Validation of the Pre-licensure Examination for Pre-service Teachers in Professional Education Using Rasch Analysis

Jovelyn Delosa

Introduction

Teachers play a crucial role for students to demonstrate the expected learning outcomes. Teachers are curriculum planners, assessors and curriculum implementers (McTighe and Wiggins 2005). Teachers need to establish a strong nexus between the content that they are giving to their learners, the methods they used and the assessment they employ. One of the roles of teachers is to assess the learning process and outcomes. Tests are widely used as a form of assessment in universities of many countries to measure student learning, to rank students and issue certification. Literature also outlines arguments on how validity issues of tests. There are critics about tests but there are also groups who argued that tests can be used as long as they are reliable and valid. This prompted the researcher to look into her own context specifically in teacher education and examine the test used for the pre-service teachers in preparation for the Licensure Examination for Teachers in the Philippines and look into its psychometric qualities. Using a Mock LET instrument, this paper discusses the strengths of the Rasch model as a psychometric tool and analysis technique, referring to person-item maps and differential item functioning.

The pre-service teachers of Xavier University enroll themselves in a subject called Education 60, a Refresher Course for the Licensure Examination for Teachers. At the end of the course, they are given a test that sets similar to the real Licensure Examination for Teachers (LET). This paper aimed to examine the validity of the test and determine whether this mock test measures the knowledge and skills expected of them. Validation is important to ensure that the given tests are appropriate and the results are trustworthy.

J. Delosa (✉)
Xavier University, Cagayan de Oro, Philippines
e-mail: jingdelosa@gmail.com

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_8

Literature Review

Role of Assessment

Literature confirms the various roles of assessment in learning. Assessments inform all stakeholders about instruction and learning outcomes. Specifically in higher education, assessment is very crucial especially when results of such are the bases of ranking and certification. It drives instruction to important goals and standards (Brew et al. 2009; Rieg and Wilson 2009). Assessment is a component equally important if schools and stakeholders want to increase educational outcomes (Guskey 2003). Assessment is defined as the process of observing, interpreting and making decisions about learning (Griffin 2009). In addition, this process of collecting of evidences can take many forms which include tests, performances, work samples and many others (Black and Wiliam 1998; Griffin 2009). Assessments are given to facilitate learning, facilitate teaching, for school and professional requirements. They provide feedback, motivate students and inform how teachers deliver content to their students and improve their methodologies.

Assessing student performance is one of the most critical responsibilities of classroom teachers (Stiggins as cited in Mertler and Campbell 2005). Assessment information must be correct, reliable, and valid because these information can do a lot to improve classroom instruction (Mertler and Campbell 2005). Teachers are expected to show expertise in assessment (Campbell, Murphy, and Holt as cited in Mertler and Campbell 2005). However, there are issues with assessment. Research has documented that teachers' assessment skills are generally weak (Brookhart; Campbell, Murphy, and Holt as cited in Mertler and Campbell 2005). Among the various roles of teachers, assessment of student learning is somehow left out. Teachers experience inadequacy and difficulty in carrying this role (Murray as cited in Mertler and Campbell 2005), thus, a need to review new frameworks of assessment.

Tests

One of the widely utilized types to assess student outcomes is the use of tests. It could be multiple-item tests, matching type, and true/false tests or fill in the blank. Research shows an important relationship between the quality of classroom assessments and achievement as measured by standardised tests (Mertler and Campbell 2005). Teachers trust the results of tests because of their direct relation to classroom instructional goals and results are immediate and easy to use for analysis since it is still on the student level (Guskey 2003). Webber and Lupart (2012) argued that classroom assessment is the most important kind of assessment.

Multiple-choice items are widely used on classroom tests in colleges and universities (Mavis et al. 2001; McDougall 1997 as cited in DiBattista and Kurzawa 2011). A typical multiple-item test consists of a question and a set of two or more

options that includes the correct answer and distracter options. Multiple-choice tests are commonly used in the classroom and for licensure purposes because grading is easy (DiBattista and Kurzawa 2011) and allows for broader coverage of the topics (Bacon 2003 as cited in DiBattista and Kurzawa 2011). The fact that tests are widely used and inferences are made from the test results raises the challenge for teachers to give valid tests to students to ensure fairness (Popham 2002, 2004 in Webber and Lupart 2012).

Characteristics of a Good Test

A good test should be relevant to the needs of the learners which means that testing is not just an end in itself. It has an educational impact to both learners and teachers and it matches with the curriculum. In constructing tests, teachers should consider the feasibility of the test which includes the time for construction, time for administration, time for scoring and time for reporting (Fuentealba 2011). A good test has validity. It refers to the ability of an instrument to measure the attributes which could be knowledge or skill that it is aiming to measure (Fuentealba 2011; Purya and Nazila 2011). Validity includes looking on the importance of content to me measured, instructions, wording of the questions, spelling and grammar, level of difficulty, arrangement of items, number of items, time and the errors in scoring (Fuentealba 2011). Validity is probably the most important criterion in judging the effectiveness of a measurement tool (Alagumalai and Curtis 2005). Validity has four types: content, predictive, concurrent, and construct validity. Construct validity is the focus of this paper. Construct validity is concerned with the extent to which a test reflects the underlying construct the test is supposed to assess (Purya and Nazila 2011). Valid tests are reliable tests. Reliability is the ability of a test to measure the attributes consistently (Al-Sabbah et al. 2010; Griffin 2009). Reliable assessment when tasks get the same results regardless of when they are administered (Al-Sabbah et al. 2010).

