IS PILOT TESTING AN INDICATOR OF STUDENTS' PERFORMANCE?

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

Abstract

This study assessed the average or typical performance of selected fourth year high school students in a proposed University of San Carlos College Admission and Placement Test (USCCAPT) in Cebu City, Philippines. The college admission test is an aptitude test which is a locally-made admission and placement test composed of 259 items for all college freshman applicants. Item development was done by the different department representatives from various departments, namely; English, Mathematics, Biology, Chemistry, Physics and Psychology. Content-related validity evidence was initially established through the development of test's Tables of Specifications and a systematic review of its content and objectives based on experts' judgment. Pilot testing was administered to selected high school students who served as samples of the study. Computation of students' mean scores per school and their quartile ranks were used as methods of data analysis. Further, item analysis was utilized to assess the degree of difficulty, discrimination and effectiveness of the constructed items.

Keywords: Admission and placement test, item analysis, mean scores, quartile ranks

Introduction

The need for interdisciplinary research and research-based development of tests and other measures is highly recognized in various research agenda in higher education. Ochave (2004) underscored the importance of instrumentation or the development and validation of tests and other research instruments as a priority area in Philippine educational research.

For practitioners, test development is preparing an assessment or other individual measures for administration. It involves planning, assembling, writing, editing and administering of test items. Test administration, on the other hand, is aimed to assess students' understanding on the language used, concepts learned, time allotment and results once the measure is over. Results are revelations of students' performance whether or not they got the item right or not. Further, students' performance on administered tests serves as a significant indicator to check whether or not the items are fitted to their ability levels and desired competencies.

It is in this context that this study was undertaken to gather the validity evidence of this proposed locally-made admission and placement test for college applicants using pilot groups of Filipino fourth year high school students. This study adopted the modern unified view of validity which regarded validity as not simply a property of a test but the extent to which the inferences and decisions made based on test scores were appropriate and justified. Validity was established by using item analysis and quartile ranking to measure the pilot group's performance on the proposed college admission and placement test.

Objectives of the Study

This study aimed to assess the average or typical performance of selected fourth year high school students in the various components of the proposed college admission test, namely; English Language Proficiency, Mathematics, Science and Reasoning Ability.

Methodology

This is a quantitative research employing computation of the students' mean scores and their quartile ranking. Further, item analysis was utilized to measure the content validity of the proposed college admission and placement test.

Results and Discussion

The need to conduct a pilot test of any test is essential and serves several purposes. First, it gives an opportunity to try out the items well before the test is finalized. Pilot testing, as pointed out by Fink (2003), is a process of simulating the use of the test in its intended setting. It helps in error identification needed for item redesigning and predicts possible problems that will be encountered in its administration (Litwin, 2003).

In this study, pilot testing was conducted in five high schools in Metro Cebu. The five high schools included a non-sectarian laboratory school of a state university in Cebu City; a public science high school; a public high school; and two private sectarian schools. The scores obtained by the students in the pilot test in these five selected high schools were then analyzed by subject areas to provide a picture of students' performance.

Table 1 presents the students' performance by school in the four subject areas. The English Proficiency test was not administered to students in schools 3 and 5 due to the limited time given by the principals for pilot test administration.

Table 1 Mean Scores and Rank of the Five Schools by Areas of USCCAPT

School Code	*English*		*Mathematics*		*Reasoning*		*Science*	
	Mean Score	**Rank**	**Mean Score**	**Rank**	**Mean Score**	**Rank**	**Mean Score**	**Rank**
1	74.80	1	50.93	3	59.26	1	51.51	2
2	71.51	2	57.11	1	56.44	3	53.95	1
3	-	-	36.70	5	34.17	5	39.79	5
4	70.25	3	52.39	2	57.30	2	44.01	3
5	-	-	41.84	4	51.05	4	39.96	4

The typical or average performance of the students in these subtests is expressed in terms of mean percent correct. As shown in Table 1, performance in English is relatively higher with mean scores ranging from 70.25 to 74.80. Among the five schools pilot tested, the students from the public science high school rank first in Mathematics and Science Tests, second in English and third in Reasoning. The students from the laboratory school perform better than the others as manifested by their mean scores. Further, the students' mean scores from a private high school indicate good performance in all tests. The students of a second private school rank second in both Mathematics and Reasoning and third in both English and Science. On the other hand, students from a public high school rank lowest in terms of performance in all tests.

