PERFORMANCE ASSESSMENT: IDEAS FOR REDESIGN AND REVITALIZATION OF ITS ROLE IN INTERNATIONAL COMPARATIVE STUDIES

HsingChi A. Wang
University of Calgary, Canada

Abstract

In addition to the written mathematics and science achievement tests administered to assess students' content knowledge worldwide, TIMSS 1995 also collected data from Performance Assessment. In 1999 when the TIMSS-Repeat was conducted, Performance Assessment was not administered. The objective of this study is to rethink, and perhaps through redesign to revitalize the role of performance assessment in international comparative study. There is a need to invest in deeper discussions about what has been learned in the TIMSS 1995 curriculum to achievement analyses, and to consider what newer research findings on performance assessment can contribute to the understanding of how to know what students know. The TIMSS 1995 analyses on curriculum to achievement and the numerous textbooks examined, show that there is a link between how science content is presented, and the outcomes in content exams. The performance assessment results show that some of the test items in relation to the content and performance expectation link validate the results of written TIMSS assessment. This leads to an important implication for future international studies on science education. In order to best portray students' learning in science, the inclusion of performance assessment is critical. However, test items in the future performance assessment should consider the content-performance expectation pairs found in this study to enhance the accuracy of studying what students learn in science across nations. The sensitivity to content and performance expectation pairs in the redesign of performance assessment will bring richer and more informative results to the community of international comparative studies.

INTRODUCTION

In the past three decades, the research findings from cognitive scientists, neuroscientists, and educators have begun to converge and provide important insights into understanding how people learn (NRC, 2000).

> "...the goal of education is better conceived as helping students develop the intellectual tools and learning strategies needed to acquire the knowledge that allows people to think productively. ...Fundamental understanding about subjects, including how to frame and ask meaningful questions about various subject areas, contributes to individuals' more basic understanding of principles of learning that can assist them in becoming self-sustaining, lifelong learners" (p.5).

This stated goal of education should not be just an aspiration. Curriculum goals will remain as ideology if curriculum alignment does not exist. To reach curriculum alignment, there should be a 'cohesive' relationship among educational standards, instruction and assessment. That is, to help students become independent problem solvers, the educational system needs to provide opportunities for students to learn the necessary skills; skills of how to ask questions, to locate resources, to study and research, and to answer the identified questions. Teachers should be provided with help to create these opportunities. Assessment systems need to be designed to encourage classroom instruction that supports the curriculum ideology. The focus of this study is on the design of assessment systems. I will discuss ideas pertaining to the design of performance assessment that insures assessment aligns with recent curriculum ideology[1].

Shavelson, Baxter, and Pine (1991) argued that the central issue in creating assessment is its content representativeness—to what extent do the assessments represent important concepts within a subject matter domain, and to what extent do they fit with local or state curricula? In order to ensure students learn to become critical thinkers and problem solvers, the assessment system needs to reflect this ideology. Shavelson et al., further argue that "good instructional activities can be translated into assessments and good assessments can be used as instructional activities" (p.357). Martin and Miller's (1996) study shows that when teachers used the performance assessment format to design science instruction, there was an increase in students' logical thinking ability. Vogler's (2002) findings support Shavelson et al.,'s argument. He found that the established state-wide performance assessment system did result in a significant increase in instructional practices that focused on enhancement of critical thinking and problem solving skills.

RATIONALE OF THE STUDY

Since the release of the Third International Mathematics and Science Study (TIMSS) in 1997, many secondary analyses of content knowledge attainment have been conducted worldwide, and many reform movements have taken place around the world based on information gained from TIMSS. Naturally, a lot of the discussions have centered on the content knowledge organization and teaching in these content areas, but very few discussions are about teaching for inquiry, which constitutes one of the major goals for most of the TIMSS countries.

