Cantrell, P. (2012). Using test blueprints to measure student learning in middle school science classrooms. *The Researcher*, *24*(1), 55-71

Using Test Blueprints to Measure Student Learning in Middle School Science Classrooms

Pamela Cantrell
Brigham Young University

Abstract: *This paper describes test blueprints as a method for developing end of unit tests as one measure of the effects of science teacher professional development on student content knowledge. The process for constructing test blueprints is first described relative to large-scale testing and is then modified and described for small-scale tests developed by teachers. The role of alignment to national and local standards for increasing the degree of test validity is also described. Models for subdividing the content and cognitive domains into hierarchical levels are provided from the literature and use of these levels relative to test item development is discussed. Two examples of the use of test blueprints in science teacher professional development programs are described, and overall implications are discussed.*

Keywords: Middle School Science, Teacher Professional Development

With the authorization of No Child Left Behind Act of 2001 (HR1) came the call for "nurturing and reinforcing a scientific culture of educational research" (Feuer, Towne, & Shavelson, 2002, p. 4) as underscored by use of the phrases "evidence-based practices" and "scientific research" mentioned over 100 times in the Act. Many saw this language as the call for randomized trials as the new gold standard for educational research. The response was predictable and impassioned. Berliner (2002) commented, "To think that his form of research is the only "scientific" approach to gaining knowledge—the only one that yields trustworthy evidence—reveals a myopic view of science in general and a misunderstanding of educational research in particular" (p. 18). He further stated that scientists in fields such as physics, chemistry, and geology would never tolerate the conditions under which educational researchers do their science with having to deal with the ubiquity of powerful contextual interactions that thwart the ease of replication, thus giving rise to the importance and necessity of qualitative inquiry in educational research.

Over the past decade extensive debates have ensued over the merits of using randomized trials in educational research vs. a more broad and diversified approach to methodology including quantitative, qualitative, and mixed methods (Henson, Hull, & Williams, 2010; Johnson & Onwuegbuzie, 2004; Tashakkori & Dreswell, 2008). Although the debate is still ongoing, a number of positive outcomes continue to emerge in educational research because of this new scientific orthodoxy (Howe, 2008) including a focus on better preparation of educational researchers (Henson et al., 2010), closer

Pamela Cantrell is an Associate Professor of Science Education and can be reached at the David O. McKay School of Education; Teacher Education, 206-G MCKB; Brigham Young University; Provo, UT 84602; Office: 801-422-4129; pamela_cantrell@byu.ed

attention to context in the research process (Berliner, 2002; Feuer et al., 2002), revisiting the foundational importance of how to collect reliable, unbiased and meaningful evidence (Slavin, 2008), and new online products from national forums to guide the researcher in merging randomized controlled trials with multiple methods (American Educational Research Association, 2009).

The purpose of this paper is to describe test blueprints as one method for generating valid and meaningful classroom-based evidence of student achievement useful for studies that measure the impact of professional development outcomes on student learning. Although the examples provided below are from science professional development programs employing mixed methods research using a variety of data sources to measure impact, the focus here will be only on the construction and use of test blueprints that could be developed by teachers with some guidance and applied across a number of content areas or disciplines. First, test blueprints will be situated in the measurement literature with a focus on alignment and validity. Test blueprint construction will be addressed in the next section, followed by detailed examples illustrating how test blueprints were successfully developed by teachers and used in two different professional development programs. The final section will discuss lessons learned and provide implications for the use of test blueprints in professional development studies.

LITERATURE REVIEW

A test blueprint is a table of specifications that provides the framework for a test. There are several ways to construct test blueprints that generally map, in some type of matrix design, three or more of the following elements: (a) content learning goal/objectives, (b) types or levels of knowledge (i.e. Bloom's Taxonomy), (c) item type, or (d) degree of difficulty. General considerations for test blueprints include ensuring that the desired coverage of topics and level of objective are addressed. Although the type of test blueprints addressed in this paper are intended for development by teachers with guidance from facilitators with some assessment expertise, the major use of test blueprints today is exemplified in large-scale testing such as state tests for science and mathematics, or national and global testing such as the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS). For these large-scale tests, very detailed test specifications are stipulated (Olson, Martin, & Mullis, 2007) and careful attention is given to establish alignment and validity. While the test blueprints design process for much smaller scale measurement does not require nearly the detailed involvement as does the process for large-scale testing, the design process for large-scale tests provides an important model for consideration at the small-scale level, particularly for alignment and content evidence to increase validity.

