A CROSS-COMPARISON OF PAPER AND PENCIL TESTS AND HANDS-ON TASKS RESULTS IN MATH AND SCIENCE

Ali Reza Kiamanesh
University for Teacher Education

Abstract

Hands-on performance assessment tasks were used for the first time at large scale in the Third International Mathematics and Science Study (a comparative study which was carried out by the International Association for Evaluation of Educational Achievement) with the purpose of assessing certain complicated mental processes such as problem–solving, hypothesis forming, predicting, hypothesis testing , generalizing, and making conclusions. These tasks were administered along with traditional paper and pencil tests in two fields, i. e., science and mathematics. The testees of this study were two target populations: 9–year–old students from 9 countries and 13–year–old students from 19 participant countries. In this article, the findings of the study were examined from different points of view and a cross–comparison has been made between the performance of students in paper–and– pencil tests and hands–on tasks in different domains. According to the findings, the eighth and fourth grade Iranian students' average scores in "scientific problem solving and applying concept knowledge" category of hands–on tasks were higher than the average score of the other participant countries in science. Further, fourth grade Iranian students' average score on items measuring "problem solving and mathematical reasoning" was higher than the international average in math. Moreover, the performance of Iranian students in both grades revealed that they did not perform well in "interpreting investigational data". In addition, no statistically meaningful difference was observed between the Iranian girl and boy students' performance. Furthermore, the high positive correlation between the data gained from paper–and–pencil tests and hands–on tasks and also the high amount of common–variance between the two instruments are indicative of the fact that through the use of these two different types of instruments similar information is gained regarding the students' capabilities.

Formal testing in Iran`s educational system has mainly been carried out through the use of two types of teacher–made tests: selected response items and constructed response items. The former, i. e., the multiple–choice achievement tests used in large–scale testing, measure the students` factual knowledge with regard to different subject matters and also their comprehension ability. In fact, these items reveal the testees` success in undertaking a couple of tasks which do not require complicated mental processes on the part of the testees. Hence, in these items, little attention is paid to more demanding tasks such as problem–solving ones. The latter, i.e. the constructed response items, are apparently effective tools for measuring more complex and higher–level learning outcomes. Of course, great care should be exercised to measure the desired behavior effectively. However, the use of essay–type questions dose not guarantee the successful measurement of the intended objectives since the prerequisite to obtaining reliable pieces of information through these items is investing a lot of time, money, and energy. In other words, the scoring procedure of these items is quite time–consuming. The unreliability of the obtained test scores and also the sampling problem are from among the other shortcomings which paved the way for using objective tests and at most the restricted response essay–type questions.

No matter what the objective and restricted essay–type items measure, they are not really effective tools for measuring the testees' abilities in handling problem–solving tasks. Recent studies on students' achievement and schooling in many countries including Iran, indicated that the tests used at schools do not 1) properly cover the content of the textbooks, 2) properly emphasize measuring higher–order abilities, and 3) provide specific information on what students actually know and are able to do.

In order to effectively measure the testees' abilities to undertake complex problem–solving tasks, some pieces of information regarding the process skills and also how and when to apply knowledge and process are required. In fact, if interest lies in knowing how well and effective students use their prior knowledge in problem–solving situations, then alternative technologies for assessing achievement are needed.

[DOR: 20.1001.1.25382640.2000.7.1.5.1]

All achievement test items, including multiple-choice and free response ones, assess students' performance but "performance assessment" is the technical term most often used in the literature for assessment tasks in which students are required to carry out hands–on activities with equipment to show how well they are able to apply strategies and procedures to investigate and solve problems in practical settings. Other equivalent terms used in the literature include "alternative assessment", "practical assessment", and "authentic assessment" (Garden, 1997, p.3).

Even though performance assessment is not a new technology in educational contexts, "the assessment of actual skills and abilities that we want students to learn *and master has until very recently been limited to* informal classroom assessments by teachers" (Bond, 1995, P. 21).

Performance assessment tests have been divided into three major categories by Ryans and Frederiksen (1951). More specifically, they have been classified into recognition, simulation, and work sample tests. One of the most important characteristics of having an ability is being able to transfer that ability and this property is mainly measured through the use of the third type of performance assessment tasks, i.e., work sample ones. In these tests, the students are engaged in actual job tasks under controlled conditions and the results of these high quality assessments can be indicative of what students can do in a broad knowledge or skill domain. Further, the skills that students exhibit in these assessment situations should be transferred to other solutions and problems as well (Herman, Aschbacher, & Winters, 1992).