To provide an overview of the distribution of the students' scores in the various subtests, quartile ranking was used to divide the frequency distribution into equal fourths (Kaplan & Saccuzzo, 2001). The fourth quartile is the upper 25%, followed by the third quartile, then second quartile and first quartile, the lower 25%.

Table 2 shows the distribution of students based on quartile ranking by subtests and schools. The data provide relevant information on how well high school students from different types of high schools perform in the subtests of the college admission and placement test.

Table 2
Performance of Students in the Different Areas of USCCAPT by Schools

Subtests	***School Code***	***Quartile Ranks***								***Total***
		4th Quartile (Upper 25%)		**3rd Quartile**		**2nd Quartile**		**1st Quartile (Lower 25%)**		
		n	**%**	**N**	**%**	**N**	**%**	**n**	**%**	
English	1	11	26.83	10	24.39	10	24.39	10	24.39	41
	2	8	25.81	8	25.81	7	22.58	8	25.81	31
	4	11	25.58	10	23.26	11	25.58	11	25.58	43
Math	1	18	40.00	11	24.40	11	24.40	5	11.11	45
	2	19	**63.33**	8	26.67	3	10.00	-	-	30
	3	3	5.56	5	9.26	14	25.92	32	**59.26**	54
	4	12	27.27	24	54.55	7	15.91	1	2.27	44
	5	1	2.63	12	31.58	11	28.95	14	36.84	38
Reasoning	1	9	20.00	19	42.22	5	11.11	12	26.67	45
	2	10	**33.33**	4	13.33	9	30.00	7	23.33	30
	3	18	32.14	11	19.64	15	26.79	12	21.43	56
	4	12	28.57	10	23.81	10	23.81	10	23.81	42
	5	7	18.42	15	**39.47**	7	18.42	9	23.68	38
Science	1	12	26.67	10	22.22	11	24.44	10	22.22	45
	2	4	21.05	7	**36.84**	3	15.79	5	26.32	19
	3	12	22.22	19	35.19	8	14.81	15	27.78	54
	4	10	23.81	11	26.19	12	28.57	9	21.43	42
	5	10	26.32	7	18.42	12	31.58	9	23.68	38

Table 2 shows that for the English Proficiency test, there is an almost equal distribution of students per quartile in the three schools, indicating that the performance of the students in the English Test is almost comparable across the three schools. Moreover, for the Mathematics Proficiency Test, students from the public science high school (school 2) performed much better than the others with 63.33% belonging to the upper quartile and none of them belong to the lower quartile. On the contrary, it is notable that most students from the public high school (school 3) performed very poorly with 32 (59.26%) of them belonging to the 4th or lower quartile. In Reasoning, students from a state university's laboratory school got the highest percentage, (62.22%) in the upper two quartiles and they also got the highest mean score in this subtest. For Science, the results showed an almost equal distribution of students in both upper and lower quartiles. The students from the laboratory school have the highest percentage in the upper quartile, followed closely by students in a private high school. Although the mean scores of the students from the public science high school is the highest among the group, only 4 (21.05%) of these students belong to the upper 25% while the 36.84% (7) of these students belong to the third quartile. This indicates the presence of extremely high scores in the group which could have pulled up the mean score.

Aside from a description of the students' performance in the various areas of the test provided in the pilot test results, pilot testing has also served in this study as a means to screen the items in the test and to serve the purpose of initially testing the items for validity. Natemeyer, et.al., (2003) explained that pilot testing is done to reduce the number of items in an initial pool to a more manageable number by deleting items that do not meet certain psychometric criteria.

Item Analysis Results Based on the Pilot Testing. Item analysis is the process of evaluating the effectiveness of items in a test by exploring the examinees' responses to each item (Ho Kim, 1999). It gives information on difficulty, discriminating power of the item/s (Calmorin, 1994) which serves as basis for retaining, deleting or revising them. In this study, item analysis procedure was done for all content areas of the proposed test. The degree of difficulty based from the first item analysis results is shown in Table 3a.

