ASSESSING AND VALIDATING SCALE INSTRUMENT FOR MATHEMATICS TEACHERS

[1]*Edmundo C. Lopez*

Abstract

This study developed and validated scale instrument for mathematics teachers, which can be used for an initial step in recruiting competent mathematics teachers. The scale items were developed by intensive and extensive reading, and soliciting ideas from students, teachers, and school administrators about their perceptions on competent mathematics teachers. The scale items were finalized and subjected to three try-outs to establish their reliability and validity. The SPSS version 14.0 was used to analyze the gathered data. Results showed that the scale items were valid and reliable at 0.987 Cronbach's alpha; hence, this developed and validated scale instrument served its purpose.

Keywords: *development, validation, scale instrument, mathematics teachers*

1.0 Introduction

The critical stage of scale instrument development is the validation. In this stage, the dependability and usability of the scale instrument are gauged for its acceptability or rejection. Several precautions have to be considered in order to make the pre-determined scale instrument suits to what it intends to measure and the consistency it has to establish. Since development and validation of scale instrument is crucial, there is a need to adapt some procedures in attaining unbiased scale instrument. These procedures involve in establishing the reliability and validity of the scale instrument. As it is said that a scale instrument may be reliable but not valid has to be dealt with. Validity, as a unitary concept, represents all of the evidences that support the intended interpretation of the measure (Kaplan & Saccuzzo, 2001). It establishes the lucidity of the scale instrument in terms of its contents, its comparison to other scale instruments, the variation of responses among scale takers, and its construct itself.

In any educational system, scale instruments to measure effectiveness and efficiency of a competent teacher are very necessary. If the instrument is, indeed, precise and accurate, it can discriminate a teacher who demonstrates effective teaching from another teacher who is ineffective (Ochaves, 2004). An instrument that is precise and accurate includes prominent aspects of a competent mathematics teacher, which are orderly classified and analyzed for the purpose of segregating teachers either competent or incompetent based on what is manifested in the assessment using the said instrument.

In order to recruit competent mathematics teachers in the teaching force, a scale instrument for mathematics teachers has to be developed and validated. This scale instrument is an indispensable tool. The Inventory of Essential Teaching Skills, as Kozloff (2002) explained, can help: (1) assess educational school students as they move through and complete the curriculum; (2) guide the evaluation and

[1]*Surigao del Sur State University*

improvement of education school curriculum; and (3) evaluate the quality of classroom instruction.

The issue of teaching effectiveness of mathematics teachers is quite interesting; however, it may be jeopardized if there is no appropriate scale instrument to measure it. This scale instrument unfolds the characteristics of a competent mathematics teacher that can help educational managers to determine mathematics teachers who can carry out the thrust of mathematics instruction in the educational system. This scale instrument also can point out the strengths and weaknesses of the mathematics teacher, which can be the basis for more improved performance in the teaching job. With these advantages that can be brought in this scale instrument, the researcher is arduously determined to develop and validate this scale instrument for mathematics teachers.

2.0 Theoretical/Conceptual Framework

This study is anchored on generalizability theory (Ochaves, 2004) that is a behavioral measurement theory, which effect brought classical reliability theory into a stage of fuller and more coherent development. It acknowledges and identifies multiple sources of error. It is a test score to be a single sample from a universe of possible scores, and the reliability of that score is the precision with which it estimates a more generalized universe value of the score ("true score"). Its computation involves the application of analysis of variance of statistical techniques to determine the generalization, or dependability of test scores as a function of changes in the person(s) taking the test, different samples of items comprising the test, the situations under which the test is administered, and the methods or people involved in scoring the test (Aiken, 1994). Generalizability coefficient can be determined using coefficient Alpha by Cronbach, although G coefficient has its own formula.

