Establishing Quality Instrument for the Summative Assessment of Pre-Service Elementary Teachers

Apolinaria Daquioag- Andres[1]

_____

[1]Ph. D, Professor I, College of Teacher Education Cagayan State University, Andrews Campus, Tuguegarao City 3500, Philippines

Establishing Quality Instrument for the Summative Assessment of Pre-Service Elementary Teachers

Apolinaria Daquioag- Andres[1]
[1]Ph. D, Professor I, College of Teacher Education Cagayan State University, Andrews Campus, Tuguegarao City 3500, Philippines

ABSTRACT
The study aims to establish the quality of assessment instrument that is administered in the final term examination among pre- service elementary teachers. The test is a 50- item multiple choices type that covered the topics discussed during the term. It was administered among 106 second year students who were enrolled in the course. Quality of the test was determined through face validity, content validity, reliability, index of difficulty, index of discrimination, and item distracters of the test. Face validation of the test was done using the criteria and guidelines while content validation was determined through the Table of Specifications (TOS) of the test. Internal consistency of the test was established using Kuder- Richardson Formula 21 (KRF21). The index of difficulty, index of discrimination, and measure of attractiveness of item distracters of the test were determined using the item analysis procedures and applying the corresponding formulas needed. The results revealed that the test has its face validity, acceptable content validity, and high internal consistency. Item analysis also disclosed that most of the items have moderate index of difficulty, are discriminating and have effective distracters.

Keywords: Assessment, Face validation, Content Validation, Reliability, Item analysis

1. INTRODUCTION

Student assessment is an integral part of the teaching- learning process. As such, it is not just limited to determining what the students have learned as a result of instruction but it is also being integrated with instruction to become an opportunity for learning. As explained by Balagtas and Ferido (2007), it is the process of gathering and organizing qualitative and quantitative data as basis in evaluating students' learning and in making decisions on how to improve teaching or a curriculum.

Valid assessment decisions could substantially improve student performance, guide the teachers in enhancing the teaching- learning process and assist policy makers in improving educational system. At the same time, however, poor assessment procedures could adversely affect the students, teachers and administrators. In fact, the results of the teachers' assessment to students could either make or break their students' future. (Balagtas, et. al, 2010).

Teachers across all levels of education are expected to be competent in developing quality assessment tools to be able to evaluate learners validly and reliably. The Philippine Professional Standards for Teachers, which is built on National Competency-Based Teacher Standards (NCBTS), complements the reform initiatives on teacher quality from pre-service education to in-service training. It articulates what constitutes teacher quality in the K to 12 Reform through well-defined domains, strands, and indicators that provide measures of professional learning, competent practice, and effective engagement. Based on PPST, there are 7 domains that are required by teachers to be effective in the 21st Century in the Philippines. One of these domains requires that teachers apply a variety of assessment tools and strategies in monitoring, evaluating, documenting and reporting learners' needs, progress and achievement. Also, they use assessment data in a variety of ways to inform and enhance the teaching and learning process and programs and provide learners with the necessary feedback about learning outcomes that informs the reporting cycle and enables teachers to select, organize and use sound assessment processes (PPST, 2017).

The emphasis on outcome- based education (OBE) requires students to demonstrate their knowledge and skills. With this, the assessments used to evaluate their performances become critically important. Indeed, the challenge of outcomes- based education is not only on how to teach but equally important is on how to assess learning (Akhmadeeva, Hindy, & Sparrey, 2013). It should be noted that despite the promotion of OBE, traditional pen- and paper assessment is still indispensable as it supports higher order thinking skills. Also, it cannot be denied that board examinations are still in the form of traditional written tests (K12 Academics, 2019).

Few teachers have been given the opportunity to develop their assessment skills and capabilities on the job. Cognizant of the needs of teachers in the basic education level as well as the competencies for all teachers across levels and disciplines that are deemed necessary in assessment of learning, the Commission on Higher Education (CHED) issued a guideline on what ought to be taught in the teacher education curriculum. As stipulated in CHED memoranda series of 2017, one of the emphases of the teacher education curriculum is assessment. In the new teacher education curriculum, CHED requires two courses on assessment. The first course (Assessment in Student Learning 1) focuses on pen-and- paper assessment while the second (Assessment in Student Learning 2) focuses on alternative methods of assessment. The focus of these assessment courses addresses the need on assessment as learning for would-be-teachers. (Balagtas&Ferido, 2007).