Establishing the psychometric qualities of a test is highly essential and to help teachers we need to look into classical and item response theories. They serve as guiding principles in decision-making of teachers with student learning. Teachers should be equipped with a degree of test literacy which includes test construction test analysis and testing theories. The ability to select and design assessment tools is highly expected of every teacher (Rieg and Wilson 2009).

Classical Test Theory

Test theories provide a framework about the relationship of test and item scores to true scores and ability scores (Hambleton and Jones 1993). Test theories are important in educational measurement because they provide a guiding post for

considering issues of handling measurement errors. “Different models and theories will handle error differently”. One may assume normal distribution of errors and the other may have another assumption (Hambleton and Jones 1993).

Classical test theory introduces three concepts: observed score, true score, and error score. True score is the difference between test score and error score. CTT is a “psychometric theory laid by Charles Spearman in 1904 that allows the prediction of outcomes of testing such as the ability of the test takers and the difficulty of items” (Alagumalai and Curtis 2005, p. 5). This theory explains the concept of an observed score that is manifest and this score is composed of a true score and an error which are both latent in nature. It is a model for testing which is widely used in constructing and evaluating fixed length tests. Although the major focus of CTT is on test-level information, item statistics, like item difficulty and item discrimination, are also important. The p-value, which is the proportion of examinees that answers an item correctly, is used as the index for the item difficulty. A higher value indicates easier items. The item discrimination index is the “correlation coefficient between the scores on the item and the scores on the total test and indicates the extent to which an item discriminates between high ability examinees and low ability examinees”. Similarly, the point-biserial correlation coefficient is the “Pearson r between the dichotomous item variable and the continuous total score variables” (Alagumalai and Curtis 2005, p. 7). However, CTT has limitations (Alagumalai and Curtis 2005). The two statistics which are item difficulty and item discrimination are group and test dependent. The increase or decrease and the homogeneity or heterogeneity of the group affects the results; and test difficulty has a direct effect on test scores (Hambleton as cited in Hambleton and Jones 1993; Boone and Scantlebury 2006). There is no basis to predict how an examinee may perform on a particular item. The true score is not an absolute characteristic of a test taker since it depends on content. A simple or more difficult test would result in different scores for examinees with different levels of ability. Therefore, it is difficult to compare test takers’ results between different tests (Boone and Scantlebury 2006). In the discussion about tests above, testing for many years has a key role in assessing learning (DiBattista and Kurzawa 2011).

However, there are also issues with the use of tests. Guskey (2003) argued that for assessment to be of use, teachers should change their views on how to interpret results. Particularly for multiple-item tests, critics argued that this type of test can be subject to guessing (DiBattista and Kurzawa 2011) and questions on validity and bias (Boone and Scantlebury 2006; Stiggins 1999). Guskey (2003) added that despite of the importance of assessment education today, few teachers receive much formal training in assessment design and analysis.

With this given fact of how tests in universities and countries are used to rank and certify, it is logical to look carefully at the tests that we construct. The purpose of an examination is to infer about students knowledge, skills and values, make inferences about an overt behavior to a covert quality and this poses a problem with scores. There are groups which questioned the reliability of raw scores in representing a person’s true ability.

It is sensible that with the limitations of CTT as cited above, teachers can use another theory, the Item Response Theory (IRT). The past 50 years has not only seen the strengths and limitations of the classical test theory but also acknowledged the use of new approaches to educational measurement.

Psychometricians were interested in psychometric theory which would describe examinees' achievement as independent of the particular choice of items that were used in a test. Classical item statistics such as item difficulty and item discrimination and test statistics such as test reliability are sample dependent; however, thousand of excellent tests have been constructed in this way. Classical test theory and related models have been used and are still used successfully for over 60 years and many testing programs are deeply founded in classical measurement models (Hambleton and Jones 1993).

IRT and Rasch Model

Item response theory is a theory about the relationship of an examinee's performance in an item and with his ability. The items are discreet or continuous and the scores are dichotomous or polytomous. This theory argues that an item can measure single or multiple abilities (Hambleton and Jones 1993).

IRT was originally developed to overcome the issues with CTT. IRT assumes that the latent ability of a test taker is independent of the content of a test. The relationship between the probability of answering an item correctly and the ability of a test taker can be shown in different models depending on the nature of the test. It also assumes that it does not matter which items are used making the possibility to compare test takers (Wiberg 2004).

One of the models in IRT is the Rasch model, named after the Danish mathematician and statistician George Rasch, which is a probabilistic model with two distinguishing properties: invariance (Boone and Scantlebury 2006) and interval scaling which are obtained if unidimensionality occurred that is when the data fit the model (Purya and Nazila 2011). Unidimensionality is achieved when the instrument measures one trait at a time (Wolfe and Smith 2007 in Purya and Nazila 2011).

It specifies the probability of a correct response on an item as a function of the difference between the ability of person and the difficulty of a test item (Webber and Lupart 2012).

When the student ability equals the difficulty of the item, there is a 50 % probability that the student will answer the item correctly (Webber and Lupart 2012). The model is probabilistic based upon logits (Lamb et al. 2011). If data fit the model, the scale is defined as being unidimensional; one can be confident that the item measures are independent of the person measures and vice versa (Purya and Nazila 2011). However if the data do not fit the model, this can be because the

instrument items may be measuring another construct and Rasch analysis allows these items to be identified. When scales are multidimensional, summing of item scores may cause misleading assumptions to be made (Belvedere and de Morton 2010).

