Development of parallel tests for numerical ability for student teacher in public university

[1]Thattharit Thiamtham

ABSTRACT--*The objective of this research is to develop three parallel tests and establish the norms for interpreting the scores from numerical ability test of teachers college students in public universities. The sample group in this study consisted of 1,993 fourth-year teachers college students in public universities the academic year 2018. They were selected purposive sampling and multi-stage sampling. The results of the research show that the 3 parallel tests, each one consisting of 20 items, are of high quality in terms of content validity. Each item has IOC of 1.00. The difficulty level of the first, the second, and the third test were 0.68, 0.67, and 0.68, respectively. These three tests were parallel to one another. The score from each test was related to that of the other tests at statistical significance of 0.01. Correlation coefficient between Test 1 and Test 2, Test 1 and Test 3, and Test 2 and Test 3 were 0.96, 0.96 and 0.94 respectively. The means and the variances from the three tests did not differ. The norm was shown in a table presenting the relation between raw scores, percentile, stanine, and* ***normal*** *scores; the latter were ranging from 25-68.*

Keyword--*Parallel Score; Numerical Ability; Teachers; College students*

I. INTRODUCTION

Education is one of the basic rights that the Thai government needs to provide to all Thai people in different ages, in order that they may thrive in good ways and have their own intellect for developing skills, competency, and abilities to do honest work and live together with others peacefully. This will lead to stability and security of the nation, which is being pressed to prosper and compete with other countries in the global stage, amidst the rapid changes of the 21st Century (Office of the Education Council, 2017). Therefore, teachers are very important persons because they are the frontiers and the major mechanic in developing the quality of education management in Thailand during the time of social changes. That is why it is a necessity to develop teachers to become more modern and prepare them for a higher profession. Sumeth Yaemnoon (2011) said that the learning outcomes of modern teachers must consist of (1) ethics and morality: understanding others, understanding the world, having public mind, sacrificing one's own benefits, and being a good example, (2) cognitive skills: the ability to analyze deeply and assess information and concepts from various types of data, and being initiative, (3) Knowledge: having general knowledge, knowledge specific for becoming teachers in the 21st Century, and profound systematical knowledge about the subject that one teaches, (4) interpersonal relationship and responsibility: having emotional and social maturity, knowing how to be a good leader and a good follower, having good relationship with students, (5) Numerical analysis ability, communication and information technology skill, the ability to analyze statistical

[1] *College of Teacher Education, Phranakhon Rajabhat University, 9 Changwattana Road, Bangkhen Bangkok, Thailand 10220,Tel.: (+662)-544-8291, Fax.: (+662)-544-8610,thattharit1603@gmail.com*

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

figures, spoken language and written language, having good judgement in processing and interpreting the data, and (6) learning management skill: having expertise in student-center learning management in various forms, having knowledge and ability in foreign language to be used in teaching, having the ability to provide learning that is suitable for the 21st Century.

We can see that numerical ability is an important skill for teachers. This goes in line with the fact that the Higher Education Commission (2009) established Qualifications Framework for Higher Education B.E. 2552 that has, at minimum, 5 expected learning outcomes that the graduate must possess. Numerical analysis and communication and information technology are ones of the expected outcomes that the students must develop during their years of study. In selective examination for teachers arranged by Office of the Teacher Civil Service and Education Personnel Commission (2018), it is required that general aptitude test must include numerical ability, logical ability, and language ability. Thus, numerical ability is also part of general aptitude test. In Thailand, Phannachawee Prayunphrom (2008), has developed an aptitude test for Thai professional teachers that can be used in selective examination. Every year, all institutes involved with the arrangement of selective examination for teacher students and open exam for teacher assistants produce the test for selecting teachers and teacher assistants, but general aptitude test was not used to find out about general abilities of the candidates. Therefore, this test can lead to further development of general aptitude of the students, by the teachers and by the student themselves. This is why the researcher has become interested in developing a numerical test that measures someone's ability in calculating and making logical summaries of figures and data. The content was improved by designing the test based on interviews and opinions from the experts. This research was to create 3 parallel tests. They are objective tests with four choices to be used in testing numerical ability during the beginning, the middle, and the ending of the semester when they have been fully taught by the teachers. Each test will provide the faculty, the department, the teacher and the students with the information from the examination and use them to improve and develop the students' numerical ability. This will benefit them when they take the selective examination after they graduate. Therefore, this study of numerical test development is one of the inventions that can drive the mechanics of quality development in teacher students. It will bring great benefits to the development of national education in terms of improving capability of learners to go in line with the current principles and conditions.

