed so that items in the test that have more than one dimension can be given equally to the test takers.

Content balance can be applied to various psychological tests, one of which is the aptitude test, the Academic Potential Test (TPA). In its development, various educational institutions have used TPA as an entrance selection test, one of which is a private university entrance selection test. The TPA has then been developed independently and adapted to the needs of each university. One of the universities that developed their own TPA is College X which adapted it from the Indonesian Collective Intelligence Test (TIKI) and the Scholastic Aptitude Test (SAT).

In administering the test, University X still uses Computer-Based Tests (CBT) using the CTT approach. But going forward, Higher Education X plans to develop an adaptive-based Academic Potential Test (TPA). Based on the aforementioned matters, the researcher is interested in conducting research on the content balance in the Academic Potential Test (TPA) in Higher Education X, in its development to become an adaptive-based TPA test.

2 RESEARCH METHODS

The research variable raised in this study is the balance of content. Content balancing is a set of one or several additional item selection rules based on non-statistical content or feature items (Segall, n.d.).

This research was conducted on a limited sample as a preliminary study of content balance for the development of an adaptive-based Academic Potential Test (TPA). The data used comes from the selection of prospective new students who have taken the Academic Potential Test (TPA) in Higher Education X in 2016, namely 2041 data samples consisting of Subtraction Abstraction, Substance Logic, and Verbal Subtest. The design of this study is a descriptive study that is useful as a preliminary study (that is, in the initial stages of research) (Gravetter & Forzano, 2012).

The data processed were obtained from the results of the Academic Potential Test (TPA) for prospective new students of Higher Education X in 2016 with a total of 2041 samples. TPA has three sub-tests, namely the abstract sub-test consisting of 35 question items, the Logic sub-test consisting of 30 question items, and the Verbal sub-test consisting of 35 question items.

The result you want is the balance of content on the adaptive-based TPA test. It's just that the sample data obtained is still in the form of data based on the Classical Test Theory (CTT) approach, while the required data is the ability estimation and item difficulty level. Therefore, it is necessary to change sample data into Item Response Theory (IRT) with the Rasch Model. The Rasch Model is a model that predicts the possibility of a true answer from the relationship between the ability estimate (theta) and the item difficulty level parameter (Partchev, 2004).

By using the Rasch Model, data cleaning is carried out which causes the elimination of the estimated capability of the sample and items. Data cleaning on capability estimation is done due to inconsistent answers, ie samples with MNSQ criteria above 2. Then, data cleaning on items is done because items do not measure the ability to be measured, ie items with MNSQ above 2, and the presence of maximum or minimum estimated measure.

The content balance is then obtained by reprocessing by reducing the five items carried out twice to obtain three different numbers of items for each sub-test. The way to determine the five items reduced questions is to choose based on the high, medium and low measure or difficulty level of each item. The measure results or estimated abilities obtained have a negative to positive measure or ability estimate with the highest scale of 4 and -4, therefore changed on a scale of 100.

After each sub-test has three different items, then the most effective number for the balance of contents will be searched by correlating between the three items in each sub-test, using the Correlation analysis technique.

Processing of these data is processed using the help of the WINSTEPS software program, which is software for item analysis with the Rasch Model associated with the calculation of the 1PL parameter model (Linacre, 2004), as well as the help of Microsoft Excel, and the Statistical SPSS version 20 software program.

3 RESULTS AND DISCUSSION

There was a discussion about the description of research data for each sub-test after being processed into IRT and the description of the data based on gender, regional origin, and origin of the study sample schools. In addition, there are correlation test results to find the most effective number of items for abstraction, and logic subtest. For verbal sub-tests do not use correlation because there is only one number of items.