Fit indices are used to check the relevance of the test content to the intended construct. Misfitting items may be measuring a totally different and irrelevant construct. Moreover, person-item map and item strata are two important criteria for checking the representativeness of the items (Wolfe and Smith 2007 in Purya and Nazila 2011).

Test takers scores are expressed in logit measures which are the conversion of raw scores to logits through use of the Rasch model. If researchers or educational practitioners do not convert raw scores to equal interval measures then the results of their analysis may provide incorrect and/or incomplete information on student performance. If a researchers uses only raw scores, then incorrect conclusions may be reached by using raw score data for parametric tests of student (Boone and Scantlebury 2006).

Rasch statistics provide similar psychometric information to traditional analyses. A point biserial expresses item discrimination, and a "person separation index". Classical test theory provides a single standard error of measurement (SEM). However, in Rasch measurement each item and test taker is provided an error term. The error in each item is considered and the range of each person's error before item removal. One technique utilized in Rasch measurement is an evaluation of an individual person's responses to test items to the model known as "fit" statistics. Fit statistics are used to assure whether the test is unidimensional and guide one to decide upon the way the test should be scored. However, in case of multidimensionality, separate scores should be reported for each dimension. Thus, fit statistics provide helpful evidence with regard to the structural aspect of construct validity (Beglar 2010; Purya and Nazila 2011). Item fit statistics evaluate the predictability of test takers' answers, given their overall ability (Boone and Scantlebury 2006).

Differential Item Functioning

Rasch analysis also facilitates the assessment of differential item functioning (DIF). DIF occurs when persons of the same ability have items that operate differently based on another variable, such as age or gender. Assessment of DIF is important as it improves generalisability of the instrument by testing that item response patterns are similar across groups. Rasch analysis also facilitates the investigation of item thresholds. If the probability of each item response category is not in the expected order, this results in a disordered threshold (Belvedere and de Morton 2010).

Hambleton and Jones (1993) summarized the main differences between classical test theory and item response theory. Classical test theory is linear, the level is for the whole test, assumptions are easy to meet test data, item-ability relationship is not specified, test scores or estimated true scores are reported on the test-score scale,

item and person parameters are sample dependent, while on the other hand, item response theory is nonlinear, level of analysis is by item, assumptions are difficult to meet with test data, item characteristics functions are available, ability scores are reported on a transformed scale, item and person parameters are sample independent if model fits the data. CTT is test based while IRT is item based. CTT permits no consideration of how participants respond to a specific item. IRT permits the analysis of the probability of an examinee answering an item. In item response theory; the measurement specialist is allowed a greater flexibility.

IRT has limitations too because of its complex mathematical formulations; however, with technology nowadays it is now very possible for teachers to analyze tests using software. Another thing is if a test is poorly designed, computing an overall measure using all test items may be impossible and results may only be evaluated at the item level. A Rasch analysis may take longer than a traditional analysis, but it provides a deeper understanding of instrument's strengths and weaknesses (Boone and Scantlebury 2006).

Licensure Tests

After the comparison between CTT and IRT, let us take a look at a specific kind of test, licensure tests. Many countries including the Philippines have practiced the national examination for licensing of teachers before teachers are considered professional teachers. The role of teachers in education has been identified as the most significant of all school factors that affect student learning and with this belief, policymakers want to guarantee a level of quality through a licensure system. Teachers pass licensure tests given by the government before they can work in the classroom. Licensure is defined by the US Department of Health "as a process by which an agency of government grants permission to an individual to engage in a given occupation upon finding that the applicant has attained the minimal degree of competency required to ensure that the public health, safety, and welfare will be reasonably well protected" (Shimberg 1981). Teacher preparation programs are being held to high standards in order to prepare the best teachers to meet the challenges of today's diverse classrooms (Rieg and Wilson 2009). The main mission of teacher education in the Philippines is the training and preparation of globally competitive teachers who are equipped with the principles, aspirations and values and possess pedagogical knowledge and skills (CMO 30, 2004). With this goal, teacher education institutions are challenged to develop and guide pre-service teachers towards this direction. To professionalize the teachers, RA No. 7836, known as Professionalization Act for Teachers is implemented to "strengthen, regulate and supervise the practice of teaching profession in the Philippines by prescribing a license" to teachers certified by the Professional Regulation Commission (PRC). The Professional Regulation Commission works hand in hand with the Commission in Higher Education (CHED) with the Teacher Education Institutions (TEI's) in the implementation of this law. CHED issued CHED

Memorandum 30, s. 2004 (CMO 30, 2004) which provided a list of the desired competencies and subject areas to be taken by pre-service teachers. The list included General education Courses (Science, Math, English, etc.); Professional courses which have three subgroups, theory subjects, strategies subjects and field subjects; and Specialization courses. PRC issued a list of competencies based on the National Competency-Based Standards (NCBTS) and their weights. For elementary education (BEED), 40 % is allotted for general education and 60 % for professional education; general education (20 %), professional education (40 %) and specialization (40 %). The TEI's created partnership with the Department of Education in coming up with the Experiential Learning course for the pre-service teachers which provided students with actual learning experiences.