Performance assessment tasks or hands-on problem solving, "are integrated, practical tasks that involve instruments and experiments and are designed to generate information that is not easily assessed via paper–and–pencil tests. Such information enables us to better understand students' cognitive process and problem solving strategies" (Robitaille et al., 1997, p.49).

Purpose of study

The first aim of this study was to analyze and interpret the results of the Iranian students' performance in hands-on assessment tasks in mathematics and science , and also compare the obtained results with their scores on mathematics and science paper and pencil achievement tests. The second aim was to compare the Iranian students' performance on achievement tests and hands-on performance assessment tasks with their international counterparts. The required data were obtained by hands – on tasks and paper and pencil tests which were administered as a part of the Third International Mathematics and Science Study (TIMSS) conducted by the International Association for the Evaluation of Educational Achievement (IEA).

Sample design

Following the TIMSS sampling manual, a two–stage stratified cluster sample design was used for two different populations named population 1 (all the students enrolled in the two adjacent grades-- fourth and third grades--that comprised the largest population of 9-year olds at the time of testing) and population 2 (all students enrolled in the two adjacent grades--eighth and seventh grades--that comprised the largest population of 13-year olds at the time of testing). More specifically, the first stage involved choosing samples of schools and the second stage involved selecting samples of classrooms from each target grade in sampled schools (Martin, & Kelly,1996; Harmon, et al., 1997). In the first stage, 191 junior high schools (guidance schools) and 180 primary schools were selected. In the second stage, for population 2, at each school one class of eighth graders and one class of the seventh graders were randomly selected, and for population 1, the same procedure was replicated for the population of the fourth and third graders. This approach yielded a representative sample of 3,694 eighth and 3,385 fourth graders, respectively. All the sample members took the mathematics and science achievement paper and pencil tests which consisted of multiple choice, short answer, and selected response items.

The sample of schools and students for the performance assessment tasks was a subsample of the schools and students that participated in the main written (paper and pencil) achievement tests. In order to choose the sample the following steps were undertaken: From among the previously selected schools for the written assessment tests, 50 schools were randomly selected for each population, and within each school from among the already selected students, two samples of 9 fourth and eighth graders were randomly selected. This approach yielded a representative sample of 436 and 440 students in the eighth and fourth grades, respectively (Harmon, et al., 1997).

Test design

Students in population 1 took the main achievement tests which totally contained 199 items (102 items in mathematics and 97 items in science). Students in population 2 also took the main achievement tests that consisted of 286 items (151 items in mathematics and 135 items in science).

The performance assessment tasks comprised 12 tasks which required mathematics and science knowledge and performance skills on the part of the testees. The 12 tasks administered at each school were presented at 9 different stations. The assignment of the tasks to the stations resulted in three stations with one short science and one short mathematics task each, two stations with one long science task each, two stations with one long mathematics task each, and two stations with one combined science–mathematics task each. Each student visited three stations according to a rotational plan. Since each station required 30 minutes working time and since the total testing time was 90 minutes. each student had time to visit three of the 9 stations. Clusters of 9 students from each school took part simultaneously in the tasks and changed stations at 30 minutes interval so that each student attempted three, four, or five tasks (Martin & Kelly, 1996). The allocation of stations to students was also random (see Table 1). Therefore, each task was attempted by approximately equal number of randomly selected students. Almost 150 students responded to each performance task in each population.

Data analysis

A: By tasks

The average percentage scores overall on six mathematics tasks (including plasticine) in the eighth grade showed that the percent of the Iranian eighth graders' correct response was lower than the international means (54 vs. 59). From among the above -mentioned tasks, the average performance of Iranian students in the plasticine task was greater than 'the international mean (81 vs. 60), and in the packaging task, Iran's mean was close to the international one (43 vs. 44).

The average percentage scores overall on six mathematics tasks in the fourth grade revealed that the percent of the Iranian students' correct responses was greater than the international mean (40 vs. 36). Moreover, in the plasticine task, the Iranian students performed the best in comparison to their counterparts in other participant countries (63 vs. 37). The same case holds true for folding and cutting (50 vs. 38) and packaging (34 vs. 17) tasks as well.