Considering the indispensable role of pen- and paper test in the summative assessment of learners, the goal of this study is to develop and establish a teacher- made summative assessment instrument that is appropriate for instructional decisions since assessing the performance of every learner is a very critical task for classroom teacher. Test constructors believe that every assessment tool should possess good qualities(Kubiszyn and Borich, 2007). Most literatures consider that the most common technical concepts in assessment, whether traditional or authentic, are the validity and reliability. Validity, according to Linn and Gronlund (2000), means the degree to which a test measures what it intends to measure. Reliability, on the other hand, means the extent to which a test is consistent and dependable. Hence, this study aimed to examine the face validity, determine the content validity, establish the internal consistency and conduct an item analysis of the test administered by determining the Index of difficulty, Index of discrimination and measure of attractiveness of the distractors of the items.

2. RESEARCH METHOD
2.1 Research Design
The quantitative- descriptive research design was used in the study since the main objective was to establish the quality of the developed summative assessment tool through validity, reliability, and item analysis. Face and content validation were done prior to the administration of the test. Meanwhile, reliability was determined, and item analysis was conducted using the test result of the students.

2.2 Respondents of the Study
The participants of the study were the 106 second year pre- service elementary teachers who were enrolled in one of their content area courses.

2.3 Research Procedure
Prior to the administration of the 50- item multiple choice assessment instrument, validation was conducted to establish its quality. Face validation of the 50- item summative assessment was done using a checklist that contains the set of criteria and guidelines for test construction. On the other hand, content validation was conducted using the Table of Specifications (TOS) as a guide. The TOS served as the basis to check whether the items in the test adequately reflected the specific content of the subject and as a guide in the distribution of the test items in so far as the cognitive process dimensions are concerned (Gabuyo, 2012). Lastly, after the administration of the instrument, test for reliability and item analysis were employed.

2.4. Data Analysis
Reliability of the test through its internal consistency was established using the Kuder- Richardson Formula 21 (Popham, 2017). For the item analysis, the upper 1/3 and the lower 1/3 of the group based on score after they have been arranged from highest to lowest were considered. Index ranges were used as bases to describe the difficulty level and discrimination level of each item of the test.

3. RESULTS AND DISCUSSION
Face validity was done by examining the assessment tool in terms of its appropriateness, suitability, and mechanics in the construction. Based on the criteria, the test had observed the criteria along face validity. First, the directions of the test were complete, clear, and concise. Second, mechanics of the test such as the spacing of items, font size and font style used, margins, alignment of options, and use of other symbols such as subscripts were followed. Finally, language mechanics such as grammar, spelling, and punctuation marks were also observed.

Table 1: Distribution of Test Items Based on the Cognitive Process Dimensions

Dimensions	Frequency	Percent
Remembering	9	18
Understanding	12	24

Applying	14	28
Analyzing	10	20
Evaluating	5	10
Total	50	100

Content validity is a type of validation that refers to the relationship between a test and the instructional objectives. It establishes content so that the test measures what it is supposed to measure. As Airasian (2000) has stated it, this is the most important type of validity for a classroom test and the evidence of the content validity of a test is found in the Table of Specification (TOS). All the cognitive process dimensions were represented by the items in the test. This means that the test measured higher order thinking skills of the students. This attests that the best type of traditional test is the multiple- choice type since it can assess learners in multiple competencies through the different cognitive skills (Dixson and Worrell, 2016).

Table 2: Reliability Coefficient Obtained Using KRF 21

Minimum	17
Maximum	42
Total Items	50
Mean Score	28
Standard Deviation	5.92
KRF 21 Correlation Coefficient	0.81
Interpretation	High reliability; good for classroom test. There are few items that need to be improved.

The internal consistency of the test scores of the students is 0.81 which means that it has high reliability. The highest score obtained in the test was 42 out of 50 items and the lowest score was 17 with mean score of 28 and a standard deviation of 5.92. As mentioned by Gronlund, et. al (2002), a correlation coefficient of 0.70 or higher means that the test is reliable.