Examining the PLET Items

Pre-licensure Examination for Teachers Test (PLET)

This section discusses the background of the practice test used before the teachers take the real Licensure exam for teachers. This particular study focused on the professional courses for LET. These items are constructed by the different teachers teaching the subjects. The LET coordinator compiled all the items and gives the test as the final examination for the subject Education 60. It is a 6-unit course. The class runs for 14–15 Saturdays and the sessions from 1–7 pm. The sessions cover all the competencies indicated in the LET Primer and address the three components of General Education, Professional Education and Major area of concentration.

To demonstrate content validity, it is important to establish that the questions on a test represent a content domain that the test sample (Shimberg 1981). The questions of this test are based on the list of competencies as seen in the LET Primer. When the researcher reviewed the items, 90 % of the items are reflected in the competencies in the LET Primer. However, nothing much has been done to examine the construct validity of this test except that of looking at the overall scores of the students and checking what items were not answered right. This then is the main purpose of this paper, to investigate the one construct validity of the test using the IRT perspective.

Item Analysis Using the Rasch Model

The Rasch model is used for item analysis. It has features which Rasch labeled as 'specific objectivity' and unidimensionality. Unidimensionality is when the items measure the same construct while specific objectivity states that two person who are taking the tests are compared and such comparison is not based on that items are included in the test. Item analysis using Rasch model can give practitioners the

answers why particular tests are not functioning as they should and can guide them on which items to include or to omit (Choppin 1983). Validity was assessed by evaluating the fit of individual items to the latent trait as per the Rasch model and examining if the pattern of item difficulties was consistent with the model expectancies.

The Data

The test is composed of 200 items. These items tried to measure the students' understanding of professional education which is one of the major components in the real Licensure Examination for Teachers. This test investigated their knowledge about theories, assessment skills and other pedagogical areas of the teaching profession The 200 items were subjected to Rasch analysis to check their fit and if all these items measure one construct which is professional education. Fit indices were examined closely to check the relevance of the items as part of content validity. However, the results of the analysis showed that the 200 items are composed of other constructs because of the variety of patterns of responses no matter how many times the test was rerun in Conquest. The researcher decided to check the items again. There are 7 constructs that emerged from the 200 items based on the content analysis of the whole test vis a vis the list of teaching competencies: understanding of curriculum concepts, understanding of teaching profession concepts, knowledge and application of educational technology concepts, understanding of social dimension concepts, knowledge and application of assessment concepts, application of teaching principles, and understanding of the theories of learning. Curriculum items summarise topics about the nature of curriculum development and there are 6 items. The teaching profession domain is about knowledge on the laws that govern the teaching profession and the essence of the profession; it has 7 items. Educational technology items include competencies on the use of technology in the classroom with 14 items. Social dimension concepts include understanding of the role of society in education with 23 items. The assessment construct which describes various concepts of assessment knowledge and skills have 41 items; principles of teaching which is about the strategies in teaching has 47 items and theories of learning has 62 items. The last domains have more items compare with the rest since these topics belong to subjects which are credited for 6 units in the entire pre-service education.

Findings

Item Analysis of the Curriculum Items

The curriculum items were subjected to Rasch analysis using the residual-based fit statistics. The important information considered were the Infit Weighted Mean Square (IWMS) and the t-statistic (T) which determined whether an item followed

the requirements of measurement. A range 0.80–1.20 (Wright and Linacre 1994) was used for IWMS since LET is a test, and −2 to +2 for calculated T (Wu and Adams 2007) to indicate acceptable mean fit. The items with mean square values falling above 1.2 were considered under fitting and suggests the presence of unexpectedly high variability (Bond and Fox 2007 in Franchignoni et al. 2011) and do not discriminate the students with high ability from those with low ability while values below 0.8 were over fitting items and gave redundant information and too predictable pattern (Wright and Linacre 1994). The items that do not fit the model were removed one at a time. In this paper, only the initial and final analyses results are presented. The reliability was evaluated in terms of 'separation', defined as the ratio of the true spread of the measures with their measurement error (Franchignoni et al. 2011).

The initial analysis included all the curriculum items and 152 test takers. The results are tabulated in Appendix A. In the first run of the analysis, all items belong to the 'good' fit so there was no need to rerun it. The item difficulty values are shown in Table 1 and they are expressed in terms of logit, the unit used in Rasch logit interval scale which allows person and item to be placed on a common scale (Wright and Linacre 1994). The scale consists of numbers from $-\infty$ to $+\infty$ with 0 in the middle indicating average difficulty for item and person (Bond and Fox 2007). Items with estimates above 0 (positive values) are more difficult items and those below 0 (negative values) are easier items. However, the items level of difficulty and arrangement is needed to be revisited because the items were not arranged well. Item 1 was easy then item 2 was very difficult. Additionally, items in this construct showed good discrimination ability.

The Rasch model transforms raw item difficulties and raw person scores to equal interval measures. These measures are used to map persons and items onto a linear scale. Items ranged from easiest which is located at the base of the graph and to hardest located at the top of the graph. Persons are plotted as a function of their ability, with the more able students at the top of the graph, and less able students at the base. Items plotted above any person are harder than the person's ability level and items below a person are those items for which the person has a greater than 50/50 chance of correctly answering (Santelices and Wilson 2012).

Item Analysis of the Teaching Profession Items

The 7 items of teaching profession were subjected to Rasch analysis using the residual-based statistics. Their IWMS and T-values were examined for fit to measurement requirements. The complete analysis is reported in Appendix B. The infit weighted mean square was used to identify the rating of the items that deviate from expectations (Wright and Linacre 1994). All items were in 'good' fit with IWMS falling within the range of 0.80 to 1.20. This is the final results since all items conformed to the measurement requirement. Table 2 presents the results of the analysis showing that all items belong to the desired IWMS and T-values. Similar

to the curriculum items, the arrangement of items according to the level of difficulty needed some attention. Item 3 is the most difficult item; and item 6, the easiest. Majority of the students were in the average level of ability.