In addition to the written mathematics and science achievement tests administered to assess students' content knowledge worldwide, TIMSS 1995 also collected data from Performance Assessment (Harmon, Smith, Martin, Kelly, Beaton, Mullis, Gonzalez, & Orpwood, 1997). In 1999 when the TIMSS-Repeat was conducted, Performance Assessment was not administered. The high cost and difficulties in administration of performance assessment in international comparative studies can be prohibitive to reintroducing performance assessment. Because there have been very few analyses of the performance assessment data collected in 1995, very few educational implications have resulted. Thus, the chances that performance assessment will continue to feature in international comparative studies are slim.

Yet, the value of performance assessment in educating critical thinkers and problem solvers is immense (Ayala, Yin, Schultz, & Shavelson, 2002; Linn, 2003; Shavelson, Baxter, & Pine, 1991). Thus, the objective of this study is to rejuvenate discussions around performance assessment in international comparative studies.

CURRICULUM TO ACHIEVEMENT FINDINGS

With the release of *Why Schools Matter*, Schmidt, McKnight, Houang, Wang, Wiley, Cogan, & Wolfe (2001) revealed how aspects of curriculum impact learning mathematics and science. Textbooks and teachers are what these researchers defined as implementing aspects of curriculum. They found that in science textbooks, among earth science, life science, and environmental science, physical science is the most likely content to be presented with cognitive demanding expectations-theorizing, analyzing, problem solving, and science investigation. Fourteen percent (14%) of science books couple the above cognitive demanding expectations with physical science content. When summarized across all topics for each country, Denmark, New Zealand, and Singapore have over twenty percent (20%) of their middle school science textbook space devoted to cognitive demanding expectations coupled with science content.

Furthermore, Schmidt et al., (2001) showed that among all the curriculum variables examined in *Why School Matters,* the only variable that was found to be significantly related to achievement gain in the science content knowledge test is the textbook coverage related to more demanding performance expectations. This result means that, in terms of learning outcomes, what is covered in the science textbook does not have as significant impact as how a topic is covered. Another finding from the curriculum assessment can help explain how textbooks may impact students' learning: Schmidt et al., (2001) found that science textbook space is related to both instructional time and a proportion of teachers teaching a topic.

The correlation of determination also shows that middle school science textbook space accounts for from around ten to forty percent (10%-40%) of variance in instructional time devoted to the specific science topics, and the proportion of teachers teaching a topic. Although there is no significant direct impact of instruction on students' science performance, which might be a result of too many subtopics involved, it is important to consider this correlation of determination in science textbook coverage to science instruction.

Science involves various aspects of inquiry—designing, conducting, analyzing, and interpreting data from empirical investigations. These various aspects of inquiry are the conceptual heart of today's science education. The relationship between the textbook coverage and teacher variables in middle school science, and the relationship between performance expectation (per textbook coverage) and learning not only signals a very critical direction in the understanding of measuring science achievement, but also reinforces the idea of promoting teaching science for inquiry.

THE PERFORMANCE ASSESSMENT

In TIMSS 1995, a sub-sample of 9- and 13-year-old students who participated in the main TIMSS assessment also participated in a session of performance-based testing. The results of the TIMSS Performance Assessment were of great interest to the U.S. educational community for two primary reasons. First, calls for educational reform from groups such as the National Council of Teachers of Mathematics (NCTM, 1989, 2000), the American Association for the Advancement of Science (AAAS, 1993), and the National Academy of Sciences (NAS, 1996), have included demands for more performance-oriented assessment practices that allow students to demonstrate applications of complex thinking and reasoning. Second, the TIMSS Performance Assessment represents the only attempt for a cross-national study in performance assessment. The results of the Performance Assessment, therefore, allow educators to gauge the performance of a particular country's students against the performance of students worldwide in skills deemed important by the global educational community. The results also provide the educators who are interested in measurement with a chance to determine the feasibility of developing and implementing a large-scale performance assessment and to evaluate the quality and value of the information gained from such an approach.