II. OBJECTIVES

- To develop parallel tests for measuring the numerical ability of students in teachers college of public universities
- To establish norm for interpreting the scores from the numerical ability tests for students in teachers college of public universities
-

III. SCOPE OF STUDY

3.1 Population and Sample Group

Population

The population used in this study were 20,710 fourth-year teacher students who studied in the academic year 2018 in 54 public universities in Thailand.

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

Sample Groups

The sample groups used in this study consisted of 1,953 fourth-year teacher students who studied in the academic year 2018 in 54 public universities in Thailand. They were divided into groups. One group was for assessing the appropriateness of language used in the tests and for measuring the time spent in completing each test. These 10 students were from College of Teacher Education in Phranakhon Rajabhat University. They were selected by the method of purposive selection. Another group was for checking the quality of each item in the tests and for establishing the norms of the test. They were 200, 200, and 1,543 students, respectively. They were selected by using multi-stage random sampling.

3.2 Variables

The variables used in this study were the quality of the tests in terms of content validity, difficulty, discrimination, reliability, and parallelism of the tests. It was also used in establishing the norms for interpreting the scores.

3.3. Content

The content of these numerical ability tests was constructed by using the content framework of general aptitude test in numerical ability established by Office of the Teacher Civil Service and Education Personnel Commission. There are 3 aspects to be tested: numerals, logics, and language. According to the interviews and the opinions of 5 experts, the content that can be used in creating general aptitude test in numerical ability were in 9 topics: percentage, proportion/ratio, algebra, rule of three in arithmetic, equation, geometry, distance, mean, and logics related to numerals and data.

3.4 Steps

This research aimed to develop 3 parallel tests that assess the general aptitude of numerical ability and establish the norms for interpreting the scores from the tests. The following steps were used in this study:

Step 1: Constructing a general aptitude test in numerical ability

1. Establish the content framework used in constructing the test, based on the content framework stated by Office of the Teacher Civil Service and Educational Personnel Commission (OTEPC) under the Ministry of Education, as well as the interviews of and the investigation by 5 experts, which consist of 3 experts in mathematics and 2 experts in Educational Evaluation. Then the parts of the content that was highly appropriate were selected to produce the test.

2. Make the test blueprint and item specification from the content selected in 1, then check for the appropriateness of the specification by using the same 5 experts. Adjust the tests based on their suggestions to make it even more suitable.

3. Create a general aptitude test in numerical ability, an objective test with 4 choices based on the content and specification selected in no.2. It consisted of 3 parallel tests, each one with 20 items. There were 60 items in total. However, this time the test constructed will have the content in double of that amount, or 120 items.

Step 2: Checking the quality of the items in the tests

1. Content Validity of all the 3 tests (40 items each) was checked by the same 5 experts. The IOC was analyzed according Rovinelli and Hambleton (1997), quoted in Luan Saiyod and Ankhana Saiyod (2000). The test with the IOC value from 0.50 or higher was selected to be adjusted and improved according to the suggestions of the experts, before a real sample tests were constructed.

2. The three tests were used with 10 fourth-year students of College of Teacher Education of Phranakhon Rajabhat University, which had been selected by purposive selection. This was in order to check the appropriateness of language and check the amount of time that students spent on each test, then used the results as the basic data for specifying the right amount of time for each test in the next examination.