Table 3a
Distribution of the University of San Carlos College Admission and Placement Test (USCCAPT) Items According to the Degree of Difficulty by Content Areas

	Content Areas							
	English Proficiency		**Mathematics**		**Science**		**Reasoning**	
Degree of Difficulty	**No. of Items**	**%**	**No. of Items**	**%**	**No. of Items**	**%**	**No. of Items**	**%**
Very Difficult	3	2.42	2	3.32	6	10.17	1	6.67
Difficult	6	4.84	7	11.67	5	8.47	2	13.33
Moderately Difficult	51	41.13	40	66.67	40	67.80	8	53.33
Easy	38	30.64	10	16.67	8	13.56	4	26.67
Very Easy	26	20.97	1	1.67	-	-	-	-
Total	124	100	60	100	59	100	15	100

As to the level of difficulty, these results show that most (roughly 52%) of the items of the English Proficiency Test range from *Very Easy* to *Easy*. *Difficult* to *Very Difficult* items account for around 7% of the items while *Moderately Difficult* items account for roughly 41%. Educational testing literature suggests that in an ideal test, around 70% of the items should be *Moderately Difficult* and the remaining 30% should be distributed more or less equally between the two extremes (Linn and Gronlund, 2000). These results suggest the need to revise the *Very Easy* and Easy items.

On the other hand, the distribution of Mathematics Test items with regard to the degree of difficulty is within the recommended or ideal distribution as 66.67% of the items are *Moderately Difficult*, which is not a far deviation from the ideal 70%. The percentage of *Difficult* and *Easy* items is also within acceptable range.

Further, the distribution of the Reasoning Test shows that 53.33% of the items are *Moderately Difficult.* There are no *Very Easy* items indicating that very few needs improvement.

The degree of difficulty of the Science Proficiency Test reveals that most of the items are *Moderately Difficult* (67.80%) and the actual distribution of items is not far from the ideal *70%-Moderately Difficult* rule of thumb.

Table 3b presents the level of discrimination of the University of San Carlos College Admission and Placement Test (USCCAPT) based on the first item analysis.

Table 3b
Distribution of the USCCAPT Items According to the Degree of Discrimination by Content Areas

Degree of Discri-mination	***Content Areas***							
	English Proficiency		**Mathematics**		**Science**		**Reasoning**	
	No. of Items	**%**	**No. of Items**	**%**	**No. of Items**	**%**	**No. of Items**	**%**
Very Poor	47	37.90	10	16.67	8	13.56	6	40.00
Poor	28	22.58	12	20.00	9	15.25	3	20.00
Moderate	25	20.16	19	31.66	14	23.73	3	20.00
Good	17	13.71	9	15.00	8	13.56	3	20.00
Very Good	7	5.65	10	16.67	20	33.90	-	-
Total	124	100	60	100	59	100	15	100

The discrimination power of the English Proficiency Test reflects the item’s ability to distinguish between those who know and those who do not know, from among the test takers, or that is, between the high performing and the low performing students in the group. The results reveal that there are more *Very Poor* items that need to be discarded from the item pool. Only few *Very Good* items need to be retained. *Very Good* items are items that discriminate highly from the item pool and have discrimination indices of 0.40 and above. *Very Poor* items are items that need to be rejected or improved by revision whose values are below 0.20.

In the English Proficiency Test, there were 26 V*ery Easy* and 47 V*ery Poor* items which were discarded from the item pool. These were presented to the item developers during a committee meeting for reconstruction of new items.

As shown, the discrimination power of the Mathematics Test items reveals that there are more items that *Need Improvement* (31.66%). These items are subjected for revision so as to improve the manner in which either the stems or the alternatives in the multiple-choice format are stated. Items which are either *Easy* or *Very Easy* and have low or poor discrimination power were also discarded from the item pool. On the contrary, there are more good items compared to poor ones.

In terms of discrimination power, 60% of the Reasoning Ability Test items are between *Good* to *Very Good.* Further, there are no *Very Poor* items indicating that most of the items in the Reasoning Subtest are ideal and very few needs revision.

Moreover, the Science Test discrimination power reveals that there are more items that need improvement. Almost half of the items have discrimination power ranging from *Very Poor* to *Poor*, indicating the need for some items to be discarded, revised and improved. Again, these items were presented to item developers for improvement, revision and reconstruction.