The average percentage scores overall on six

Table 1: Assignment of Tasks to Stations and Students to Stations

Students' Sequence Number	Students (Rotation 1)	Tasks*
1	A, B,C	S_1, M_1 , S_2 , M_2 , SM_1
2	B, E,D	S_2 , M_2 ,S_4 , S_3 ,M_3
3	C,F, E	SM_1 , M_5 , S_4
4	D, G, H	S_3 , M_3 , S_5 or S_6 , M_4
5	E,A, G	S_4 , S_1 , M_1 , S_5 or S_6
6	F,H, B	M_5 ,M_4 , S_2 , M_2
7	G, I, F	S_5 or S_6 , SM_2 , M_5
8	H, C, I	M_4 , SM_1 , SM_2
9	I, D, A	SM_2 , S_3 , M_3 , S_1 , M_1 ,

Notes:
S_1=Pulse; S_2=Magnets S_3=Batteries S_4=Rubber Band S_5=Solutions (population 2 only)
S_6=Containers (population 1 only) M_1=Dice M_2=Calculator M_3=Folding and Cutting
M_4=Around the Bend M_5=Packaging SM_1=Shadow SM_2=Plasticine S= Science Task
M= Mathematics Task SM= Combined Science and Mathematics Task

science tasks (including shadows) in the eighth grade indicated that the percent of the Iranian students' correct responses was lower than the international mean (50 vs. 58). From among these tasks. the average performance of Iranian students on two tasks was greater than the international means. These tasks are pulse (55 vs. 44) and shadows (43 vs. 35). In the solution task. Iran's mean is almost close to the international mean (47 vs. 50)

The average percentage scores overall on six science tasks in the fourth grade showed that the percent of the Iranian students' correct responses was lower than the international mean (36 vs. 43). Iranian students performed the best. relative to those of the other countries. in pulse (41 vs. 36) and close to the international mean in batteries (40 vs. 41) (Kiamanesh, 1997).

B: By performance expectations

International attention to the importance of the processes of inquiry such as understanding. investigating, and communicating, as intended educational goals in mathematics and science. is rapidly increasing (Robitaille. et al.. 1993).

In TIMSS. the term "performance expectation" describes the kind of performance that students will be expected to demonstrate--problem solving or using scientific or mathematical procedures. reasoning and conjecturing. the ability to plan. conduct. and interpret an investigation--while engaged in an activity. The performance expectation categories as defined for TIMSS were derived from Robitaille et al.(1993). More specifically. the performance expectation categories for eighth and fourth graders in the performance assessment tasks were derived from the five main categories proposed for mathematics and science. In fact. there were three and two categories of performance expectation for science and mathematics tasks. respectively. These categories according to Harmon et al. (1997) are as follows:

- Scientific Problem Solving and Applying Concept Knowledge.
- Using Scientific Procedures.
  Scientific Investigating.
- Performing Mathematical Procedures
- Problem Solving and Mathematical Reasoning.

Here it should be noted that since some of the tasks and items were complex. and the students were usually involved in more than one performance expectation at a time. in the analysis of the results only one performance expectation was regarded as the primary performance category associated with each task item.

The average percentage score for each of the five performance expectation categories was calculated based on the percentage scores for each item within the category averaged across all items within the same category (Harmon, et al., 1997).

B1: Science performance expectations

The results of science performance expectations for eighth graders, presented in Table 2. revealed that even though Iranian students performed significantly better in "scientific problem solving and applying concept knowledge" than the other two science categories. their counterparts' performance was significantly lower in this performance expectation category than the other two ones.

Further. the Iranian eighth graders' performance in the other two categories was similar to their international counterparts with average score of about

Table 2: Science Performance Expectations Means for Eighth and Fourth Grade Students

Grade	Performance Expectation	Number of Items	Iran's Mean %	International Mean %	Iran's Rank
Eight	Scientific Problem Solving And Applying Concept Knowledge	12	61	47	1 out of 19
	Using Scientific Procedures	7	53	59	15 out of 19
	Scientific Investigating	16	56	60	15 out of 19
Four	Scientific Problem Solving and Applying Concept Knowledge	14	34	23	1 out of 9
	Using Scientific procedures	8	57	58	8 out of 9
	Scientific Investigating	13	37	43	8 out of 9

56% in comparison to the 60% international average score. In the fourth grade, students performed significantly better in "using scientific procedures" than in either of the two categories at both national and international levels.

