USING RELIABILITY MEASURES IN TEST VALIDATION

Rufina C. Rosaroso, PhD
Associate Professor, College of Arts and Sciences,
Cebu Normal University, Cebu City, Philippines

Abstract

This study aimed to assess the use of reliability measures in test validation, an essential part of test standardization. Since reliability of a test is a prerequisite to validity, the results of reliability analysis using internal consistency measures with coefficient alpha, also called the Cronbach's a*lpha*, as the measure of homogeneity of items was used. Reliability coefficients above 0.80 met the minimum acceptable standards for the proposed University of San Carlos College Admission and Placement Test (USCCAPT) in English, Science and Mathematics, respectively. Further, reliability analysis was done based on the second pilot testing scores of 262college freshman students from a university in Cebu City.

Keywords: Test validation, reliability analysis, Cronbach's alpha, reliability coefficient

Introduction

Test validation as posited by Cronbach (in Cracker & Algina, 1986) is the process in which a test developer gathered evidences to support the test scores' interpretations. He suggested that when an examiner planned to conduct a validation study, clear identification of the intended inference is required.

Test validation process may be viewed as a continuous and endless activity. While test validity may be established on the basis of past data, for every new inference that is made, fresh empirical support must be obtained. The validity of a test is thus evaluated on the strength of the accumulated supportive empirical evidence. This implies that validity is not an all-or-none property but, rather it occurs in varying degrees, just like reliability (Carlota, 1987).

Test validation involves administration and revision over and over again until acceptable levels of validity and reliability are achieved and item analysis statistics are satisfactory. The process involves instrument validation prior to its use in the intended evaluation. Test revision is normally based on the results of the actual administration to the intended students (Gay, 1991). Further, Gay (1991) added that the issue on analyzing the adequacy of a test after its administration is essential in test validation. Analysis of test results, on the other hand, may or may not support the initial validity judgments. However, the use of techniques such as item analysis, construct-related evidence, evidence based on internal structure of the test and reliability analysis utilizing internal consistency measures help identify specific ways in which test items can be improved. Thus, analysis of results provides better score/s interpretations making test items undergo validity and reliability.

In a general sense reliability is defined as the consistency of measurement (Linn and Gronlund, 2000) or the precision of measurement (Carlota, 1987). It is a measure of consistency of test scores from one measurement to another. It also describes the scores consistency obtained by the same set of test takers when given the same test on different time.

Reliability is a critical quality of any test, whether it is a written test, a performance assessment or an informal observation or question. It provides the consistency that makes validity possible (Linn and Gronlund, 2000).

As applied to testing and assessment, reliability pertains to the results obtained with an assessment instrument and not to the instrument itself. It is the reliability of the test scores that educators looked into rather than the test or the assessment. The estimate of reliability demonstrates consistency.

Procedurally, there are a number of ways by which reliability may be estimated. This can be estimated through test-retest, alternative forms, inter-rater reliability and internal consistency.

Kerlinger and Lee (2000) explained test-retest reliability as a measurement of stability over time. This is done by using the same test to the same group of examinees on two different occasions. If the same examinees who originally got the highest scores obtained the same scores, likewise the middle and the low performing examinees got the same results, the test is consistent (Oosterhof, 2001). Correlation between the first set of scores and the second set of scores is computed in test-retest reliability.

Alternate or equivalent forms involve the administration of two forms of a test to the same examinees where each student get two scores, one on each test form (Oosterhof, 2001). The test scores obtained using either of the two administrations are treated in a correlational manner and the resulting statistical index, called the reliability coefficient, is computed to check the level of reliability which the test possessed. Reliability coefficient values range from 0% to 100% for different measures. Thus, an instrument is said to be perfectly reliable if no measurement error is present. When measurement error is present, an examinee’s score would deviate from his true level of the attribute (Carlota, 1987).

Another type of reliability is inter-rater consistency which is obtained when two or more raters did independent scores on student’s performances (Linn & Gronlund, 2000). Consistency is obtained by correlating the scores from one judge to the other. In the classroom setting, inter-rater consistency is used when two or more teachers had independent scores to each student’s performance, making each student receiving two scores. The correlation coefficient is then computed between the teachers’ scores for each student’s performance.