Item Analysis of the Educational Technology Items

The results revealed that in the initial analysis all items achieved the required IWMS and T-values. The separation reliability which is defined as the ratio of the true spread of the measures with their measurement error (Bond and Fox 2007, pp. 40–41 as cited in Franchignoni et al. 2011) is high (0.986). It can be observed that the average ability of the students is above the difficulty of the items. Item 3 is the most difficult item and item 6, the easiest. Majority of the students were in the average level of ability. It is essential to review the difficulty level of the items because items 3, 13, and 14 are too easy and far below the ability of the students and items 1 and 6 are too difficult for the students.

Item Analysis of the Social Dimension Items

There are 23 items for social dimension. The results showed that in the initial analysis, there was 1 underfitting item which was removed and the analysis was rerun. This process was done for the second time. In the second analysis, all the 22 items conformed to the fit requirement. All the items have IWMS and T-values that are within the range of the required measurement values. Table 3 shows the second and final analysis of the items, their particular estimates, error, IWMS and T-values. The social dimension items have a separation reliability of 0.988 which conformed to the required value of equals to and more than 0.90 (Wright and Linacre 1994). Checking at the estimates, the items were not arranged well from easy to difficult. Item 2 which was removed in the second run has a T-value of 2.2 which means that this is an under fit item (Wu and Adams 2007). Half of the items are difficult items with estimates of positive values. Most of the items have a discrimination value which was quite good except for item 11 which has a negative value which means that this item did not discriminate the students with high ability from the students with low ability. Even if the items have reached the required measurement fit it is important to reexamine the structure of the questions to improve them.

Item Analysis of the Assessment Items

The initial analysis included 41 items which were examined for measurement fit. The item with the ‘worst’ fit (based on the T-value and IWMS) was first removed

and the analysis was done again. This step was performed until no misfitting items were found. One item was found to be over fitting and two to be under fitting based on their T-values. The under fitting item was first removed and the analysis was rerun. This process was done until all items fit the measurement requirement. Two more analyses were done to come up with 39 'good items' whose IWMS and t-statistics fell under the required range of measurement. Item 26 was the most difficult item that no one got it right and items 10 and 35 were the easiest that all students got them right.

Item Analysis of the Principle of Teaching Items

Similar with the other constructs, the 47 items of the principle of teaching were subjected to the Rasch analysis for dichotomous items. In the initial analysis, two under fitting items were found. The item with the biggest T-value was removed first and the data was rerun. The second analysis showed that all the 46 items were 'good' items as based on their infit weighted mean square and T-value. Forty-five percent of the items or 21 items out of 46 items were difficult items and the separation reliability is 0.987. Some items need to be reviewed because they are too difficult (PT 24, 35, 36, 43) and too easy (PT 4, 15, 17, 25 and 37).

Item Analysis of the Theories of Learning Items

The 62 items of the theories of learning were subjected to the Rasch analysis for dichotomous items. In the initial analysis, three under fitting items were found. The item with the biggest T-value was removed first and the data was rerun. The third analysis showed that all the 59 tems were 'good' items as based on their infit weighted mean square and T-value. The separation reliability is 0.987. Most of the items were also below the average level of difficulty. There were also outliers, items which were extremely difficult that no student got the right and extremely easy that even students with ability below average got them right.

Differential Item Functioning (DIF) of the Pre-licensure Examination for Teachers (PLET) Items

The 7 constructs of the 200 item pre-Licensure test were submitted to Differential Item Functioning (DIF) after examining that the items are good items. DIF was done to examine if the items behave well across the two groups: male and female in this particular study. DIF is necessary to check whether test items have same response patterns to ensure generalizability (Boone and Scantlebury 2006).

All items from the 7 constructs which have acceptable fit were analyzed. There were 6 items for curriculum, 14 items for educational technology, 7 items for teaching profession, 22 items for social dimension, 38 items for assessment, 46 items for principle of teaching and 59 items for theories of learning. Items of each construct were examined separately using Conquest 2.0 software (Wu et al. 2007). In DIF detection, these indicators were considered: an approximate Z-statistic (calculated T-value) calculated by dividing the estimate by the standard error; comparing the standard error with the parameter estimate (Wu et al. 2007); checking if the chi-square value is significant (Wu et al. 2007) and verifying the difference between the estimates. A calculated T-value less than −2.0 or greater than +2.0 points out significant DIF between two groups and in this test chi-square value of equal to and less than 0.05 is significant.

The 7 constructs of the 200 item pre-Licensure test were submitted to Differential Item Functioning (DIF) after examining that the items are good items. DIF was done to examine if the items behave well across the two groups: male and female in this particular study. DIF is necessary to check whether test items have same response patterns to ensure generalisability (Boone and Scantlebury 2006). All items from the 7 constructs which have acceptable fit were analyzed. There were 6 items for curriculum, 14 items for educational technology, 7 items for teaching profession, 22 items for social dimension, 38 items for assessment, 46 items for principle of teaching and 59 items for theories of learning. Items of each construct were examined separately using Conquest 2.0 software (Wu et al. 2007). In DIF detection, these indicators were considered: an approximate Z-statistic (calculated T-value) calculated by dividing the estimate by the standard error; comparing the standard error with the parameter estimate (Wu et al. 2007); checking if the chi-square value is significant (Wu et al. 2007) and verifying the difference between the estimates. A calculated T-value less than −2.0 or greater than +2.0 points out significant DIF between two groups and in this test chi-square value of equal to and less than 0.05 is significant.