Checking the quality of the tests in terms of difficulty and discrimination by making the first try-out with the 200 fourth-year students in College of Teacher Education of public universities that had been selected by multi-stage random sampling. Then the results were used in analyzing the difficulty and the discrimination of the tests. The items that had the difficulty level from 0.20-0.80 and discrimination level from 0.20 or higher would be selected (Natthaporn Laothong. 2016; Luan Saiyod and Ankana Saiyod 2000; Sirichai Kanchanawasi, 2013).

Step 3 – Produce the parallel tests

In this step, the real numerical ability tests were produced by selecting 60 items according to the structure of the test. The items selected must have the same specifications, and similar difficulty level and discrimination level. The items were separated into 3 parallel tests; each one contains 2o items by random.

Step 4 – Quality check in reliability and parallelism of the 3 tests

In this step, the 3 tests produced in step 3 were used in the second try-out with 200 fourth-year students in College of Teacher Education of public universities in order to check the reliability by using KR-20 formula of Kuder-Richardson Method (Luan Saiyod and Ankana Saiyod, 2000). The tests that pass must have the value of reliability at 0.70 or higher. Then the basic statistics of the scores from each test were calculated, which are the mean and standard deviation. Then the parallelism of the 3 tests were measured by finding out the correlation coefficient (Chusri Wongrattana, 2017). The differences in the variances of the 3 tests were measured by using the F-test.

Step 5 – Establish the norms

To establish the norms, one test is randomly selected from the three tests as a source for data collection that will be used in establishing the norms. The test was used with the sample group, which were 1,543 people selected by multi-stage random sampling. The results were analyzed to find the percentile, t-score, and stannine score. Then they were used to establish the norms and produce the numerical ability test for students of College of Teacher Education in public universities.

IV. CONCLUSION AND DISCUSSION OF RESULTS

4.1 Conclusion

Part 1 – the results of the development of parallel tests for numericity ability testing of students in College of Teacher Education in public universities.

Results of creating numerical aptitude test

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

The results of synthesizing the content, the interview and the opinions from 5 experts about the appropriateness of the content used to construct the test. It was found that 9 topics should be included in the tests: percentage, proportion/ratio, algebra, the Rule of Three in Arithmetic, equation, geometry, distance, mean, and logics related to numerals and data. The results show that the appropriateness of the content used in these tests were at the highest ($\bar{x} = 4.73$). Each topic has appropriateness of the highest level, except for geometry and distance, which was at high level. The overall appropriateness of the tests was at the highest level ($\bar{x} = 4.87$). Each item was also at the highest level, except for the complicated problem that requires the knowledge in geometry, which was at high level.

Results of Test Quality Check

1.1 The results of content validity of 120 items show that IOC of each item is equal to 1.00.

1.2 The results of difficulty and discrimination of all items are: difficulty of the overall test was from 0.37 to 0.88, discrimination of the overall test was from 0.40 to 0.84. When the criteria for selecting the items for the test at difficulty of 0.20-0.80 and discrimination of 0.20 or higher, it was found that there were 104 items that met the criteria. They had difficulty level of 0.37-0.08 and discrimination level of 0.40-0.84. There were 16 items that did not meet the criteria. Their difficulty level was between 0.82-0.88 and discrimination level was between 0.50-0.73.

Results of selecting items for the parallel tests

After selecting 3 items from each test that have similar specifications, with difficulty and discrimination very close to the specification, they were put together randomly into 3 parallel tests, each one with 20 items. After the tests were put together, it was found that Test 1, Test 2, and Test 3 have difficulty level of 0.52 – 0.80, 0.50 – 0.80 and 0.53 – 0.80 respectively. The averages of difficulty were 0.68, 0.67 and 0.68, respectively. The discrimination of the three tests were 0.50 – 0.80 , 0.44 – 0.78 and 0.42 – 0.84, respectively. The averages of the discrimination were 0.65, 0.67 and 0.66, respectively. The details are as followed:

Table 1 Difficulty and discrimination in the 3 parallel tests

No.	Quality by item in Test 1		No.	Quality by item in Test 2		No.	Quality by item in Test 3	
	p	r		p	r		p	r
1	0.74	0.62	1	0.79	0.66	1	0.79	0.68
2	0.79	0.50	2	0.78	0.44	2	0.77	0.64
3	0.55	0.64	3	0.51	0.74	3	0.53	0.84
4	0.54	0.80	4	0.57	0.62	4	0.58	0.76
5	0.77	0.68	5	0.75	0.68	5	0.79	0.72
6	0.62	0.64	6	0.65	0.74	6	0.63	0.72
7	0.78	0.56	7	0.80	0.58	7	0.79	0.60
8	0.73	0.70	8	0.71	0.72	8	0.70	0.66
9	0.73	0.74	9	0.74	0.78	9	0.73	0.74
10	0.71	0.66	10	0.65	0.72	10	0.67	0.58

No.	Quality by item in Test 1		No.	Quality by item in Test 2		No.	Quality by item in Test 3	
	p	r		p	r		p	r
11	0.54	0.76	11	0.64	0.50	11	0.57	0.68
12	0.52	0.74	12	0.50	0.70	12	0.59	0.74
13	0.71	0.58	13	0.76	0.54	13	0.79	0.66
14	0.55	0.72	14	0.56	0.46	14	0.63	0.64
15	0.80	0.64	15	0.77	0.44	15	0.80	0.54
16	0.69	0.54	16	0.65	0.66	16	0.61	0.60
17	0.64	0.70	17	0.63	0.68	17	0.60	0.70
18	0.68	0.64	18	0.61	0.74	18	0.65	0.70
19	0.67	0.56	19	0.54	0.56	19	0.57	0.42
20	0.79	0.58	20	0.74	0.72	20	0.71	0.64
Average	0.68	0.65	Average	0.67	0.63	Average	0.68	0.66

Results of reliability check and parallelism of the three tests After three parallel tests (20 items each) have been put together, they were used with the sample group of 200 students. Then the basic statistics of the test, the mean and the standard deviation, were calculated. After calculating the quality of reliability (r_{tt}) of the three tests, it was found that reliability of Tests1, 2 and 3 were 0.90, 0.89 and 0.90, respectively. The details are shown in the table below.

Table 2: Basic statistics and reliability of the three parallel tests

Test	K	$\bar{x}$	S.D.	r_{tt}
Test 1	20	13.57	5.50	0.90
Test 2	20	13.23	5.24	0.89
Tests 3	20	13.48	5.58	0.90

Results of parallelism between the three tests

1) Results from the analysis of correlation coefficient of the three parallel tests

The results of the three parallel tests in correlation coefficient show that these tests are related to one another at statistical significance of .01. The correlation coefficients of Test 1 and Test 2, Test 1 and Test 3, and Test 2 and Test 3 equal to 0.96, 0.96 และ 0.94, respectively. The details are shown in the table below.

Table 3 : Correlation coefficients of the three parallel tests

Test	Test 1	Test 2	Test 3
Test 1	1.00		
Test 2	0.96**	1.00	

Test 3	0.96**	0.94**	1.00

**p < .01

2) The results on differences of the average scores from the 3 parallel tests

The results on the differences of the average scores from the three tests show that Test 1, Test 2, and Test 3 do have different average scores, as shown in details in the table below.

Table 4 : The test to find the differences of the average scores of the parallel tests

Variance Source	SS	df	MS	F
Between groups	12.81	2	6.41	0.215
Within groups	17776.30	597	29.78	
Total	17789.10	599		

The results of the difference of the variances of the scores from the 3 parallel tests show that Test 1, Test 2, and Test 3 did not differ in variances. The details are presented in the table below.