However, the classical reliability theory (Aiken, 1994) stated that a person's observed score on a test is composed of "true" scores plus some unsystematic error of measurement. True score is defined as average of the scores that would be obtained if a person took the test an infinite number of times. It cannot measure exactly but must be estimated from the person's observed score on the test. The reliability of the test is not influenced by systematic changes in scores that affect all examinees similarly, but only by unsystematic changes that have different effects on different examinees. It tends to be higher when the variance of the variables of interest (test scores, item scores, ratings) is large than when it is small. It is noted that an instrument can be reliable but not valid. On the other hand, a valid instrument is reliable.

Furthermore, the theory of validity, as requisite for unbiased scale instrument, has to be given equal consideration with the theory of reliability in which it has to be enforced in the development and validation of scale instrument. It explicates the concept of measuring of what it tends to measure. As this scale instrument possesses the inherent characteristics of validity, content validity, construct validity, predictive validity, convergent validity, divergent validity, and known-group validity have to be established.

Based on the information apparent in the theories gathered in this study, the conduct of this study would be systematically and properly initiated and executed. As it would be finished, it would be assumed that this tool can predict future

teaching performance. To validate the scale instrument, it was correlated with the Teachers Performance Rating, Teaching IQ and Attitude Towards Movies. The output of the study is the final validated scale instrument.

3.0 Research Design and Methods

This is descriptive-development study because it develops an instrument or a scale using primary data derived from responses in survey, interviews, and questionnaires. It is directed to develop and validate a scale instrument for the inventory of the mathematics teacher's competencies. To successfully achieve the objectives of this study, procedures of instrument development and validation was followed, which were Phase I - Pooling of Items, Phase II - Content Validation, Phase III - Try-outs and Finalization of the Scale Instrument, and Phase IV – Validity Establishment.

The final scale instrument would be a 75-item questionnaire, which consisted of three components: Preparation of Teacher, Instructional Competence, and Affective Competence. Each component contained twenty–five scale items.

The sampling design used in this research was purposive sampling in order to serve the purpose of this study. Purposive sampling, according to McMillan (1992), is a judgment or judgmental sampling. It is selecting particular elements from the population that will be representative or informative about the topic. It is based on the knowledge of the researcher about the population in which samples are selected to provide the best information to address the purpose of the research.

The respondents of this endeavor were primarily elementary, secondary, and tertiary mathematics teachers of Caraga Region 13. Other respondents were regional and division mathematics supervisors, college deans who are teaching mathematics, Mathematics Department chairmen, school mathematics coordinators, school administrators, and fourth year high school students of Agay National High School.

To get the S and Q values, Edwards (1957) said that a minimum of 50 judges can be used to judge the items as having negative or positive perspective on the construct being measured. Thus, this study only used a minimum of 50 judges.

For reliability and validity test, many authors before suggested ten times the number of final test items. However, with the advent of computers and the modern psychometricians, they suggested a minimum of 50 respondents. In this study, 105 teachers were used for the reliability measure and finalization of the inventory scale, and 300 teachers for the validity establishment.

4.0 Results and Discussions

4.1 Phase I - Pooling of Scale items

Pooling of scale items was done through readings and survey from teachers and students. As a result of extensive and intensive readings and the responses of the mathematics teachers and students, 151 scale items were formulated. The scale items were classified into three components namely: Preparation of Teacher, Instructional Competencies and Affective Competencies. According to Arends (1998), effective teachers possess good personal qualities and dispositions, and are prepared with a vast of teaching practices and personally disposed to problem solving. Thus, the three classifications of the scale items were derived from Arends' theory.

There were fifty scale items contained in

the Preparation of Teacher and another fifty items for Instructional Competencies, and fifty-one scale items in Affective Competencies component, making a total of 151 scale items. These scale items were subjected to content validation.

4.2 Phase II – Content Validation of Scale Items by Experts

Content validity is the most important validity of an instrument. It measures the relevance of the scale items to the construct or variable being measured. It is to determine whether the instrument measures what it intends to measure. To obtain content validity of the instrument, the scale items were subjected to evaluation by six experts. The six experts were all mathematics major namely; the Regional Supervisor of Mathematics, Dean of the Graduate School, a Mathematics mentor of student teachers, a Director of Research, a Chairman of the Mathematics Department and a teacher in Mathematics teaching strategies.