Salvia and Ysseldyke (1998) defined internal consistency as reliability for generalizing to other test items. Internal consistency of a test is determined from a single test administration.

One measure of internal consistency is Cronbach’s alpha which allows the rater to estimate the reliability and knows the score variance and the covariances among all its components (Crocker and Algina, 1986). Further, Bryman and Cramer (1997) specified that Cronbach’s alpha essentially calculated the average of all possible split-half reliability coefficients. As a rule of thumb, reliability coefficients of 0.8 or above are within acceptable standards.

Thus, based on the foregoing literature review, reliability and validity evidence in this study were gathered from various sources as part of the process of standardizing a locally-made university admission and placement test.

Objectives of the Study

The purpose of this study was to determine the extent on how reliable the students’ test scores in the University of San Carlos College Admission and Placement Test (USCCAPT) as bases in making decisions, admission and placement of qualified freshman students to various college academic programs.

Methodology

This is a quantitative study employing reliability analysis using internal consistency measures with Cronbach’s alpha, as the measure of homogeneity of test items. Reliability analysis using internal consistency measures was computed to check whether the items that made up each area of the USCCAPT was internally consistent.

Results and Discussion

The test was done to a sample of 262 college freshmen taking different degree programs such as Education, Psychology, Nursing, Pharmacy, Engineering, Fine Arts and Hotel and Restaurant Management. The purpose of the second pilot testing was to estimate test reliability via internal consistency measures. The results of reliability analysis and item analysis are discussed below.

Reliability of the USCCAPT. As employed in the study, reliability analysis using internal consistency measures was done in reference to the question whether the items that made up each area of the test was internally consistent. As a reliability coefficient, Cronbach's alpha estimates the reliability of the scale by determining the internal consistency of the test or the average of all the correlations between each item and the total score (Fink, 2003). It shows how well a set of items (or variables) measures a single unidimensional latent construct. A reliability coefficient was generated whose value ranges from 0 and 1, with values closer to 1 indicating a more reliable measure. The standard acceptable value for reliability coefficients is 0.8 and above (Bryman and Cramer, 1997).

The reliability coefficients of the four areas of the proposed college admission and placement test and its overall reliability are shown in Table 1. The Cronbach's alpha values for the English, Science and Mathematics Proficiency Tests meet the acceptable standard for reliability coefficient, with values greater than 0.80.

The Reasoning Ability Test consisting of only 12 items falls below the acceptable standards. According to Ary (2002), one factor that affects reliability coefficient is the length of the test, that is, the longer the test, the greater the reliability. This could be one factor that explains the low reliability of the Reasoning Ability Test.

Table 1 Reliability Coefficients of the USCCAPT

Areas of the USCCAPT	**No. of Items**	**Cronbach's *Alpha***	**Interpretation**	**No. of Items to be Revised**
English	114	0.912	high	17
Reasoning	12	0.666	low	3
Science	41	0.872	high	12
Mathematics	37	0.902	high	6
Overall Test	**204**	**0.968**	**high**	**38**

Hatcher (in Santos, 1999) explained that Cronbach's *alpha* is an index of reliability associated with the variation accounted for by the true score of the "underlying construct." The larger the overall alpha coefficient, the more likely the items contribute to a reliable scale. From the results in Table 1, English Proficiency Test has the highest reliability index, followed by Mathematics and Science, respectively. This implies that there is a degree of internal consistency among the items in the three tests. Further, the overall test reliability supports these findings.

Item Analysis of the Revised USCCAPT.

After performing the reliability analysis of the test, item analysis was performed for the second time to determine the degree of difficulty and discrimination of each item. The results of item analysis can be used to select items of desired difficulty that best discriminate between high and low achieving students. Moreover, these can be useful in identifying faulty items and can provide information about students' misconceptions and topics that need additional work (Linn & Gronlund, 2000).

Table 2a presents the distribution of the items by area of the revised proposed college and admission test based on the difficulty indices in the second item analysis.