Table 5 : Difference test of variances of scores from the parallel tests

Test	$\bar{x}$	**S.D.**	S^2	F
Test 1	13.57	5.50	30.25	
Test 2	13.23	5.24	27.46	1.13
Test 3	13.48	5.58	31.14	

Part 2: **The norms for interpreting the scores from the numerical ability test of students in Teacher** College in public universities.

The norms for interpreting the scores from the numerical ability test can be shown as raw scores, Stannine, normal scores, as in the table below

Table 6: Norms for the numerical ability test of students from Teacher College in public universities

Raw scores	**Stannine**	**T-Test**	**Numerical Ability**
16 - 20	7 - 9	59 or higher	High
9 - 15	4 - 6	43 – 58	Moderate
1 - 8	1 - 3	43 or lower	Lowe

V. DISCUSSION

This research aimed to develop 3 parallel tests and establish the norms for interpreting the scores from the numerical ability test of students from Teacher College in public universities. The results can be discussed as followed.

1. Content validity of the tests, which had been checked by the experts, using the method of Rovinelli and Hambleton (1977) as quoted on Luan Saiyod and Ankhana Saiyod (2000), shows that in every item, index of item objective congruence between items, specifications of items, and specifications of content equal to 1.00. Therefore,

all the items have content validity, which is in accordance with the statement in the work of Luan Saiyod, Ankana Saiyod, and Phuangrat Thaweerat (2000) that said if the index of item objective congruence equals to or higher than 0.50, it means that the test really measures what it had been designed for.

2. Difficulty level and discrimination level of the 3 tests, based on the first try-out, was found to be between 0.37-0.88. There were 16 items that did not meet the criteria, with discrimination between 0.40-0.84. Therefore, it all the items used in the test have the satisfactory level of discrimination. Then the researcher selected 60 items out of 104 items that met the criteria. There were 3 items for each specification, with similar difficulty and discrimination level. They were randomly put them together into parallel tests. The result was a test that have difficulty level between 0.50-0.80, which means a moderate level to easy level. This is accordance with Natthaporn Laothong (2016), Luan and Ankana Saiyod (2000), Sirichai Kanjanawasi (2013) which stated that a good test must have difficulty level between 0.20 and 0.80, and discrimination level at 0.20 or higher. It also agrees with the statement from Luan and Ankhana Saiyod (2000) which said that a test with difficult level between 0.40-0.59 is considered an average test, while a test that has difficulty level between 0.60-0.60 is considered an easy test. Discrimination level of more than 0.40 is considered very good. It was found that the difficulty level of Test 1, Test 2, and Test 3 were 0.52 – 0.80 , 0.50 – 0.80 and 0.53 – 0.80, respectively. The average difficulty of each test was 0.68 , 0.67 and 0.68, respectively. The discrimination of the tests in between0.50 – 0.80, 0.44 – 0.78 and 0.42 – 0.84, respectively. The average discrimination were 0.65 , 0.67 and 0.66, respectively. Therefore, the parallelism of all these tests have similar level of difficulty and discrimination. This is in accordance with Louis and Marylin (1978) who stated that parallel tests measure the same thing with different questions. They have the same level of difficulty and quality of the questions.

3. Reliability of the 3 parallel tests were calculated by using KR-20 formula of Kuder-Richardson. It was found that the reliability of Test 1, Test 2, and Test 3 equal to 0.90, 0.89 and 0.90, respectively. Therefore, all of the 3 tests have high level of reliability and they are deemed as acceptable and reliable tests. This conforms with Luan and Ankhana Saiyod (2000); Frankel & Wallen, 1999 which said that the reliability level should not be lower than 0.70 in order to be acceptable. It also agrees Suwimon & Wallen (1999) who said that a reliability level that is close to 1.00 shows that the test has good quality.