The six evaluators judged each scale item according to its relevance to the three components of the scale instrument. A scale of 0 to 5 was used to judge each scale item, where scale of 0 means not relevant and scale of 5 means highly relevant. The evaluators indicated their comments/ suggestions on the spaces provided for as revisions of the particular item. The result of the content validation shows the number of items retained and revised based on the ratings of the evaluators in each item. The arithmetic mean of the evaluators' ratings for each item was computed. Those items with a mean score of 4.00 to 5.00 were considered as relevant items.

Under Preparation of Teacher, there were 46 items with mean scores from 4.00 - 5.00 and four items were rejected. So, 92.00 percent of the original items were retained. Twenty-seven items were revised because the statements were very long, have double concepts, and were not clearly stated.

In Instructional Competence, there were 49 items with a mean score from 4.00 to 5.00 while one item was rejected. So, 98.0 percent of the original items were retained. Thirty items were revised due to a long statement, double concept, and not clearly stated.

Furthermore, all scale items in Affective Competence had mean scores from 4.00–5.00. No item was discarded, however, twenty-four items were revised which is 47.0 percent of the original scale items. Revisions were due to a long statement, double concepts, and not clearly stated.

There were five discarded items in this process. As a result, 146 scale items were left for the first try-out following the revisions suggested by the experts.

4.3 The S-Values and Q-Values of the Scale Inventory

First Try-out of Scale Items. This try-out was undertaken in order to establish the S–values and the Q–values of the 146 scale items. The instrument is a Thurnstone scale; thus the Q-values is the basis for item discrimination while the S-values will be an item value score. To determine these values, 50 "judges" were made to evaluate each statement as to its negative or positive perception to the construct or variable being measured. The scale of 1 up to 11 was used by the judges to rate each scale item. Scale of 1 means the statement was highly negative with reference to the construct competencies of a mathematics teacher, and a scale of 11 means the item is highly positive. The ratings of the judges were the basis for the S-value and Q-value of each scale item.

All statements of the scale instrument were positively stated since the S–values were above 6. This condition does not invalidate the construct of the scale instrument. Netemeyer et al., (2003) said that negative statements do not show as high reliability as positive statements do, they do not cause confusion to respondents. However, items with Q-values of 2.0 and above were discarded from the list of scale items in the inventory scale but those Q-values below 2.0 were retained in the list, since Q-values of 2.0 and above showed wider ranges than Q-values below 2.0.

There were 39 scale items retained in Preparation of Teacher, which are 84.78 percent of the original number of scale items, 40 items in Instructional Competencies or 81.63 percent and still 51 items in the Affective Competencies or 100 percent of the original number of scale items. A total of 130 items were retained, which are 89.04 percent of the over-all number of scale items.

However, there were seven (7) scale items discarded from the Preparation of Teacher component, nine (9) scale items from the Instructional Competencies component, and none from the Affective Competencies component. A total of sixteen scale items was discarded from the inventory scale which is 10.96 percent of the overall total of scale items. As a result of the first try-out, there were 130 scale items retained. This result would show that majority of the items were relevant to the construct. These scale items were subjected to the second try-out.