The results of TIMSS' Performance Assessment have attracted interest from others besides the American educators. Japanese concern over their students' achievement in performance assessment has led the Japanese government to form a national task force that is conducting investigations of the problematic areas where Japanese students did not do well. Despite the high cost and difficulties involved in administering performance assessment, there are merits to conducting large-scale performance assessment, because "Hands-on assessments distinguish students' experienced in hands-on science from students who have received a more traditional textbook approach. . ." (Shavelson, et al., 1991, p.358). Performance assessments and multiple-choice, traditional paper-and-pencil type assessments measure different aspects of science achievement.

Twenty-one countries participated in the TIMSS performance assessment for 13-year-old students. Ten of the 21 countries also participated in the performance assessment for 9-year-old students. Table 1 shows a list of the countries.

The TIMSS performance tasks were designed to reflect the TIMSS mathematics and science curriculum frameworks (Robitaille, McKnight, Schmidt, Britton, Raizen, & Nicol, 1993), and to be feasible for administration in a large-scale international assessment. In particular, attention was focused on developing tasks that represented the range of performance expectations in the TIMSS curriculum frameworks (Harmon, et al., 1997).

Table 1: Countries Participating in Performance Assessment of TIMSS

	Population 1 *(average age: 9 years old)*	*Population 2* *(average age: 13 years old)*
Australia	1	1
Canada	1	1
Columbia	0	1
Cyprus	1	1
Czech Republic	0	1
England	0	1
Hong Kong	1	1
Iran, Islamic Rep.	1	1
Israel	1	1
Netherlands	0	1
New Zealand	1	1
Norway	0	1
Portugal	1	1
Romania	0	1
Scotland	0	1
Singapore	0	1
Slovenia	1	1
Spain	0	1
Sweden	0	1
Switzerland	0	1
United States	1	1

Note: '0' in the cell means the country did not participate
'1' in the cell means the country participated

Harmon and Kelly (1996) have described in detail the history of TIMSS performance assessment tasks development. Following is a brief summary of their report.

Among the challenges faced by the Performance Assessment Committee (PAC), students would have only 90 minutes of testing time; yet, the tasks needed to cover a range of mathematics and science content. The PAC began by reviewing tasks submitted from Australia, Canada, England, New Zealand, and the U.S., as well as from two prior international studies. In the end, 26 tasks were selected or created for initial piloting in six countries. Twenty-two tasks at each testing population were then used for field testing in 19 countries. Tasks for the final instrument were selected based on the following criteria:

- *Difficulty level.* Selected tasks had no more than 50% incorrect or missing responses on easier items, and no more than 70% on more difficult items in the field trial. For items used in both populations, these difficulty criteria were applied with Population 2 provided that Population 1 students showed some achievement on some of the items of a task.

- *Subject-matter expert ratings.* All tasks selected received ratings of 2 or higher (on a scale of 1 to 4, 4 as the highest) from mathematics and science experts within each country, based on interest, feasibility, content quality, and congruence with curriculum and instruction within the country.
- *Administrators' ratings.* All tasks selected received ratings of 2 or above on a scale of 1 to 4 (4 as the highest) from administrators of National Research Centers in participating countries.
- *Balance.* Tasks were selected to maintain a balance between the number of mathematics and science tasks, and between tasks estimated to take 10 to15 minutes to complete and those estimated to take 25 to 30 minutes to complete.
- *Framework representation.* As much as possible, the selected tasks sampled across the subject-matter content of the TIMSS curriculum frameworks. Content coverage was necessarily selective, since only 12 tasks were to be administered in each population. Coverage of all age-relevant performance expectation categories in the frameworks was achieved.
- *Linkage between Populations 1 and 2.* Four complete tasks were identical and seven more were similar for the two populations to facilitate comparisons between the two populations. The primary differences were re-ordering of items or the addition of one or two items at the end of the task.
- *Professional judgment of task quality.* Tasks with lower than 50% correct or partially correct responses in the field trial were retained only if minor revisions would render them more accessible to younger students without destroying their assessment intent, and if they yielded rich information about common approaches to tasks and common errors and misconceptions.