Table 1. Description of Abstraction Subtest Data

Input		*Measured*	
Items	***Persons***	***Items***	***Persons***
		29	1861
35	2041	24	1849
		19	1835

Source: Results of WINSTEPS Data Processing

Based on table 1 above, it can be seen that the data of the abstraction subtest entered for processing are 35 question items and consist of 2041 sample data. However, after processing the data, there are sample data and items that must be deleted because they do not meet the criteria, so that the data that can be processed is only 29 question items with 1861 sample data. In addition, there were 24 question items with a total of 1849 sample data that could be processed and 19 question items with a total of 1835 sample data that could be processed. The reduction in the number of samples is because every time there is a reduction of five items, there is an estimate of the ability of the sample that does not meet the criteria.

Table 2. Description of Logic Subtest Data

Input		*Measured*	
Items	***Persons***	***Items***	***Persons***
		26	1724
30	2041	21	1705
		16	1685

Source: Results of WINSTEPS Data Processing

Based on table 2 above, it can be seen that the data of the logic subtest entered for processing are 30 question items and consist of 2041 sample data. However, after processing the data, there are sample data and items that must be deleted because they do not meet the criteria, so that the data that can be processed is only 26 question items with 1724 sample data. In addition, there are 21 question items with a total sample of data that can be processed as many as 1705 people and 16 question items with a total sample of data that can be processed as many as 1685 people. The reduction in the number of samples is because every time there is a reduction of five items, there is an estimate of the ability of the sample that does not meet the criteria.

Table 3. Description of Verbal Subtest Data

Input		*Measured*	
Items	***Persons***	***Items***	***Persons***
35	2041	17	1212

Source: Results of WINSTEPS Data Processing

Based on table 3 above, it can be seen that the verbal subtest data entered for processing are 35 question items and consist of 2041 sample data. However, after processing the data, there are sample data and items that must be deleted because they do

not meet the criteria, so the data that can be processed is only 17 question items with 1212 sample data.

To illustrate the distribution of sample data for all sub-tests, namely abstraction, logic sub-test, and verbal subtest, dominated by male samples, originating from DKI Jakarta, and originating from private schools.

Table 4. Abstraction Subtest Correlation Test Results

Correlations				
		Abstraction29 Items	Abstraction24 Items	Abstraction19 Items
Abstraction29 Items	Pearson Correlation	1	.998**	.995**
	Sig. (2-tailed)		.000	.000
	N	1861	1849	1835
Abstraction24 Items	Pearson Correlation	.998**	1	.996**
	Sig. (2-tailed)	.000		.000
	N	1849	1849	1835
Abstraction19 Items	Pearson Correlation	.995**	.996**	1
	Sig. (2-tailed)	.000	.000	
	N	1835	1835	1835

**. Correlation is significant at the 0.01 level (2-tailed).

Source: SPSS Data Processing Results

Based on table 4 of the correlation test results above, it can be seen that the correlation between the 29 items abstraction and the 24 items abstraction is 0.998 (r = 0.998) with a significance of 0.000 (p = 0.000 <0.01). Then, the correlation between the 29 item abstraction and the 19 item abstraction is 0.995 (r = 0.995) with a significance of 0.000 (p = 0.000 <0.01). Furthermore, the correlation between the abstraction of 24 items with abstraction of 19 items is 0.996 (r = 0.996) with a significance of 0.000 (p = 0.000 <0.01).

From the results of the correlation analysis shows the number of the most effective items to be used in the abstraction subtest is the number of 24 items, which shows the greatest correlation when correlated with 29 items and 19 items.

Table 5. Subtest Logic Correlation Test Results

Correlations				
		Logic26 Items	Logic21 Items	Logic16 Items
Logic26 Items	Pearson Correlation	1	.993**	.988**
	Sig. (2-tailed)		.000	.000
	N	1724	1705	1685
Logic21 Items	Pearson Correlation	.993**	1	.989**
	Sig. (2-tailed)	.000		.000
	N	1705	1705	1685
Logic16 Items	Pearson Correlation	.988**	.989**	1
	Sig. (2-tailed)	.000	.000	
	N	1685	1685	1685

**. Correlation is significant at the 0.01 level (2-tailed).

Source: SPSS Data Processing Results

Based on table 5 of the correlation test results above, it can be seen that the correlation between logic 26 items with logic items 21 is 0.993 (r = 0.993) with a significance of 0.000 (p = 0.000 <0.01). Then, the correlation between logic 26 items with logic 16 items is 0.988 (r = 0.988) with a significance of 0.000 (p = 0.000 <0.01). Furthermore, the correlation between logic 21 items with logic 16 items is 0.989 (r = 0.989) with a significance of 0.000 (p = 0.000 <0.01).