4.4 Establishing reliability coefficient

Second Try-out of the Scale Items. The 130 scale items were tried out to 105 elementary and high school mathematics teachers in the schools of the Divisions of Agusan del Norte and Butuan City. This second try-out was to establish the reliability of the scale and to select the final items of the scale inventory. Reliability is a characteristic of the instrument that describes the consistency of the scores of the examinees. When the scores of the examinees are consistent then it is assumed that the test items are powerful and can measure the abilities of the examinees without constraints of time and context. The reliability measure of the instrument was established using the Cronbach Alpha with the use of the Version 14 Evaluation copy of SPSS (Statistical Package for Social Sciences). The scale inventory was administered in the Likert scale form, so as to give respondents the freedom to rate each scale item to the degree of agreement they want to. The descriptive scales used were SA for Strongly Agree, A for Agree, SWA for Somewhat Agree, SWD for Somewhat Disagree, D for Disagree, and SD for Strongly Disagree. Each descriptive scale of the questionnaire had corresponding numerical equivalent: 6 for SA, 5 for A, 4 for SWA, 3 for SWD, 2 for D, and 1 for SD.

Table 1. Reliability coefficient of the inventory scale

Scale Competencies	Number of Items	Cronbach Alpha	Cronbach Alpha based on Standardized Items	Remarks
Preparation	39	0.963	0.984	Reliable
Instruction	40	0.943	0.977	Reliable
Affective	51	0.977	0.978	Reliable
Total	**130**	**0.984**	**0.989**	**Reliable**

As shown in table 1, the Cronbach's Alpha is 0.984 and the Cronbach's Alpha based on Standardized Items was 0.989. These two reliability values mean that the instrument has a high level of reliability, which means that the newly developed inventory scale is reliable (Aiken, 1994).

The table shows further the alpha values of the inventory scale in each component. The alpha value of Preparation of Teacher was 0.963, which showed that the sub-component on Teacher Preparation was reliable. While the alpha value of the Instructional Competencies was 0.943, which showed that scale items of this component were reliable, so with the alpha value of Affective Competencies, which was 0.977. The alpha values were all within the ranges of high reliability 0.29 – 1.00.

As to finalization of the scale instrument, the number of scale items was pegged at 75, where each component will have 25 scale items. To select the best items for the final scale, the SPSS output was analyzed. Using the column "Alpha if item deleted", those that were higher than the overall alpha of 0.984 were discarded. This result simply tells that if the item is deleted, the overall alpha will increase. Hence these items should be discarded from the scale.

However, the differences of the values of the Cronbach's Alpha if Item Deleted were insignificant because almost all of the values were the same. Since there was a need to pick out twenty-five scale items in each component of the inventory scale, the values of the *Corrected Item-Total Correlation* were used. These values were ranked in each component of the inventory scale. The first twenty-five in the rank were selected as the final scale items for each component of the inventory scale.

The original number of scale items of this stage comes from the number of scale items retained in the first try-out. Twenty-five scale items retained in the Preparation of Teacher which are 64.10 percent of the original number of scale items while the twenty-five scale items retained in the Instructional Competencies are 62.20 percent of the original number of the scale items and the twenty-five scale items retained in the Affective Competencies are also 49.02 percent of the original number of scale items. The total number of scale items that are retained is 75, which is 57.69 percent of the overall number of scale items.

However, 14 scale items were discarded in the Preparation of Teacher component which were 35.90 percent of the original number of scale items; in Instructional Competencies, 15 items or 37.50 percent of the original number of scale items were discarded; and in the Affective Competencies, 26 or 50.98 percent of the original number of scale items were excluded. There was a total of 55 scale items that were discarded, which was 42.31 percent of the overall number of scale items.

The twenty five items per component of the scale are within the principle of scale items in a scale instrument that a research instrument should have a minimum of 20 items and a maximum of 25 items in order to obtain a valid representation of indicators. Too long instrument would likely have biases on administrability when a respondent does not answer the items objectively due to its length. Although the whole scale has 75 items, yet, each component could be administered separately if so desired.

4.5 *Establishing the concurrent, predictive, divergent, convergent and known-group validity*

To establish the concurrent, predictive and construct validity of the scale, the scale was administered to 300 mathematics teachers of the eight divisions of Caraga

Region 13 in the three different educational levels, elementary, secondary and college. Another three instruments were administered to be used for validation of the newly developed scale – (1) Self Administering Performance Evaluation for Teachers, (2) Teaching IQ Test, and (3) Attitude Towards Movie Scale by Thurstone (1930). Through the four sets of questionnaires answered by the respondents/subjects, the following were established: concurrent validity, predictive validity, convergent validity, divergent validity, and known-group validity. Table 2 presents the summary of the validation.