DIF in Curriculum Items

The overall results of DIF analysis of the curriculum items by gender shows that the LET curriculum items exhibited no DIF as evident in its T-value of 1.48 and the parameter estimate is lower than twice its standard error. The results also show that on average female pre-service teachers perform higher the males with a logit difference of 0.148 but this difference is not significant (chi-square p value = 0.141; calculated T-value = −1.48).

The item level results indicate that females achieved higher in 3 items and in the same manner, males achieve higher in 3 items; however, the difference is not statistically significant. Curriculum items behaved the same between the two groups.

The result shows a minimal difference of 0.296 of performance between males and females yet it is to be taken into account that this difference is not significant.

DIF in Teaching Profession Items

The results show that the LET teaching profession items exhibited DIF based on the calculated T-value which is more than −2 (T-value = −4.98). The female pre-service teachers on average scored higher than the male pre-service teachers in the knowledge and application of concepts about the teaching profession with a difference of 1.006 logit which indicates a gap of 2 years (Griffin 2009). The parameter estimate is more than twice its standard error and the fact that the chi-square value is smaller than 0.05 (significant level = 0.001) indicate that this difference is statistically significant. These results have some implications on revisiting the items about the teaching profession. These items try to measure understanding about teaching as a mission, a vocation and profession. The items should function fairly to both groups to establish fairness even if in the Philippines, the teaching profession is mostly embraced by women.

However, at the item level, the results varied slightly. Based on the calculated T and the estimate compared with twice of the standard error, there is only one item that shows DIF and the rest of the items fall along the range of −2 to + 2 for calculated T. Out of the 7 items, males performed higher than females in 4 items. Nevertheless, the results strongly suggest that the items behaved differently between the two groups based on gender and therefore, it is important that these items need to be reviewed. The result shows that the overall teaching profession items manifest an achievement difference of 1.006 between males and females with the female teachers outperforming the males.

DIF in Educational Technology Items

DIF is not evident in the LET educational technology items. The overall results reveal that females on average got a higher achievement than the males with a logit difference of 0.20 which is minimal. DIF analysis in the item level demonstrates a degree of varied item responses by gender. There are 14 items in this construct and females performed higher in 7 items as well as male did. In determining for DIF in the item level, some items showed no DIF as based on the Z-statistic (calculated T-value) and the result of comparing the estimate with twice of the item's standard error like items 4, 6, and 7; however, there is a need to review these items because even if the big difference between estimates even if the difference is not statistically

significant. The test makers of this item can improve the word structure of the item or syntax of the item. The result validates that DIF did not exist in the educational technology items. With the sweeping influence of technology in education, people are trying to cope with these changes, both males and females.

DIF in Social Dimension Items

DIF is not also present in the LET educational technology items. The overall results as shown in Table 11 reveals that females on average got a higher achievement than the males with a logit difference of 0.23; however, their difference between their estimates was not statistically significant. Items in social dimension construct tried to ask students their understanding of the society and the function of school in the different system that governs human activities.

The Millennium Development Goals of the Philippines (MDG) advocated for gender equity as both males and females tried to portray their roles as teachers. The notion that teaching is regarded as a women work has a long history (Skelton 2002). There is a need of male teachers to serve as good role models for males. Schools have the responsibility to ensure gender equity between males and females. Gender equity should be fundamental in educational practices. DIF analysis in the item level demonstrates a degree of varied item responses by gender. There are 14 items where females performed higher and 8 items where the males achieved higher. In determining for DIF in the item level, some items showed no DIF as based on the Z-statistic (calculated T-value) and the result of comparing the estimate with twice of the item's standard error like items 3, 4, and 8, 9, 11, 13, 19, 22; however, there is a need to review these items because even if the big difference between estimates even if the difference is not statistically significant. The test makers of this item can improve the word structure of the item or syntax of the item or the arrangement of the items. It is recommended to recheck the items. The result validates that DIF did not exist in the educational technology items. The females on average performed better than the males. Even DIF did not significantly exist, it is interesting to observe that in item 8, males performed much better than the females who were even below the average ability level but in most of the items, both groups gathered in the middle near the average ability and difficulty.

DIF in Assessment Items

The overall DIF analysis for the assessment items show that the items function the same across the 2 groups ($T = -0.65$). Results revealed that females on average got a higher achievement than the males with a logit difference of 0.26; however, their

difference between their estimates was not statistically significant. Assessment items tried to measure knowledge on both traditional and alternative forms of assessment. Every teacher is expected to acquire the skills on how to assess learning in two methods: the use of tests and authentic assessment. The pre-service teachers of the university were constantly exposed to avenues to hone their assessment knowledge and skills. The results showed that the ability of the test takers is higher than the average ability.

DIF analysis of assessment items in the item level demonstrates that females got a higher achievement in items in 17 items while males achieved better in 29 items, however, in the overall analysis, the females did well than the males in the test. It is also noted that even in the overall analysis, DIF did not exist, there are items that should be reviewed because of the big difference between estimates of males and females though the difference is not significant. The results revealed that items 26, 14 and 24 were items too difficult for the test takers to answer and items 23, 1, 35, 10 and 37 were too easy that everyone got it right. Similarly with teaching profession items, principles of teaching items showed a significant DIF. The content analysis of the topics for these two areas showed that many topics were related much with each other. Again, it was the group of the female which achieved higher than males with a difference of nearly 0.5 logit which is equivalent to 1 year of learning.