Table 2a Distribution of Items by Areas of the Revised USCCAPT
According to the Degree of Difficulty

	Areas of the Test							
	English Proficiency		**Mathematics**		**Science**		**Reasoning**	
Degree of Difficulty	**No. of Items**	**%**	**No. of Items**	**%**	**No. of Items**	**%**	**No. of Items**	**%**
Very Difficult	2	1.53	-	-	-	-	1	6.67
Difficult	13	9.92	2	4.65	5	9.43	1	6.67
Moderately Difficult	104	79.39	41	95.35	47	88.68	12	80.0
Easy	3	2.29	-	-	1	1.89	1	6.67
Very Easy	9	6.87	-	-	-	-	-	-
Total	**131**	**100**	**43**	**100**	**53**	**100**	**15**	**100**

As revealed in the table, most of the tests are of moderate difficulty indicating an ideal distribution as Linn and Gronlund, (2000) suggest. Further, the results show that the students' difficulty level is indicative of their English, Mathematics, Science and Reasoning performance. The English Proficiency Test still contains *Very Easy* items (6.87%) which are subject for deletion. Table 2b presents the degree of discrimination based on the second item analysis of the revised test.

Table 2b Distribution of Items by Areas of the Revised USCCAPT
According to the Degree of Discrimination

Degree of Discri-mination	***Areas of the Test***							
	English Proficiency		**Mathematics**		**Science**		**Reasoning**	
	No. of Items	**%**	**No. of Items**	**%**	**No. of Items**	**%**	**No. of Items**	**%**
Very Poor	37	28.25	2	4.65	2	3.77	3	20.00
Poor	19	14.50	3	6.98	10	18.87	-	-
Moderate	29	22.90	7	16.28	14	26.42	2	13.33
Good	22	16.03	8	18.60	10	18.87	4	26.67
Very Good	24	18.32	23	53.49	17	32.07	6	40.00
Total	**131**	**100**	**43**	**100**	**53**	**100**	**15**	**100**

As shown above, based on the discrimination indices of the items, the Mathematics sub-test, among the four sub-tests, has the highest cumulative percentage of items with *Moderate*, *Good* and *Very Good* discrimination indices (88.37%). This indicates that the Mathematics sub-test items may be able to discriminate well among high performing college applicants to academic programs that require more mathematical proficiency than others, such as Engineering, Accountancy, and others. Science and Reasoning sub-tests have also high percentage of items with *Moderate*, *Good* and *Very Good* discrimination indices. This implies that the items in these tests are also able to differentiate or distinguish between more knowledgeable and less knowledgeable students in these fields. According to Frisbie (cited in MacDonald, 2002), highly discriminating items contribute substantially to test score reliability for they distinguish between students of different achievement levels.

On the other hand, the English Proficiency Test has the most number of items with *Poor* and *Very Poor* (42.75%) discrimination power and thus, these items are subject for deletion or revision. One reason for this could be the length of the test which helps in discriminating or differentiating the students' perceived knowledge and abilities (Ary, 2002).

Description of the Proposed USCCAPT After the Second Pilot Testing.

The second item analysis results of the four areas of the Proposed USCCAPT per item had resulted to screening of items to be retained, revised or discarded.

Table 3 shows the distribution of the revised, discarded and retained Proposed College Admission and Placement Test items per area.

Table 3 Distribution of the Proposed USCCAPT Items After the Second Pilot Test

Areas of the Proposed USCCAPT	**Sub-areas**	**No. of Items to be Revised**	**No. of Items to be Discarded**	**No. of Retained Items**	**Percent of Retained Items**
English Language Proficiency	Spelling	4	0	9	6.87
	Finding Errors	3	1	11	8.40
	Vocabulary	10	6	17	12.98
	Grammar	12	7	28	21.37
	Reading Comprehension	8	7	8	6.11
Sub-total		**37**	**21**	**73**	**55.73**
Mathematics	Arithmetic and Measurement	0	0	9	20.93
	Algebra and Trigonometry	1	1	26	60.47
	Geometry	2	1	3	6.98
Sub-total		**3**	**2**	**38**	**88.37**
Science	Biology	3	2	15	28.30
	Chemistry	5	1	12	22.64
	Physics	2	1	12	22.64
Sub-total		**10**	**4**	**39**	**73.58**
Reasoning	Verbal Reasoning	0	1	7	46.67
	Visual-Spatial Reasoning	3	0	4	26.67
Sub-total		**3**	**1**	**11**	**73.33**
Total		**53**	**28**	**161**	**100.00**

As shown, 55.73% of the items in English Proficiency Test are retained. On the other hand, 37 (28.24%) of the items need to be improved and 4 (16.03%) of them had to be discarded from the item pool.