Concurrent Validity. Concurrent validity is a sizable correlation between the construct measure under development and a criterion measure collected simultaneously or "concurrently" (Netemeyer et al., 2003). When an instrument has established concurrent validity, then the test results of the examinees in the instrument to be validated has significant relationship with a similar tool of similar skills. In this study, the newly developed scale inventory was compared to the teachers' job performance for the month of April. It is assumed that teaching performance has similar content with the new instrument which also deals on the competencies of a mathematics teacher. This would further assume that the higher the scores in the inventory scale, the higher also the job performance rating of the examinees. According to Sison (1981), tests are given to supplement the interview and to determine the applicant's ability which cannot be grouped through interview. They also help make an objective comparison among applicants.

As shown in Table 2, the result reveals that the Pearson Correlation Coefficient of the scores of the respondents in the Scale Instrument for Mathematics Teachers and Teacher's Performance Evaluation is 0.391 significant at the 0.000 level (two-tailed), which confirms that concurrent validity is inherent in this newly developed scale instrument because of significant correlation. This shows that the existing scale instrument significantly correlates with the newly developed scale instrument implying that if Teacher's Performance Evaluation can measure teacher's performance, which is the existing scale instrument, then this newly developed scale instrument can also measure teaching competence of mathematics teachers.

Table 2. Summary of validity coefficients of the scale instrument

Validity	**Indicators**	**Computed Values**	**Level of Significance**	**Remarks**
Concurrent	Teaching Performance	r = 0.391	0.000	valid
Predictive Validity	Teaching Performance	t = 6.61438	0.001	valid
Convergent	Teaching IQ	r = 0.279	0.002	valid
Divergent	Attitude towards Movies	r = 0.092	0.114	valid
Known Group	Educational Levels	f = 9.974	0.000	valid

Predictive Validity. Predictive validity refers to the ability of a measure to effectively predict some subsequent and temporally ordered criterion (Netemeyer et al., 2003). When an instrument has established predictive validity, then the test results of the examinees in the instrument to be validated, has significant relationship with a similar tool of similar skills.

It is assumed that teaching performance has similar content with the new instrument which also deals on the competencies of a mathematics teacher. This would further assume that the scores in the scale inventory can predict the teachers' performance on the job.

In this study, the newly developed scale inventory was compared to the teachers' job performance for the month of April. The scores of the subjects in the newly developed scale instrument serve as the indicator to identify the upper 30 percent and lower 30 percent of the scores of the subjects in the Teachers' Performance Evaluation. These two sets of scores were compared using t-test. Result showed that the computed value for t-test was 6.61438 at 0.001 level of significance. This shows that there is significant difference in the scores of the upper 30 percent and lower 30 percent. This shows further that scale takers who score high in this newly developed scale instrument, also score high in the Teachers' Performance Evaluation and those who score low in this newly developed scale instrument, also score low in the Teachers' Performance Evaluation. In this case, mathematics teaching performance can be predicted using this inventory scale, hence predictive validity of this inventory scale was made known.

Convergent Validity. A measure is said to possess convergent validity if independent measures of the same construct converge, or are highly correlated. The evidence of convergent validity typically is provided from correlations between the new measure being developed and existing measures (Netemeyer et al., 2003).

As shown in the Pearson correlation, the newly developed Scale Instrument and Teaching IQ Test in comparison obtained 0.279 significant at the 0.002 level (two-tailed), which shows that this newly developed scale instrument when compared with another scale instrument with similar construct, there is high correlation. It shows further that the validity of this newly developed scale instrument is supplemented by another scale instrument with similar construct resulting to convergent validity.