DIF in Principle of Teaching Items

The item level analysis presented item 32 showing DIF with a T-value of −2.03 and a difference of 0.89 logits between males and female with females outscoring the males. There are also 18 items which needed some reexamination even if there is no significant difference according to gender due to big difference in the estimates. The principles of teaching items assessed the students' knowledge and application of classroom strategies and methods of teaching. The results clearly confirmed that the ability of the female test takers was above the average ability and the males' ability was below the average level of ability and this difference is statistically significant (sig level = 0.000).

DIF in Theories of Learning Items

The overall DIF analysis for the theories of learning shows that there is no error in the estimates and therefore T cannot be calculated. However, based on the chi-square test of parameter equality = 0.00, df = 1, and Sig Level = 1.000, it shows that there is no significant difference between the two groups. The separation reliability is 0.96 but the 0 chi-sqaure signaled for a careful reexamination of the

items. The value of chi-square which is 0 could be traced to the homogeneity of the group.

The item level analysis of DIF confirms the overall results showed some conflicting results from the overall results because there are two items which shoed DIF on the item level and many items which needed revalidation. There were items whose T-value was within the range of −2 to +2; however, their difference of estimate is quite big and even if the difference is not significant, it warrants some attention.

Conclusion

In summary, out of the 7 constructs, generally, on average the female achieved higher than the males in all the constructs. All the constructs got separation reliability above 0.95.

Females performed better than the male teachers; this result is to be confirmed yet as to who achieves higher in LET in the Philippines, males of females since only one study has been found investigating gender differences in LET performance.

With regards to DIF, teaching profession and principles of teaching items exhibited significant DIF.

In conclusion, each of the 7 construct measures what it is supposed to measure and each construct has good psychometric qualities. Three of the constructs (curriculum, educational technology and teaching profession) have all items which have the required fit; social dimension has only 1 item which was under fitting; assessment has 2 'worst' items, social dimension has 1 under fitting item, principle of teaching having 1 under fit item and theories of education with 1 over fitting item. The separation of all the constructs is good. Items of each dimension possess a good discrimination power of separating the student which has a high ability from the less performing students. However, putting all the items together to measure the general construct on understanding and application of professional education created a problem thus, there is a need to review the competencies in each construct vis-avis the general goal of professional education as a whole. This maybe because of the large number of items which is 200 and in reality it is difficult that all items measure one construct since the items can be related to one another, there could be overlapping of content. This clearly calls for attention especially when 2 of the constructs exhibited DIF. One important aspect that needs much time for re examination is the development of the assessment instrument. Hence, the following recommendations are hereby given.

Recommendations

The LET items can be improved, first by construct then as a whole measuring professional education which is a major component in the BEED and BSED Licensure Examination for Teachers. Based on the results of the literature review on the role of assessment and tests; and the benefits of IRT with the analyses of the test instrument following recommendations are suggested for future practice.

A great need to create an assessment committee in the School responsible for studying the validity of the test specifically content and construct validity. The team has to invite PRC and CHED to discuss the competencies for LET. Teacher just read the LET Primer and each one could have his/her own understanding of the stated competencies. It would be better that everyone has a common understanding of the content. This collaboration is also important because the subjects are not just taught by one teacher; content should be brought to a common ground too.

- The teachers developing the items have to establish first reliability of the items, do pilot testing of the items before these items be included in the final pool of items for Educ 60 and later develop an item bank.
- A review of the LET Primer whether the school is teaching all the competencies in the Primer since even in the National Licensure examination for Teacher, Professional Education is one area which acquired low performance from test takers.
- Review the level of difficulty of the items in terms of order and arrangement in the test. Many of the items did not follow a pattern of order of simpler to most difficult as they are presented in the test and the construct with DIF.
- Items which are under fitting or over fitting need to be considered for revision. Review also the items which were too difficult that no one answered them and some even skipped them and the easiest items where all the participants were highly able to answer them correctly because the difficulty of these items were far below their ability level. Revisit the theories of learning items and do further investigation of the emergence of a 0.00 chi-square and a significance level of 1.00, thus producing no error.
- Time element is test administration should be reconsidered if the test was administered the same way during the real LET.
- The mock LET instrument should be designed as following the format of the real LET like having booklets where students do the shading of their correct answer and not writing the letter of their choice. This will improve face validity of the test and consistency of the instrument.
- Feedback should be sought from the School of Education alumni who have taken the National Licensure Examination for Teachers on content and procedures. It is not a matter of asking them what came out during the test; it is a matter of soliciting feedback if the subject matter taught in the pre-service are consistent with the knowledge sought during the national test.
- Investigate the consistency of the number of items in the PLET to the items given in the national test.