These item analysis results obtained by pilot testing or pretesting the proposed college admission and placement test provided the test item developers a chance to correct possible errors identified in the test and predicted respondents' difficulties that might arise during instrument administration. Identification of these potential impediments in advance will eventually lead in assessing its implications to testing (Litwin, 2003).

Description of the Revised University of San Carlos College and Admission Test (USCCAPT) after the First Pilot Testing.

After item analysis of the results of the first pilot test, items with *Easy* and *Very Easy* degree of difficulty and those with *Poor* and *Very Poor* discriminating power items were discarded from the item pool. Results of the first item analysis were presented to the item developers during a committee meeting for reconstruction and editing.

Table 4 presents the distribution of the revised items (subtests, number of items and time allotment) of the proposed college and admission test intended for second pilot testing for freshman college students.

Table 4
Distribution of Revised Items of the USCCAPT after the 1st Pilot Test

Content Area	Sub-areas	No. of Items	Time Allotment
English Language Proficiency	Spelling	13	1 hour and 10 minutes
	Finding Errors	15	
	Vocabulary	33	
	Grammar	47	
	Reading Comprehension	23	
Mathematics	Arithmetic and Measurement	8	50 minutes
	Algebra and Trigonometry	29	
	Geometry	6	
Science	Biology	20	50 minutes
	Chemistry	18	
	Physics	15	
Reasoning	Logical Reasoning	8	10 minutes
	Visual Spatial Reasoning	7	
TOTAL		242	3 hours

The same four subject areas are covered in the revised college and admission test intended for the second pilot testing. However, the section on Probability and Statistics under the Mathematics subtest was deleted since almost all respondents did not answer them correctly. From the students' feedback, it was found out that this area on Probability and Statistics is either not in their Mathematics curriculum or is taken last during the academic year and thus, the topic was not yet taken up when the test was administered. Except for Science High Schools, basic concepts of Probability and Statistics are included as part of the Mathematics curriculum and not as a stand-alone elective subject. Likewise, the items on Statistical Reasoning in the Reasoning Ability subtest were deleted.

Findings

The typical or average performance of selected fourth year high school students comprising the pilot group revealed that students from the public science high school ranked first both in Mathematics and Science Tests, second in English and third in Reasoning. Further, students from a laboratory school performed better than the others as manifested by their mean scores. They ranked first both in English and Reasoning, second in Science and third in Mathematics. On the other hand, students from a public school ranked lowest in all tests. Students from an urban private school indicated good performance in all tests where they ranked second in Mathematics and Reasoning, third in both English and Science Tests. On the contrary, students from a rural private school ranked second to the lowest among the five pilot tested schools.

Conclusion

The typical or average performance of selected high school students as pilot group for the proposed college admission test utilizing the initial pilot testing is dependent on the type of school.

Recommendations

In as much as any test or assessment tool has its own implicit assumptions, limits of applicability and potential hazards of misinterpretation, the following recommendations are put forth for future courses of action:

1. Administration of the second set of items should be conducted for freshman students for validation purposes.
2. Further studies need to be conducted for second item analysis which is a requisite for reliability, validity and analyses of psychometric properties of the test. An item pool for the proposed college admission and placement test will be developed to provide alternative items for those that need to be replaced.
3. An office or group of researchers will take charge in the continuous review and evaluation of the test items for improvement.

References:

Calmorin, L. P. 1994. *Educational research measurement and evaluation*. 2nd ed. Mandaluyong City: National Book Store.

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

Ho Kim, S. H. A computer program for classical item analysis. Paper Presented at the University of Georgia, June 1999:1-3. Retrieved January 15, 2015 from shkim.myweb.uga.edu/epm99.htm

Kaplan, R. M. & Saccuzzo, D. P. 2001. *Psychological testing: Principles, applications and issues.* 5th ed. Australia: Thomson Learning.

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

Litwin, M. S. 2003. *How to assess and interpret survey psychometrics.* 2nd ed. Thousand Oaks, California: Sage Publications, Inc.

Natemeyer, R.G., Bearden W.O. & Sharma, S. 2003. *Scaling procedures: Issues and applications.* Thousand Oaks, California: Sage Publications, Inc.

Ochave, J. A. (2004). Research program for graduate students in education: A prototype for other behavioral disciplines. *The PAGE Journal-Special Issue*. 108-116.