From the results of the correlation analysis shows the number of items that are most effective for use in sub-logic logic is the number of 21 items, which shows the greatest correlation when correlated with 26 items and 16 items.

Table 6. Descriptive Analysis of Verbal Subtest

Statistics	
Verbal17Items	
Mean	32.66
Median	33.88
Std. Deviation	9.47
Minimum	3.63
Maximum	65

Source: SPSS Data Processing Results

On table 6 shows the descriptive statistics for verbal sub-tests, showing that the mean value of 1212 sample data is 32.66, median 33.88, with standard deviation (std. Deviation) 9.47, and the smallest (minimum) value the sample data is 3.63, and the largest value (maximum) sample data is 65.

4 CONCLUSIONS AND SUGGESTIONS

Based on the results of data processing and calculations that have been done, obtained the number of items that are most suitable for the balance of contents in each sub-test. In the abstraction sub-test, the most effective number of items for content balance is 24 question items, then in the logic sub-test, the most effective number of items for content balance is 21 question items, and in the verbal sub-test the most effective items for balance content are not obtained because limited number of items due to a large number of items that must be eliminated.

Based on the conclusions that have been obtained, the number of 24 items in the abstraction sub-test, and the number of 21 questions in the logic sub-test is the most effective number of items to balance the contents of the Academic Potential Test (TPA), because it is a combination of items that are closest to theta (estimated ability). Only the verbal sub-test did not obtain the appropriate number of items for the balance of contents because there was only one number of items, namely 17 items obtained directly from the results of processing without reducing items, so no correlation analysis was performed. This is because the sample data on the verbal sub-test is the most does not meet the criteria in the estimation of the ability of the sample because of the many inconsistent answers, as well as the criteria in the level of difficulty of items and items that have not really measured verbal ability.

In general, all question items in each sub-test already have a level of difficulty that is evenly distributed when processed using the IRT approach. It's just that, the correlation analysis results obtained indicate the correlation (r) which is not too large the difference. In the abstraction sub-test, the correlation of 29 items with the abstraction of 24 items is 0.998 (r = 0.998), the correlation between the abstraction of 29 items and the abstraction of 19 items is 0.995 (r = 0.995), and the magnitude of the correlation between the abstraction of 24 items and the abstraction of 19 items is 0.996 (r = 0.996). Then, the logic sub-test is known that the correlation between logic 26 items with logic 21 items is 0.993 (r = 0.993), the correlation between logic 26 items and logic 16 items is 0.988 (r = 0.988), and the correlation between logic 21 items with logic 16 items is 0.989 (r = 0.989).

In the abstraction and logic sub-test, it can be seen that the large correlation does not have a large difference, but in the abstraction sub-test the number is 29 items with the number 19 items and the number of logic items 26 with the number 16 items has the smallest correlation compared with the large correlation between the number another item. That is because the difference in the number of items reaches 10 the number of items, different from the number of other items that only have a difference of 5 in the number of items.

Based on these results, it can be assumed that there is a need for larger item differences to see large differences in more significant correlations, to be able to determine a more effective combination of content balance. It's just that it can't be done because of the limited number of items, so the conditions for the number of items cannot be fulfilled. In addition, the distribution of data, namely gender, regional origin, and school origin can also affect the results of ability estimates, where the sample data used is dominated

by men, originating from DKI Jakarta, and coming from private schools.