- Conduct a study about Let exams by collaborating with PRC. There should be bases on LET performance because there is a big gap in between evidenced-based practice and reality. PRC and teacher education institution are rich terms of data about teacher performance in national examinations yet there are no researcher available for reference.
- Incorporate IRT in assessment courses so that pre-service teachers will also be aware of this test theory and its application not just because of LET per se but for their future assessment roles in the classroom.
- Additional research should be conducted to better understand the dynamics of student views for teaching profession and principles of teaching related topics, activities, and pedagogical approaches. Of particular importance is an understanding of the factors that are most important for female and male students since these are two areas where DIF occurred.
- Teacher's teaching profession and principle of teaching topics should intensify the use of research results of gender based studies to design and develop standards based activities that appeal to males.
- The School as a center of excellence shall venture into studying new theories to testing which are now widely used which is the Item response theory, be open to changes in measurement and assessment community. The school should reexamine consistency between content, assessment and methods (Martone and Sireci 2009).

Generally, the pre-Let instrument has good measurement properties. It is an acceptable tool but the items can just be used for each construct only and not to measure the general construct which is professional education. There are only few 'misfit items' per construct however there is an urgent need to have a committee to revisit the items as a whole test for professional education. There seems a problem with unidimensionality if all the items are taken as one test. Each construct is fine but bringing all the constructs together creates some questions on validity. It could be that some items are highly dependent on other items. Another important reminder for teacher educators and other stakeholders that a good licensure test or a pre-licensure test will not remove the need of teacher evaluation and other programs to ensure competence in the teaching field (Mehrens 1987) because there is still that question of whether licensure exams can assure schools for quality teaching (Shepard 1991). CTT continues to be an important framework for test construction. Teachers need to have a clear idea of the relations between IRT and CTT. This should improve the appreciation of both theories and facilitate communication with researchers, item writers and classroom teachers who are frequently more familiar with CTT than with IRT (Bechger et al. 2003).

This study opens new directions for further research in test development of pre-licensure exam of teacher education institutions which can be used not only for this particular school but for the TEI's in the region and even in the country. Hence, it is recommended that an assessment committee in the School of Education should be created; teachers do pilot testing of the items before these items be included in

the final pool of items for Educ 60 and later develop an item bank; review of the LET Primer; review the level of difficulty of the items in terms of order and arrangement in the test; and review of the items which were too difficult that no one answered them and some even skipped them and the easiest items where all the participants were highly able to answer them correctly because the difficulty of these items were far below their ability level.

References

Al-Sabbah, S. A., Lan, O. S., & Mey, S. C. (2010). The using of Rasch-Model in validating the Arabic version of multiple intelligence development assessment scale (MIDAS). [Report]. *International Journal of Behavioral, Cognitive, Education and Psychological Sciences, 2*(3), 152.

Alagumalai, S., & Curtis, D. D. (2005). Classical Test Theory. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 1–14). The Netherlands: Springer.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27*(5), 319–334.

Belvedere, S. L., & de Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *Journal of Clinical Epidemiology, 63*(12), 1287–1297.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J: Lawrence Erlbaum Associates Publishers.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*, 253–269.

Brew, C., Riley, P., & Walta, C. (2009). Education students and their teachers: Comparing views on participative assessment practices. *Assessment & Evaluation in Higher Education*, 1–16.

Choppin, B. (1983). *The Rasch model for item analysis*. Center for the Study of Evaluation, University of California.

DiBattista, D., & Kurzawa, L. (2011). Examination of the Quality of Multiple-Choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2).

Franchignoni, F., Ferriero, G., Giordano, A., Sartorio, F., Vercelli, S., & Brigatti, E. (2011). Psychometric properties of QuickDASH—A classical test theory and Rasch analysis study. *Manual Therapy, 16*(2), 177–182.

Fuentealba, C. (2011). The role of assessment in the student learning process. *Journal of Veterinary Medical Education, 38*(2), 157.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Instructional Topics in Educational Measurement,* Module 16. Retrieved 30 November 2011 from http://www.ncme.org/pubs/items/24.pdf.

Griffin, P. (2009). Teachers' use of assessment data *Educational assessment in the 21st century* (pp. 183–208). Springer.

Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership, 60* (5), 6–11.

Lamb, R., Annetta, L., Meldrum, J., Vallett, D., Lamb, R., Annetta, L., & Vallett, D. (2011). Measuring science interest: Rasch validation of the science interest survey. *International Journal of Science and Mathematics Education, 10*(3), 643–668.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332–1361.

McTighe, J., & Wiggins, G. (2005). *Understanding by Design* (Expanded Second Edition). Association for Supervision & Curriculum Development.

Mehrens, W. (1987). Validity issues in teacher licensure tests. *Journal of Personnel Evaluation in Education, 1*(2), 195–229.

Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy Inventory.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, QC.

Purya, B., & Nazila, A. (2011). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2*(5), 1052.

Rieg, S. A., & Wilson, B. A. (2009). An investigation of the instructional pedagogy and assessment strategies used by teacher educators in two universities within a state system of higher education. *Education, 130*(2), 277–294.

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement, 72*(1), 5–36.

Shepard, L. A. (1991). Will national tests improve student learning? *The Phi Delta Kappan, 73*(3), 232–238.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36*(10), 1138–1146.

Skelton, C. (2002). The 'feminisation of schooling' or 're-masculinising' primary education? *International Studies in Sociology of Education, 12*(1), 77–96. doi:10.1080/09620210200200084.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23–27.

Webber, C. F. E., & Lupart, J. L. E. (2012). *Leading student assessment* (Vol. 15). Dordrecht: Springer, Netherlands, Dordrecht.

Wiberg, M. (2004). Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*: Educational Measurement Solutions Melbourne.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ConQuest Version 2.0. Camberwell, Victoria: ACER Press.

Wright & Linacre, (1994) http://www.rasch-analysis.com/rasch-analysis.htm.