Although the tasks and items within different categories were not the same for both grades, in particular items on "problem solving" and "investigating", the international and national eighth graders' average percentage scores for "using scientific procedures" category were comparable for the fourth grade performance.

According to the data presented in Table 2, the performance of both groups of students in the two categories of "scientific problem solving and applying concept knowledge" and "scientific investigating" significantly increased during four years of schooling, i. e., from the fourth grade to the eighth one.

The mean average percentage scores for Iranian eighth graders, in 10 out of the 12 items, in "scientific problem solving and applying concept knowledge" category were better than the international average percentage score. Further, in this category the mean average percentage score for Iranian fourth graders in 9 out of 14 items was better than the international average percentage score.

In "using scientific procedures" category, the Iranian eighth graders' average performance was better than the international average only in 1 out of 7 items, and in two other items the average performance of both groups were almost equal. For the fourth grade, the average performance of Iranian students in 3 out of 8 items of the above-mentioned category were better than the international average score.

The Iranian eighth graders' average performance in 7 out of 16 items in "scientific investigating" category were better than the international average score. The fourth grade Iranian students performed better in 2 out of 13 items than their international counterparts in "scientific investigating" category and average performance of the two groups were almost equal in 4 out of 13 items (Kiamanesh, 1997).

B2: Mathematics performance expectations

At both international and national levels, the eighth grade students performed significantly better in "performing mathematical procedures" category than in "problem solving and mathematical reasoning" one. It should be notified that even though the two categories for the mathematics performance expectations were the same for both grades, the tasks and items included within these categories were different. At the international level, the fourth graders performed better in "performing mathematical procedures" category than the other one. In Iran, however, the students performed similarly in these two categories (Table 3).

The mean average percentage score for eighth grade Iranian students only in 1 out of 13 items in "performing mathematical procedures" category was better than international average percentage score. Furthermore, in 3 items the performance of both groups was almost equal. The mean average percentage scores for the Iranian fourth graders in 3 out of 12 items in the above–mentioned category were better than the international average percentage score and in one item, the performance of the students was almost equal at national and international levels.

The mean average percentage scores for the eighth grade Iranian students in 6 out of 21 items in "problem solving and mathematical reasoning" were better than the international average, and in 2 other items the average performance of both groups was almost equal. In this category, the Iranian fourth graders performed better in 11 out of 16 items than their international

Table 3: Mathematics Performance Expectation Means for Eighth and Fourth Grade Students

Grade	Performance Expectation	Number of Items	Iran`s Mean %	International Mean %	Iran`s Rank
Eight	Performing Mathematical Procedures	13	61	70	17 out of 19
	Problem Solving and Mathematical Reasoning	21	49	52	14 out of 19
Four	Performing Mathematical Procedures	12	40	43	7 out of 9
	Problem Solving and Mathematical Reasoning	16	43	32	1 out of 9

counterparts. In two other items, the performance of both groups was almost equal (Kiamanesh).

C: Gender differences

Gender differences presented in this paper are calculated based on the proportion of students making fully correct responses to items averaged across items within a single task. Accordingly, in the case of eighth graders the average percentage scores overall 12 tasks revealed that the percent correct for boys was four percent greater than that for girls (54 vs. 50). Further, *this comparison revealed that in seven tasks boys* performed better than girls, and in four ones the girls outperformed the boys.

In the fourth grade, the percent correct for girls was two percent greater than that for boys (39 vs. 37). More specifically, in five tasks the boys' performance was higher than that of the girls', and in six tasks the girls outperformed the boys.

In general, the aforementioned observed differences between boys and girls at both grades in all tasks and were relatively small and at 0.05 level of significance no significant differences between genders were found.

D. Comparing written achievement test and performance assessment results

The use of equipment and apparatus as useful aids in teaching and learning processes is not emphasized in Iran's educational system. Furthermore, enough attention is not paid to this issue in available science and mathematics textbooks. In addition, no evidence of using hands–on tasks in Iranian schools is available. In fact, TIMSS performance assessment tasks were the first experience in conducting hands–on tests as a new technology for measuring students' knowledge and performance skills.

In order to calculate the international means for performance assessment tasks and written achievement test, the means of the 9 participant countries in the fourth grade were averaged and the same procedure was replicated for the 19 countries which participated in the eighth grade.