Divergent Validity. The divergent validity assesses the degree to which two measures designed to measure similar, but conceptually different, constructs are related. A low to moderate correlation is often considered evidence of divergent validity (Netemeyer et al., 2003).

The Pearson Correlation of newly developed Scale Instrument and Attitude Towards Movie Scale was 0.092 at the 0.114 level of significance. This was not significant. This means that these two instruments differ in construct - no relation at all and no overlap of the contents in the different constructs, which implies that the newly developed scale instrument measures what it tends to measure as other scale instrument with different construct measures what it tends to measure; thus, divergent validity is ascertained. This shows further that this newly developed scale instrument is unique from other scale instrument.

Known-Group Validity. Known-group validity involves the measure's ability to distinguish reliably between groups of people that should score high on the trait and low on the trait (Netemeyer et al., 2003). Supportive evidence of known-group validity

typically is provided by significant difference in mean scores across independent samples (Tian et al., 2001 as cited in Netemeyer et al., 2003).

The F–value of the scores obtained by the three groups of mathematics teachers: Elementary, high school and college was 9.974 significant at the 0.000 level. Result shows that there is a significant difference in the mean scores of the three educational levels. The high school mathematics teachers had high mean score, followed by college mathematics teachers, then the elementary mathematics teachers. This implies that when this newly developed scale instrument is given to scale takers or group of scale takers, they have different responses depending on what characteristics of competent mathematics teachers they possess. In this result, known - group validity is established.

5.0 Conclusion

Through the findings gathered in this study, the task of developing and validating the scale instrument for mathematics teachers has succeeded to reach its objectives because it underwent content validation through inter-raters validity, computation of the S-value and Q-value using Thurstone's scale, tryouts, reliability and validity testing. With the results of reliability and validity testing as shown in the summary of findings, this scale instrument for mathematics teachers, as a product of careful and intellectual venture, is reliable and valid. Linn & Gronlund, (2000) said that validity is an evaluation of the adequacy and appropriateness of the interpretations and uses of assessment results while reliability refers to the consistency of assessment results. Reliability (consistency) of measurement is needed to obtain valid results but we can have reliability without validity. In this regard, this scale instrument for mathematics teachers is worth using because it is geared up to serve for the intended purpose of this inventory scale.

References

Aiken, L.R. (1994). *Psychological testing and assessment*. (8th ed.). USA: Allyn and Bacon.

Arends, R. (1998). *Learning to teach* (4th Ed). Singapore; The McGraw-Hill Companies, Inc.

Edwards, A.L. (1957). *Techniques of attitude scale construction***.** New York: Appleton – Century – Crofts,

Kaplan, R.M. & Saccuzzo, D.P. (2001). *Psychological testing*. (5th ed.). Australia: Thompson Learning.

Kozloff, M. A., (2002). *"Inventory of essential teaching skills"*. Watson School of Education. University of North Carolina at Wilmington, 1-7.

Linn, R. I. & Gronlund, N. E. (2000). *Measurement and assessment in teaching.* 8th ed. Upper Saddle River, New Jersey: Prentice – Hall, Inc.

McMillan, J. H. (1992). *Educational research (fundamentals for the consumer).* USA: Harper Collins Publishers.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications.

Ochaves, J. A. (2004). The question of dependability of students' ratings on college instructor effectiveness (Application of the generalizability theory on behavior measures). *Compendium of Research – Based Papers in the PAGE 1996 – 2004*: PAGE Publication on the Occasion of CHED's Asia – Pacific Conference on Research in Higher Education.

Sison, P. S. (1981). *Personnel and human resources management.* 5th ed. Manila, Philippines: Rex Printing Company, Inc.

Tian, K.T. and Mckenzie, K. (2001). "The long-term predictive validity of the consumers' need for uniqueness scale". *Journal of Consumer Psychology, 10 (3), 171-193.*

Thurstone, L.L. (1930). A scale for measuring attitude toward movies. *Journal of Educational Research, 22, 89-94.*