The final TIMSS' Performance Assessment contained 12 tasks at each testing population. Five of the 12 tasks focused on science content for 13-year-old students. These tasks are Pulse, Magnets, Batteries, Rubber Band, and Solutions. Except the Solutions task, all tasks were modified for testing 9-year-old students. Two tasks, Shadows and Plasticine, focused on a combination of mathematics and science content and both populations participated in these two tasks.

This study focused only on the science tasks analyses for population 2 students. In the science tasks, students had to develop and conduct short "experiments" or investigations. Students needed to demonstrate the ability to collect, organize, and interpret data. They also had to predict, explain results or their reasoning, and come up with algorithms to represent situations. Each task began with the statement of a problem and consisted of a series of questions that generally increased in difficulty. Some questions relied on the results of earlier steps, and others relied on prior knowledge of content facts or concepts (Harmon, et al., 1997).

REANALYSIS RESULTS OF TIMSS PERFORMANCE ASSESSMENT SCIENCE ITEMS

Table 2 is the result of re-analyses of the science tasks. The content codes starting with 'M' are codes from the TIMSS mathematics content framework, whereas codes with 'S' are from the TIMSS science content framework. The same scheme is also applied to the performance expectation codes. The rationale for reassigning codes to the tasks is because the re-analyses of the performance assessment data will be compared to the TIMSS curriculum to achievement results reported by Schmidt et al., (2001), where the TIMSS curriculum frameworks applied in the analyses are different from what has been applied in Harmon et al.,'s (1997) work.

According to the results as presented in Table 2, 11 items across the 38 items over the seven science (science/math) tasks are found to have no content codes. Yet, every science performance assessment task contains at least one science topic, except the task "Plasticine." These items have the following performance expectation codes: interpreting investigational data (S2.4.4), sharing information (S2.5.2), designing investigations (S2.4.2), using apparatus (S2.3.1), & gathering data (S2.3.3). The most frequently appearing code among these five is the sharing information performance expectation. This communication skill is found to be the most prevalent performance expectation across the seven science tasks. The seven tasks are also found to focus on students' ability to use apparatus, equipment and computers (S2.3.1), to conduct routine experimental operations (S2.3.2), to gather data (S2.3.4), to organize and represent data (S2.3.4), and to interpret data (S2.3.5). There are very few items that intend to assess student's abilities in theorizing, analyzing and solving problems (S2.2.1, S2.2.2, S2.2.3, S2.2.4, S2.2.5).

The table also shows that apart from the Batteries task, all tasks contain mathematics codes even though out of the seven tasks only Shadows and Plasticine were designed to be math/science tasks. This reanalysis result shows that performance assessment for science in TIMSS seems inevitably to test also students' mathematics content knowledge and mathematics process skills.

REANALYSIS RESULTS OF TIMSS' PERFORMANCE ASSESSMENT

The following findings are derived from an entirely new set of analyses using TIMSS performance assessment data. The U.S. data is the focal data set in these new analyses:

1. U.S. performance compared to other countries was strongest on the "Pulse" tasks. The students seemed to do particularly well on items asking them to interpret and draw conclusions from data presented in tables – this was true on other tasks as well. U.S. students had a difficult time on the Batteries task determining which batteries were good and which were worn out, and, as was true also on the "Plasticine" task, they had difficulty using an appropriate strategy, scientific principles, or science content knowlegde to complete the task or at least in articulating what they did.