In the case of the fourth graders, the results of the calculated average percent correct response to comprising items of mathematics and science achievement tests showed that the students have respectively answered 38 and 40 percent of these items correctly (Mullis et al., 1997 & Martin, et al., 1997). The average of percent correct response for the fourth grade Iranian students in performance assessment tasks were 40 and 36 percent, respectively (Kiamanesh, 1997). The comparison of 38 and 40 percent correct responses on mathematics and science main achievement tests for fourth graders with 40 and 36 percentage scores on mathematics and science tasks showed that in general the results for the two different methods of measurement are the same. Nevertheless, from among the participant countries, the rank of fourth grade Iranian students in mathematics performance assessment tasks was better than their rank in the written test.

The eighth grade Iranian students respectively answered 38 and 50 percent of the main achievement mathematics and science items correctly (Beaton, et al., 1996a & Beaton, et al., 1996b). The average of percent *correct response for the Iranian eighth graders that* participated in the assessment performance was 54 for mathematics and 50 for science tasks (Kiamanesh, 1997). The comparison of the results obtained from the two instruments revealed that the mean performance of students in mathematics tasks was much better than their mean score on written achievement test. Hence, at the international level, the rank of eighth grade students in mathematics performance assessment tasks was better than their rank in the written test (see Table 4).

The high correlation between mean performance assessment scores on mathematics assessment tasks and achievement tests for fourth graders (r = 0.831) and also the high correlation between mean performance assessment scores on science tasks and science achievement tests (r = 0.839) indicated that those who did well on each of the achievement tests, performed well on the respective assessment tasks as well.

The correlation between country means (9 countries) derived by the two instruments at fourth grade was 0.78 for mathematics and 0.40 for science. One of the reasons for obtaining this low correlation in science was the fact that some of the tasks were difficult for this age group.

In the case of eighth graders, the obtained correlations between the students' scores on the main written tests and performance assessment tasks in mathematics and science were 0.767 and 0.779, respectively. And the correlations between country means (19 countries) derived by two instruments were 0.78 for mathematics and 0.81 for science (Garden, 1997, p.106). Again, these correlations show that the two measurement technologies have a large common variance in measuring students' ability (Garden, 1997).

Conclusion

The eighth and fourth grade Iranian students' average scores in "scientific problem solving and applying concept knowledge" category were higher than the average score of the other participant countries. Both groups' average scores on items measuring "applying

Table 4: Average Percent Correct Scores for Iranian Fourth and Eighth Graders Using both Measurement Technologies

Grade	subject	Iran's Rank		Measurement		Correlation between WT & PA at National level	Correlation between WT & PA at International level
		Written Test	Performance Assessment (PA)	Written Test (WT)	Performance Assessment		
Eight	Mathematics	18	14	38	54	.767	.78
	Science	17	16	50	50	.779	.81
Four	Mathematics	9	3	38	40	.831	.78
	Science	9	8	40	36	.839	.40

scientific principles to develop explanation" and "applying scientific principles to solve quantitative problems" were higher than the international averages. Further, the performance of Iranian students in both grades revealed that they did not perform well in "interpreting investigational data."

In the mathematical performance expectations, only fourth grade students' average scores on "problem solving and mathematical reasoning" category was better than the international average score. Both groups had average scores on most of the items measuring "applying mathematical principles to solve quantitative problems" and only the fourth grade average scores on most of the items measuring "problem solving" were higher than the international average scores. The most observable problems with regard to eighth graders were "predicting", "conjecturing", and "problem solving" tasks.

In general, the major issue which overshadowed the students' performance to a great extent was the students' disability to describe and explain different phenomena. In most of the cases, their inability in expressing themselves through writing led to their low performance.

Average performance scores obtained through mathematics performance assessment tasks for eighth grade students was 16 percent and for fourth grade students was 2 percent higher than the average percentage score obtained through mathematics written achievement tests. Average percentage science scores obtained through two forms of testing for eighth grade students were equal. But fourth grade students performed better on written achievement tests. In general, the ranks of Iranian students in the assessment performance tasks were better than their ranks in written achievement tests.

Although all differences between boys and girls for the two forms of testing were not statistically significant at 0.05 level, the girls outperformed the boys on some of the practical tasks; specially fourth grade girls' performance in mathematics tasks were notable. This evidence showed that in Iranian educational system, no matter how good the quality of the education is, it leads to the same outcome in both genders.