In the Mathematics Proficiency Test, 88.37% of the items are retained. Only few items were subject to revision while others had to be discarded. This indicates the acceptability of the Mathematics Proficiency Test containing good items in it as the results of the second item analysis revealed.

Further, Table 3 presents that 73.58% of the items in Science Proficiency Test are retained. Based from the results, 10 (18.87%) of the items need improvement while only 4 (7.55%) of them are discarded from the item pool.

In the Reasoning Ability Test, 11 (73.33%) of the items are retained. Only few items need to be revised and discarded indicating that the Reasoning Ability Test has good items.

The results of the second pilot testing of the Proposed College Admission and Placement Test provided a new profile of improved items of the test. These would help and guide the project committee members (item developers and consultants) in revising, editing and reconstructing new items. Further, an engagement in a thorough and critical review of the test parameters including continuous conceptualization, development and validation of items would be done and participated by test project committee member/s for standardization purposes.

Findings

The second pilot test provided the reliability results of the proposed University of San Carlos College Admission and Placement Test (USCCAPT) using internal consistency measures.

For the results of reliability analysis, the three areas of the proposed USCCAPT yielded the following Cronbach's alpha: 0.91 for English Proficiency Test, 0.87 for Science Proficiency Test, and 0.90 for Mathematics Proficiency Test with an overall test reliability of 0.97. Based on the acceptable reliability coefficient of 0.70 for internal consistency measures, the English Proficiency met the acceptable standard for reliability followed by Mathematics and Science Sub-tests, respectively. This implies that there is a degree of internal consistency among the items in the three tests which measure a single construct.

Conclusions

The proposed test is an admission and placement test with an acceptable reliability in the three areas; namely: English, Mathematics and Science Proficiency Tests.

Recommendations

For the first two years of implementation, the College Admission and Placement Test will be administered along with the standardized IQ test and the data obtained from these tests will be studied for concurrent validation purposes.

Continuous validation and item analyses of samples of scores obtained from the college admission and placement test will be done within the first few years of implementation. Further validation of results by colleges will be done to explore the differences in the identified factors predicting academic performance to be known. An office or group of researchers will take charge in the continuous review and evaluation of the college admission and placement test items for improvement. An item pool for the college admission and placement test will be developed to provide alternative items for those that need to be replaced.

Revision of the items are to be taken by the college admission and placement test committee project members and department representative/s.

References:

Anastasi, A. (1988). Psychological testing. New York: Macmillan Publishing Company.

Ary, D., Jacobs, L.C. & Razavieh, A. (2002). *Introduction to research in education.* USA: Wadsworth/Thomson Learning

Bryman, A. & Cramer, D. (1997). *Quantitative data analysis with SPSS for windows: A guide for social scientists.* London: Routledge.

Carlota, A. J. (1987). *Psychological measurement in the Philippines: A book of readings.* Quezon City: UP Psychology Foundation.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Fink, A. (2003). *The survey handbook.* 2nd ed. Thousand Oaks, California: Sage Publications, Inc.

Gay, L. R. (1991). *Educational evaluation and measurement: competencies for analysis and application.* Florida: Merrill, an Imprint of Macmillan Publishing Company.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling.* Cary, NC: SAS Institute.

Linn, R. L. & Gronlund, N.E. (2000). *Measurement and assessment in teaching.* 8th ed. New Jersey: Prentice Hall.

McDonald, M. E. (2002*). Systematic assessment of learning outcomes: Developing multiple choice exams.* Boston: Jones and Bartlett Publishers.

Salvia, J. & Ysseldyke, J.E. (1998). *Assessment.* Boston: Houghton Mcfflin Co.

Santos, R. A. 1999. Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of Extension*. 37 (2), 1-5. Retrieved December 10, 2014 from http://joe.org/joe/1999april/tt3.html