Table 3. Index of Difficulty (D_f) of the Items

Item Number	Index of Difficulty	Description
5 Items: 14, 17, 34, 47, 48	0.21- 0.40	Difficult
33Items: 1,2,3,4,5,7,9,10,11,12,13,15,18,19,20,21,26,27,28,29,30, 31,32,35,36,37,38,39,40,42,45,49,50	0.41- 0.60	Moderately Difficult
10 Items: 6,16,22,23,24,25,41,43,44,46	0.61- 0.80	Easy
2 Items: 8,33	0.82- 1.00	Very Easy

Distribution of items based on their index of difficulty shows that most of the items are of moderate difficulty. This comprises of 33 out of the 50 items followed by 10 items in which the index ranges show that they are easy items. This means that most of the items given in the test were within the acceptable level of difficulty. As suggested by Walsh and Betz (2001), an ideal test should contain items whose difficulty indices range from 0.41 to 0.60, but for teacher- made test, 0.30 to 0.70 could be acceptable.

Table 4: Index of Discrimination (D_s) of the Items

Item Number	Index of Discrimination	Description
1 Item: 12	Below- 0.10	Questionable Item
6 Items: 7,8,28, 33,41,45	0.11- 0.20	Not Discriminating
17 Items: 1,2,10,11,13,17,19,21,29,30,34,35, 38,39,44,49,50	0.21- 0.30	Moderately Discriminating
24 Items: 3,4,5,6,9,14,15,16,18,20,22,23,24,25,2 7,31,32,36,37,40,42,43,47,48	0.31- 0.40	Discriminating

2 Items: 26,46	0.41- 1.00	Very Discriminating

There are 24 out of 50 items that are discriminating while 17 are moderately discriminating. It should be noted that a good item properly discriminates bright students from the poor ones. Items that are very discriminating, not discriminating or questionable should be discarded in the test (Walsh and Betz, 2001). The result shows that only a few items in the test did not meet the criterion concerning the discriminating power.

Table 5: Item Category and Remarks

Item Number	Item Category	Remarks
14 Items: 3,4,5,9,15,18,20,27,31,32,36,37,40, 42	Very Good	Retain
17 Items: 1,2,10,11,13,14,19,21,29,30,35,38, 39,47,48,49,50	Good	Retain
2 Items: 17,34	Fair	Retain
11 Items: 6,7,16,22,23,24,25,28,43,44,45	Reasonably Good	Revise
6 Items: 8,12,26,33,41,46	Poor	Reject

The above presentation shows that out of 50 items, 33 were retained. This means that based on item analysis procedure, the items met the criteria of a desirable test particularly the index of difficulty and discrimination power. On the other hand, 11 items are identified to be Reasonably Good; thus, need to be revised. Finally, 6 items were considered poor; hence, to be rejected (Apple Beyerlein, 2005).

Table 6: Measures of Attractiveness of Distracters

Category	Frequency	Percentage
Effective Distracters	127	84.67
Confusing Distractors	8	5.33
Ineffective Distractors	15	10
TOTAL	150	100

Each item of the test has four (4) options; hence, three of which are the distracters. Since there are a total of 50 items in the test, there corresponds a total of 150 distracters. It clearly shows in the table that 127 or 84.67% are effective distracters while 23 are either confusing or ineffective distracters. These effective distracters are those incorrect options that have attracted the students in choosing such options as the answer (Gabuyo, 2012).

Table 7: Summary Table Showing the Item Analysis Procedure

Item #	N	Upper Group		Lower Group		D_f	Interpretation	D_s	Interpretation	Decision
		F	P_u	F	P_l					
1	87	56	0.64	32	0.37	0.51	Moderately Difficult	0.28	Moderately discriminating	Retained
2	87	62	0.71	40	0.46	0.59	Moderately Difficult	0.25	Moderately discriminating	Retained
3	87	61	0.70	33	0.38	0.54	Moderately Difficult	0.32	Discriminating	Retained
4	87	63	0.72	30	0.34	0.53	Moderately Difficult	0.38	Discriminating	Retained
5	87	67	0.77	35	0.40	0.59	Moderately Difficult	0.37	Discriminating	Retained
6	87	76	0.87	42	0.48	0.68	Easy	0.39	Discriminating	Revised
7	87	52	0.60	37	0.43	0.51	Moderately Difficult	0.17	Not discriminating	Revised
8	87	80	0.92	65	0.75	0.83	Very easy	0.17	Not discriminating	Rejected
9	87	68	0.78	36	0.41	0.60	Moderately Difficult	0.37	Discriminating	Retained