Table 2: Recoding Results of TIMSS' Science Performance Assessment Tasks

Task	*Item*	*Points*	*Description*	*Skill/concept*	*Content code1*	*Content code2*	*Content code3*	*P.E. code1*	*P.E. code2*	*P.E. code3*	*P.E. code4*	*P.E. code5*
Pulse	1	2	make table	use adequate table format	M1.7.1			S2.3.4	M2.2.2	M2.2.3		
Pulse	2	3	record data	present quality data	M1.7.1			S2.3.2	S2.3.3	M2.2.2		
Pulse	3	2	describe change in pulse	describe pattern, communicate data				S2.4.4	S2.5.2			
Pulse	4	3	explain change in pulse	explain science concept	S1.2.2.1	S1.2.5	S1.2.5	S2.2.3	S2.5.2			
Magnets	1	1	determine stronger magnet	reach accurate conclusion based on test				S2.4.4				
Magnets	2	1	explain strategy	use plausible strategy	S1.3.3.7			S2.4.2	S2.5.2			
Batteries	1	2	determine good and worn-out batteries	reach accurate conclusion based on test				S2.4.4				
Batteries	2	2	explain strategy	use systematic strategy				S2.4.4	S2.5.2			
Batteries	3	1	identify appropriate direction of batteries	use knowledge of scientific concept	S1.3.3.6			S2.1.2				
Batteries	4	2	explain why batteries go that way	explain scientific concept	S1.3.3.6			S2.2.3	S2.5.2			
Rubber Band	1	2	make table	use adequate table format	M1.7.1			S2.3.4	M2.2.1			
Rubber Band	2	3	record data	present quality data	M1.2.1			S2.3.3	M2.2.3			
Rubber Band	3	3	graph results	graph (data and format)	M1.7.1			S2.3.4	M2.2.2			
Rubber Band	4	2	determine increase in length	interpret graph	M1.7.1			S2.3.4	M2.2.3			
Rubber Band	5	2	describe increase in length	describe trend/data	M1.6.1	M1.7.1		S2.3.4	M2.4.4	M2.5.3		
Rubber Band	6	1	predict new length	use graph to predict	S1.3.1.2	M1.6.1	M1.7.1	S2.3.5	S2.4.4	S2.5.2	M2.4.5	M2.5.3
Rubber Band	7	2	explain prediction	explain prediction	S1.3.1.2			S2.3.5	S2.1.2	S2.5.2		

Table 2: (Continued)

Task	*Item*	*Points*	*Description*	*Skill/concept*	*Content code1*	*Content code2*	*Content code3*	*P.E. code1*	*P.E. code2*	*P.E. code3*	*P.E. code4*	*P.E. code5*
Solutions	1	2	describe plan	plan and communicate plan				S2.4.2	S2.5.2			
Solutions	2	1	measure temperature	use thermometer	M1.2.1			S2.3.1	M2.2.1	M2.2.2		
Solutions	3	2	make table	use adequate table format	M1.7.1			S2.3.4	M2.2.1	M2.2.3		
Solutions	4	3	record data	present quality data	M1.7.1			S2.3.2	S2.4.3	M2.2.3		
Solutions	5	2	determine effect of temperature	interpret data	M1.7.1			S2.3.5	S2.5.2	M2.2.2		
Solutions	6	2	explain effect of temperature	explain scientific concept	S1.3.4.2	S1.3.3.2		S2.2.3	S2.5.2			
Solutions	7	2	describe changes you would make to plan	alter plan based on experience				S2.4.2	S2.5.2			
Shadows	1	2	what happens to size observation	describe/communicate				S2.3.3	S2.3.1			
Shadows	2	2	why is shadow larger	explain why, concept	S1.3.3.5			S2.2.3	S2.2.4			
Shadows	3	2	3 positions, 2x	do	M1.2.1			S2.4.3	M2.2.1	S2.3.1	M2.2.3	
Shadows	4	2	describe procedure	describe method	M1.2.1			S2.5.2	M2.2.1	M2.2.2		
Shadows	5	2	present meas. Clearly	present clear data	M1.7.1			S2.3.4	M2.2.3			
Shadows	6	2	what conclusions, rule	conclude	M1.6.1	M1.7.1		S2.4.5	M2.4.3	S2.4.4	S2.5.2	M2.4.4
Plasticine	1	1	make 20g plasticine	use equipment, manipulate material	M1.2.1			S2.3.2	M2.2.1	S.2.3.1		
Plasticine	2	2	describe procedure	describe method				S2.5.2				
Plasticine	3	2	make 10g plasticine manipulate material	use equipment,	M1.2.1			S2.3.2	M2.2.1	S2.3.1	M2.3.3	
Plasticine	4	2	describe procedure	describe method				S2.5.2				
Plasticine	5	1	make 15g plasticine manipulate material	use equipment,	M1.2.1			S2.2.2	M2.2.1	S2.3.2	S2.3.1	M2.3.3
Plasticine	6	2	describe procedure	describe method				S2.5.2				
Plasticine	7	2	make 30g plasticine	use equipment, manipulate material	M1.2.1			S2.2.2	M2.2.1	S2.3.2	S2.3.1	M2.3.3
Plasticine	8	2	describe procedure	describe method				S2.5.2				