Further, for both genders, the high correlations between country means for the written achievement tests and performance assessment tasks in the two subjects and also the high correlations between Iranian students on written achievement tests and performance assessment tasks in the two subjects are indicative of the fact that the two methods of testing are measuring the same students' characteristics. Nevertheless, further research is needed to see whether information gained from using hands–on tasks in mathematics or science justifies the great demand on time, personnel, and financial resources or not.

One of the most important findings of this study is the great difference between mathematics and science performance of Iranian students and performance of students from other developing and under–developing countries. This low performance is indicative of some serious problems in Iranian educational system which should be alleviated through undertaking remedial actions. In fact, substantial changes should take place in mathematics and science programs in order to bridge the existing gap in our educational system.

Further. the results of this study indicated that Iranian students' performance on hands–on tasks is task specific and it varies from task to task depending on different variables. More specifically. Iranian students outperformed other students in some tasks and even in a number of cases they were among the best ones. This case holds true for individual performance of students as well as the total performance of Iranian students. These results are in complete accordance with findings of the pervious studies in regard to performance of students on hands–on tasks. According to Shavelson et al. (1992). the individual performance on hands–on tasks has been found to be task specific and it varies depending on different variables such as past experience and interest of students.

References

Beaton. A. E. Martin. M. O.. Mullis. I.V. S.. Gonzalez. E. J . Smith. T. A.. and Kelly. D.L. (1996a). *Science achievement in the middle school year: IEA's third international mathematics and science study.* Chestnut Hill. MA: Boston College.

Beaton. A. E. Martin. M. O.. Mullis. I. V. S.. Gonzalez. E. J.. Smith. T. A.. and Kelly. D. L. (1996b). *Mathematics achievement in the middle school year: IEA's third international mathematics and science study.* Chestnut Hill. MA: Boston College.

Bond. L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice,* 14 (4). 21- 24.

Garden. R. A. (1997). *Performance assessment in the third international mathematics and science study: New Zealand results.* New Zealand. Wellington: Research and International Section. Ministry of Education.

Harmon. M.. Smith. T. A.. Martin. M. O.. Kelly. D. L.. Beaton. A. E.. Mullis. I. V. S.. Gonzalez. E. J.. & Orpwood. G. (1997). *Performance assessment in IEA's third international mathematics and science study (TIMSS).* M. A.: Boston College.

Herman. L. J.. Aschbacher. R. P.. and Winters. L (1992). *A practical guide to alternative assessment.* The Regents of the University of California: Association for Supervision and Curriculum Development.

Kiamanesh. A. L. (1997). *Performance assessment in third international mathematics and science study: Upper grade populations 1 and 2.* Tehran: Institute for Educational Research.

Martin. M. O.. & Kelly. D. L..(Eds.) (1996). *Third international mathematics and science study: technical report. volume 1: design and development.* Chestnut Hill M. A.: Boston College.

Martin. M. O.. Mullis. I. V. S.. Beaton. A. E.. Gonzalez. E. J.. Smith. T. A.. and Kelly. D. L. (1997). *Science achievement in the primary schools years.* Chestnut Hill. MA: Boston College.

Mullis. I. V. S.. Martin M. O.. Beaton. A. E.. Gonzalez. E. J.. Kelly. D. L. and Smith. T. A. (1997). *Mathematics achievement in the primary schools years.* Chestnut Hill. MA: Boston College.

Robitaille. D. F.. Schmidt. W. H.. Raizen. S.. Knight. C. M.. Britton. E. & Nicol. C. (1993). *Curriculum frameworks for mathematics and science.* Vancouver: Pacific Educational Press.

Robitaille. D. F.. Taylor. A. R.. Brigden. S. R.. and Marshall. M. A. (1997). *TIMSS–Canada Report. Volume 3: Hands–on problem solving.* Vancouver: University of British Columbia.

Ryans.D. J. and Frederiksen. N. (1951) Performance tests of educational achievement. In E.F..Lindquist (Ed.) *Educational measurement* (pp. 455 – 494). Washington. DC: American Council on Education.

Shavelson. R. J.. Baxter.G. P.. and Xiaolong. G. (1992). Sampling variability of performance assessment. *Journal of educational measurement,* Volume 30(3).