This research is a preliminary study of content balance and is a descriptive study for the development of TPA into an adaptive-based test. In Indonesia, research on the balance of content is not yet available because Computerized Adaptive Testing (CAT) is also the first time entering Indonesia in 2014 (Dispenad, 2014). Therefore, this research is the first and the earliest step in determining the balance of content, namely designing a "specification table" or blueprint that outlines the breakdown of specific types of items and content needed in the test. The items given to test takers will later be selected based on the items that best represent what is actually needed based on the specification table or blueprint (Johnson, 2006).

This research only reached the stage of determining the specification table or blueprint test items, but to determine the most effective content balance method, and later used in the development of adaptive-based landfill can use one of the three content balance methods, namely The Constrained CAT (CCAT), The Modified Multinomial Model (MMM), or The Modified Constrained CAT (MCCAT).

Based on the research of Leung, Chang, and Hau (Leung et al., 2003) who compared the three methods of the content balance of CCAT, MMM, and MCCAT, found that the most effective content balance method among the three was The Modified Multinomial Model (MMM). This is because, among the three methods, the MMM method is the most effective in reducing the predictable item content sequence, and the number of items that are overexposed without regard to the item selection approach, test length, or target maximum exposure level. The method is the result of research with various forms of research that are different from this study. Therefore, to find out the most appropriate content balance method for the Academic Potential Test (TPA) for Higher Education X is to conduct further research by comparing the results of the three methods that exist when used on TPA.

Researchers realize this research still has many shortcomings that can be corrected to be more optimal. Therefore, researchers have some suggestions that can be done in subsequent studies, namely conducting research on content balance using other psychological tests, or can proceed by using a comparison between content balance methods based on the results of this study. Then so that the results obtained are better, it is expected to have a higher number of items and a greater difference in the number of items. In addition, in the development of Computerized Adaptive Testing (CAT) in Indonesia, it is possible to use data derived from CAT and conduct research on other CAT topics. For Higher Education X, in order to be able to implement an adaptive-based Academic Potential Test (TPA), Higher Education X must provide a bank item consisting of at least 300 items with an even distribution of difficulty levels, so that the balance of content can be achieved.

REFERENCES

Dispenad. (2014). CAT 5 HADIR DI DISPSIAD. Retrieved January 2, 2018, from DISPSIAD website: http://www.dispsiad.mil.id/index.php/en/publikasi/berita/235-cat-5-hadir-di dispsiad

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. London: Lawrence Erlbaum Associates.

Gravetter, F. J., & Forzano, L.-A. B. (2012). *Research Methods for the Behavioral Sciences*. Canada: Wadsworth, Cengage Learning.

Gregory, R. J. (2011). *Psychological Testing: History, Principles, and Applications*. Boston: Pearson.

Hardcastle, J., Errmann-Abell, C. F., & DeBoer, G. E. (2017). Comparing Student Performance on Paper-and-Pencil and Computer-Based-Tests. *Paper Presented at the Annual Meeting of the American Educational Research Association*.

Johnson, M. A. (2006). *An Investigation of Stratification Exposure Control Procedures in CATs Using the Generalized Partial Credit Model*. Austin: University of Texas.

Leung, C. K., Chang, H. H., & Hau, K. T. (2003). Incorporation of Content Balancing Requirements in Stratification Designs for Computerized Adaptive Testing. *Educational and Psychological Measurement*, *63*(2), 257–270.

Linacre, J. M. (2004). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, *5*(1), 95–110.

Partchev, I. (2004). *A Visual Guide to Item Response Theory*. Friedrich-Schiller-Universitat¨ Jena.

Segall, O. D. (n.d.). *Computerized Adaptive Testing. Encyclopedia of Social Measurement*. United States.

Sumintono, B. (2015). *Pemodelan Rasch pada Asesmen Pendidikan: suatu pengantar*. Medan: Universitas Sumatera Utara.

Veldkamp, B. P. (2014). *Some Practical Issues in Computer Adaptive Test With Response Times*. Law Admission Council.