2. U.S. eighth-grade students performed particularly poorly on measurement items. Students performed near the average on the performance assessment when asked to measure their pulse, use a metric scale, and use a metric ruler to measure the change in the length of a rubber band. However, they did not perform well when asked to measure, in centimeters, scale models of furniture and convert these to real sizes. This reflects the same findings in TIMSS content assessment that U.S. students did relatively the worst in "Units."
3. U.S. students' best relative performance was in science, particularly in the area of "science processes". This area included items on which the student had to take pulse measurements and graph data. Within-country strengths were in performing mathematical and science procedures, and weaknesses were in "reasoning" and "problem solving".
4. Using the theoretical framework described in the method section, with the assistance of the curriculum instantiations, the teacher variables were found to be one of the indicators for better student achievement in performance assessment; according to teacher questionnaire data, very few U.S. teachers reported applying strategies to promote higher-order, complex thinking in the classroom. This may be one of the many reasons why students do not do well in the reasoning and problem solving performance assessment items.

CONTENT AND PERFORMANCE EXPECTATION PAIRS IN SCIENCE TEXTBOOKS

After analyzing 156 science textbooks for students age 13 from more than 40 countries, Table 3 shows how textbooks portray the performance expectations for science content. For each Performance Expectation (PE) code, we calculated the percentage of blocks of the aggregated textbook data in each content topic area. We then selected the content area with the largest percentage for each PE. We identify and list these content areas, and juxtapose the corresponding PE code and label. For example: Topic "magnetism" (TIMSS content code"1337") is most likely to be associated with a performance expectation of "conducting investigation" when introduced in the science textbooks of all the TIMSS countries. For each content topic, we calculated the percentage of blocks in each PE. We then selected the content area with the largest percentage for each PE. We identify and list the content areas, and juxtapose all of the corresponding PE codes and labels.

This information is particularly interesting for the reasons mentioned earlier, i.e.,: (1) textbook coverage impacts on instructional time, (2) cognitive demanding content coverage impacts on students' science achievement in TIMSS written test, and (3) whether the instructional time was spent on higher-order, complex thinking tasks influenced students' achievement in performance assessment.