4. The correlation coefficients of the three parallel tests that they were related at statistical significance of .01. Correlation coefficients between Test 1 and Test 2, Test 1 and Test 3, and Test 2 and Test 3 were 0.96, 0.96 and 0.94, respectively. It means that they have high internal correlation coefficients. The test of differences between the average scores and the differences of the variances of the scores from these tests show that the average scores of all the three tests were not different. They also have the same variances. Therefore, the 3 tests can be considered parallel tests because they had been constructed with the same specifications, randomly selected from the same source, with the same number of items. There was consistency between the items and the specifications. The content, difficulty, and discrimination are at similar level as well. This agrees with Yachai Phonboribun (1979), who stated that parallel tests must produce similar average scores, and Phuangrat Thaweerat (2000) who stated that parallel tests must have similar variances of the scores.

5. The establishment of the norms for numeral ability test of students from Teacher College in public universities was conducted by transforming the raw scores that had been gathered in the percentile form, normalized t-score, and stannine scores. It was norm that comes from comparing the scores with norm-group. In

other word, the score of each test taker was interpreted by comparing to the norms. The norms used with these tests were considered appropriate because it was produced based on the 3 criteria, which was stated by Luan and ANkhana Saiyod (2000). First, it is a good representative. The sample group that was used in producing the norms came from the target group: 1,543 fourth-year teacher students. They were selected by multi-stage sampling method from the population in the target group. Therefore, the norms of these tests have representative quality. As for the size of the sample group, it should be large enough to make all the statistics constant. It means that if more people were added into the groups, the statistics after re-calculating should be similar to the results before adding. In this case, the size of the sample group was large enough for the statistics to be constant. The distribution of data is normal. Second, it has validity. The raw scores gained from testing is compared to the norms and interpreted correctly. Third, it is up-to-date because the norms have just been established.

REFERENCES

1. Office of the Higher Education Commission, (2009). Thai Quality Framework for Higher Education B.E. 2552 (online). http://www.qa.kmutnb.ac.th/upload_files/pakadout/standards/Che_TQF_52.pdf [May 24, 2018]
2. Office of the Teacher Civil Service and Educational Personnel Commission (OTEPC). (2018). The Criteria for Teacher Assistants Selective Examination (online). https://otepc.go.th/th/content_page/item/2134-21-05-61.html [May 25, 2018]
3. Chusi Wongratana (2017). The Technique of Using Statistics in Research (13th edtion). Bangkok : CU Printing House.
4. Natthaporn Laothong (2016). Building Educational Research Instruments. Bangkok: CU Printing House
5. Phanchawee Prayunphrom (2008). Development of Aptitude in Thai Teaching Profession: Master of Education Thesis, doctor’s degree, Faculty of Education, Chulanlongkorn University
6. Phuangrat Thaweerat (2000). Methods of Behavioral Sciences and Social Sciences (7th Edition). Bangkok: Educational and Psychological Test Bureau, Srinakharinwirot University.
7. Yachai Phongboribun (1979). Principles of Educational Evaluation and Assessment. Khon Kaen: Faculty of Education, Khon Kaen Unversity.
8. Office of the Educational Council. (2017). The National Scheme of Education B.E. 2560-2579. Bangkok: Phrik Wan Graffic.
9. Luan Saiyod and Ankhana Saiyod (2000). Measurement of Learning Outcomes (2nd Edition). Bangkok: The Children Club.
10. Sirichai Kanchanawasi (2013). The Theory of Traditional Testing (7th Edition). Bangkok: Faculty of Education, Chulalongkorn University.
11. Sumeth Yaemnun (2011, February). Framework for Producing Modern Teachers. Anusarn Udomsuksa, 37(392).
12. Suwimon Tirakanan. (2010). Methodology of Social Science Research: Toward Practicality. Bangkok: CU Printing House.
13. Frankel, J. R. & Wallen, N. E. (1999). How to Design and Evaluate Research in Education. (4th.Ed). Boston, MA : McGraw Hill.

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

14. Kermel, Louis J. and Marylin O. Kermel. (1978). Measurement and Evaluation in the School. New York : McMillman Publishing.