10	87	59	0.68	39	0.45	0.56	Moderately Difficult	0.23	Moderately discriminating	Retained
11	87	53	0.61	30	0.34	0.48	Moderately Difficult	0.26	Moderately discriminating	Retained
12	87	42	0.48	35	0.40	0.44	Moderately Difficult	0.08	Questionable Item	Rejected
13	87	54	0.62	34	0.39	0.51	Moderately Difficult	0.23	Moderately discriminating	Retained
14	87	52	0.60	18	0.21	0.40	Difficult	0.39	Discriminating	Retained
15	87	65	0.75	32	0.37	0.56	Moderately Difficult	0.38	Discriminating	Retained
16	87	67	0.77	39	0.45	0.61	Easy	0.32	Discriminating	Revised
17	87	45	0.52	25	0.29	0.40	Difficult	0.23	Moderately discriminating	Retained
18	87	63	0.72	35	0.40	0.56	Moderately Difficult	0.32	Discriminating	Retained
19	87	58	0.67	34	0.39	0.53	Moderately Difficult	0.28	Moderately discriminating	Retained
20	87	64	0.74	37	0.43	0.58	Moderately Difficult	0.31	Discriminating	Retained
21	87	59	0.68	35	0.40	0.54	Moderately Difficult	0.28	Moderately discriminating	Retained
22	87	72	0.83	40	0.46	0.64	Easy	0.37	Discriminating	Revised
23	87	76	0.87	42	0.48	0.68	Easy	0.39	Discriminating	Revised
24	87	78	0.90	45	0.52	0.71	Easy	0.38	Discriminating	Revised
25	87	81	0.93	50	0.57	0.75	Easy	0.36	Discriminating	Revised
26	87	69	0.79	32	0.37	0.58	Moderately Difficult	0.43	Very discriminating	Rejected
27	87	60	0.69	29	0.33	0.51	Moderately Difficult	0.36	Discriminating	Retained
28	87	53	0.61	37	0.43	0.52	Moderately Difficult	0.18	Not discriminating	Revised
29	87	47	0.54	28	0.32	0.43	Moderately Difficult	0.22	Moderately discriminating	Retained
30	87	53	0.61	34	0.39	0.50	Moderately Difficult	0.22	Moderately discriminating	Retained
31	87	56	0.64	29	0.33	0.49	Moderately Difficult	0.31	Discriminating	Retained
32	87	64	0.74	30	0.34	0.54	Moderately Difficult	0.39	Discriminating	Retained
33	87	82	0.94	67	0.77	0.86	Very Easy	0.17	Not discriminating	Rejected
34	87	45	0.52	25	0.29	0.40	Difficult	0.23	Moderately discrimineating	Retained
35	87	50	0.57	32	0.37	0.47	Moderately Difficult	0.21	Moderately discriminating	Retained

36	87	65	0.75	38	0.44	0.59	Moderately Difficult	0.31	Discriminating	Retained
37	87	58	0.67	29	0.33	0.50	Moderately Difficult	0.33	Discriminating	Retained
38	87	53	0.61	34	0.39	0.50	Moderately Difficult	0.22	Moderately discriminating	Retained
39	87	54	0.62	29	0.33	0.48	Moderately Difficult	0.29	Moderately discriminating	Retained
40	87	58	0.67	24	0.28	0.47	Moderately Difficult	0.39	Discriminating	Retained
41	87	71	0.82	58	0.67	0.74	Easy	0.15	Not discriminating	Rejected
42	87	62	0.71	30	0.34	0.53	Moderately Difficult	0.37	Discriminating	Retained
43	87	79	0.91	50	0.57	0.74	Easy	0.33	Discriminating	Revised
44	87	80	0.92	61	0.70	0.81	Easy	0.22	Moderately discriminating	Revised
45	87	50	0.57	36	0.41	0.49	Moderately Difficult	0.16	Not discriminating	Revised
46	87	75	0.86	39	0.45	0.66	Easy	0.41	Very discriminating	Rejected
47	87	44	0.51	15	0.17	0.34	Difficult	0.33	Discriminating	Retained
48	87	46	0.53	18	0.21	0.37	Difficult	0.32	Discriminating	Retained
49	87	63	0.72	40	0.46	0.59	Moderately Difficult	0.26	Moderately discriminating	Retained
50	87	60	0.69	40	0.46	0.57	Moderately Difficult	0.23	Moderately discriminating	Retained

4. CONCLUSION

The results revealed that the summative test has its face validity based on the criteria, has high content validity, and high internal consistency. Item analysis revealed also that most of the items have moderate difficulty level, have discriminating power and have effective distracters. Most of the items based on item analysis were retained. Hence, the developed assessment instrument was of quality. In this light, it is suggested that teachers constantly develop quality assessment tools to guarantee valid and reliable assessment of learners and to meet the standards of high- stake assessments that are used for board examinations.