Table 3: Content Topics With Focal Performance Expectations

Content Topic	*PE*
1235 Biochemistry of Genetics	213 Understanding Thematic Information
1322 Macromolecules, Crystals	224 Constructing, Interpreting and Applying Models
1337 Magnetism Quantum Theory &	243 Conducting Investigations
Fundamental	212 Understanding Complex Information
1344 Particles	
1351 Chemical Changes	232 Conducting Routine Experimental Operations
1354 Energy & Chemical Changes	223 Applying Scientific Principles to Develop Explanations
1364 Relativity Theory	222 Applying Scientific Information to Solve Quantitative Problems
1365 Fluid Behavior Influences of	221 Abstracting and deducing Scientific Principles
Mathematics,	234 Organizing and Representing Data
1421 technology in Science	
Influence of Society on Science,	225 Making Decisions
1432 technology	235 Interpreting Data
164 World Population	233 Gathering Data
171 Nature of Scientific Knowledge	231 Using Apparatus, Equipment, and Computer
172 The Scientific Enterprise	211 Understanding Simple Information
182 Science & Other Disciplines	211 Understanding simple information, 233 Gathering Data,
1214 Organs, Tissues	251 Communicating: Accessing & Processing Information
1234 Evolution, Speciation, Diversity	213 Understanding Thematic Information
1321 Atoms, Ions, Molecules	224 Constructing, Interpreting, & Applying Models
1332 Heat & temperature	232 Using Apparatus, Equipment, & Computers
1335 Light	234 Organizing & Representing Data, 235 Interpreting Data
	212 Understanding Complex Information,
	221 Abstracting and Deducting Scientific
Principles,	222 Applying Scientific Principles to Solve Quantitative Problems,
	223 Applying Scientific Principles to Develop
1336 Electricity	Explanations
	232 Conducting Routine Experimental Operations,
1351 Chemical Changes	243 Conducting Investigations
Influence of Science, Technology	
1431 on Society	225 Making Decisions

These findings help us understand what may have been intended in our science textbooks and how that might have indirectly impacted on student learning. For the researcher who intends to design performance assessment tasks, these pairs of content and performance expectation should be included for consideration.

The performance assessment results also show that some of the test items with (respect) to the content and performance expectation link validate the results of the written TIMSS assessment. This leads to an important implication for future international studies on science education. In order to best portray students' learning in science, the inclusion of performance assessment is critical. However, the test items in any future performance assessment should consider the content-performance expectation pairs found in this study, as this will enhance the accuracy of studying what students learn in science across nations. The sensitivity to content and performance expectation pairs in the redesign of performance assessment will bring a richer and more informative result to the community of international comparative studies.

Eisner (1999) says: "Our educational aspirations have been influenced by the fact that our children will inhabit a world requiring far more complex and subtle forms of thinking than children needed three or four decades ago" (p.658). Performance assessment can be a critical step for us to change educational practices so that our children will receive from schools what they need in life. The lessons learned from this study and all modern research on performance assessment should be considered when we design performance assessment tasks in the next large-scale international comparative study.

References

American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy.* New York: Oxford University Press.

Ayala, C. C., Yin, Y., Schultz, S., & Shavelson, R. (2002). On science achievement from the perspective of different types of tests: A multidimensional approach to achievement validation. Los Angeles, CA: UCLA/Center for the Study of Evaluation Technical Report.

Eisner, E. M. (1999). The uses and limites of performance assessment. *Phi Delta Kappan*, *80*(9), 658-661.

Harmon, M. & Kelly, D.L. (1996). Development and design of the TIMSS performance assessment. In Martin, M.O. & Kelly, D.L. (Eds.) *The Third International Mathematics and Science Study technical report volume 1: Design and development.* Chestnut Hill, MA: Boston College.

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).* Chestnut Hill, MA: Boston College.

Martin, M., & Miller, G. (1996). Instructional improvement through performance assessment. *Thrust for educational leadership*, 25(7), 10-12.

National Council of Teachers of Mathematics. (1989, 2000). Principles and Standards for School Mathematics. NCTM.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school.* Washington, D.C.: National Academy Press.

National Research Council. (1996). *National Science Education Standards.* Washington, D.C.: National Academy Press.

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

Schmidt, W. H., McKnight, C.C., Houang, R., Wang, H.A., Wiley, D., Cogan, L., Wolfe, R. (2001). *Why Schools Matter: Using TIMSS to Investigate Curriculum and Learning.* NY, NY: Jossey-Bass.

Volger, K.E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. Education, 123(1), 39-56.

NOTE

1. This study is built on works by Dr. Pamela M. Jakwerth, who started the initial analyses while she was at the U.S. National Research Center-TIMSS at Michigan State University.