REFERENCES

1. Akhmadeeva, L., Hindy, M & Sparrey, C, (2013). Overcoming Obstacles to Implementing An Outcome-Based Education Model: Traditional Versus Transformational OBE. Canadian Engineering Education Association (CEEA13) Conference.
2. American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFT, NCME, & NEA). (1990). Standards for teacher competence in educational assessment of students.Washington, DC:
3. Apple, D. K., Baehr, M., & Beyerlein, S. W. (2005). Assessment methodology. Faculty Guidebook: A comprehensive Tool for Faculty Performance. Available online:
4. https://www.webpages.uidaho.edu/ele/Scholars/Practices/Assessment/Resources/Assessment_Methodology.pdf
5. Asaad, Abubakar S. and Wilham M. Hailaya.Measurement and Evaluation (Concepts and Principles) Rex Bookstore. Manila, Philippines, 2004.
6. Balagtas, M., &Ferido, M. (2007). Building Teacher Educators' Capacity in Student Assessment: Where We Have Been and Where are We Heading?.In B. Atweh, M. Balagtas, A. Bernardo, & M. Ferido, & I. Macpherson (Eds.) Ripples of Change: a Journey of Pre-service Teacher EducationReform in the Philippines (pp. 220-249). Quezon City: Commission of Higher Education.

7. Buendicho, F. C. Ph.D. 2010. Assessment of student learning 1.Rex Bookstore, Inc., Sampaloc, Manila.

8. Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), Handbook of classroom assessment: Learning, adjustment, and achievement (pp. 1-32). San Diego: Academic Press.

9. CMO No. 74, s. 2017: Policies, Standards, and Guidelines for Bachelor of Elementary Education (BEEd)

10. CMO No. 75, s. 2017: Policies, Standards, and Guidelines for Bachelor of Secondary Education (BSEd)

11. CMO No. 76, s. 2017: Policies, Standards, and Guidelines for Bachelor of Early Childhood Education (BECEd)

12. CMO No. 80, s. 2017: Policies, Standards, and Guidelines for Bachelor of Physical Education (BPEd)

13. Commission on Higher Education (2004). CHED Memorandum Order No. 30. Revised Policies and Standards for Undergraduate Teacher Education.Curriculum. Pasig City: CHED

14. Department of Education (2004).DepEd Memo No. 33.Implementing Guidelines for a Performance-based Grading System for SY 2004-2005. Pasig City: DepEd.

15. Dixson, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. Theory into practice, 55(2), 153-159. Available online: https://doi.org/10.1080/00405841.2016.114898

16. Gabuyo, Y.A. (2012) Assessment of Learning I. Rex Book Store, Inc., Manila, Philippines.

17. Gronlund, N. E. Assessment of student achievement. Allyn and Bacon A Viacom Company, 7th Edition. 2001.

18. Impara, J. C., Plake, B. S., &Merwin, J. C. 1994.Student assessment tasks and knowledge: Comparing superintendents and elementary and secondary principals. Journal of School Leadership.

19. Linn, R. L., &Gronlund, N. E. (2000). Measurement and assessment in teaching (8thed.). Upper Saddle River, N J: Prentice-Hall.

20. K-12 Academics (2019) Education assessment and evaluation. Retrieved from https://www.k12academics.com/education-assessment-evaluation

21. Padua, Robert N. and Rosita G. Santos. Educational Measurement and Evaluation: Theory, Practice and Applicatio. Metro manila, Philippines: Katha Publishing Co. Inc, 1999.

22. Philippine Professional Standards for Teachers (2017). Department of Education - Teacher Education Council.

23. Plake, B. S., &Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), Handbook of classroom assessment: Learning, adjustment, and achievement (pp. 53-68). San Diego: Academic Press.

24. Popham, W.J. (2017) Classroom Assessment: What Teachers Need to Know, 8e. Pearson Publishing.

25. Walsh, Bruse and Nancy E.Betz. (2001)Test and assessment. 3rded. New Jersey: Prentice Hall.