ALIGNMENT

In the current climate of standards-based testing, alignment is defined as "the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (Webb, 1997, p. 4). The expectations to which assessments need to be aligned in large-scale testing are defined as the systemic reform goals for student

learning as established by national standards and frameworks. In describing the development of statewide assessments for mathematics and science in four states, Webb (1999) reported that reviewers were first sent copies of the content standards and asked to assign a depth-of-knowledge level for each objective. Once the reviewers reached consensus on these levels, then they proceeded to develop items aligned to both the content and the depth-of-knowledge level. To further examine the alignment, items were judged by four criteria on whether or not (a) they addressed the same content as found in the standards, (b) the cognitive demand (depth-of-knowledge) was equal to what students were expected to know and do as stated in the standards, (c) the breadth of knowledge necessary to correctly answer the item corresponded to the span of knowledge in the standards, and (d) the depth and breadth of items were evenly distributed across objectives. These four criteria have been adopted by others (Martineau, Paek, Keene, & Hirsch, 2007; Wise, 2004) as alignment constructs to examine the content standards across grade levels. In other words, most content standards state what a proficient student must know and be able to do at each grade level relative to a given concept. Once alignment of content standards is in place, as it is in most national content standards such as science (National Research Council, 1996) and mathematics (National Council of Teachers of Mathematics, 2000), integrating alignment into test blueprints strengthens the validity evidence of measures of achievement and "provides a solid foundation for the meaning of the score scale and individual criterion-referenced performance levels" (Martineau et al., 2007, p. 34).

Tests and a curriculum framework or standards that are in alignment will work together to provide a common understanding of what students are to learn, what instructional implications are indicated, provide a fair measure of achievement for all students. Alignment is intimately related to test validity (Webb, 1997), which is "the most fundamental consideration in developing and evaluating tests" (Joint Committee on Standards for Educational and Psychological Testing of the AERA the APA and the NCME, 1999, p. 9)

VALIDITY

In recent years, the concept of validity has moved from the traditional view that there are several different types of validity to a unitary concept of validity that requires "an evaluation of the degree to which interpretations and uses of assessment results are justified by supporting evidence and in terms of the consequences of those interpretations and uses" (Miller, Linn, & Gronlund, 2008, p. 73). This concept of validity is represented in the standards for the testing profession jointly published by the American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (Joint Committee on Standards for Educational and Psychological Testing of the AERA the APA and the NCME, 1999). Content evidence in support of validity is the degree to which items of an assessment instrument are relevant and representative with respect to the targeted content domain for a particular assessment purpose (Messick, 1993) and directly affects the validity of the inferences that can be drawn from the resulting data. As Cronbach (1971) stated, "When the items of a test are judged to adequately represent well-defined domains of content, it is permissible to view responses to these items as generalizable samples of the responses examinees would exhibit if they were tested on all of the items constituting these domains" (p. 6).

Bridge and colleagues (2003) suggest four primary principles in the development of content-based evidence for validity (see p. 415). First, the test must have an a priori purpose and measure what it purports to measure. Second, the test items must be representative of all possible items in the domain of interest. Third, the test items must be clearly written and of high quality. Fourth, an expert in the field for each content area covered by the test should review the test items, table of specifications and test blueprint for representativeness, accuracy and quality. Often, teachers not trained in test validity will construct tests that have high face validity—appearing to the teacher and the students to measure an intended outcome—but which lack the rigorous process of formal validation associated with a high degree of validity (Bridge et al., 2003; Haynes, Richard, & Kubany, 1995). Applying a systematic approach to test development by adhering to the four principles stated above will most likely yield a test with sound psychometric properties.

TEST BLUEPRINT CONSTRUCTION

In today's educational research arena, the ultimate goal of teacher professional development is the improvement of student learning (Desimone, 2009; Douglas, 2009). Like many other professional developers, the goals of our several programs are aligned with the research-based conceptual framework suggested by Desimone (2009) for studying the effects of professional development on teachers and students. And also like others, we treat the professional development as an intervention as we examine elements such as increased teacher knowledge and skills, changes in attitudes and beliefs, changes in instruction, and impact on student learning.

To examine impacts on student learning, one important measure is the content knowledge of interest. One problem with using state-level criterion referenced tests (CRTs) as often suggested by grantors as a measure the impact of teacher professional development on student achievement is first, many states do not test science at each grade level; second, it is difficult to obtain previous year's CRT scores for classroom groups as a pretest comparison; and third, most CRTs are not vertically scaled across grade levels, thus rendering CRT scores a poor choice to begin with. The science teachers attending our professional development programs often teach different grade levels and different science disciplines, which complicates the development of a single test for their students that may reflect the impact of the teachers' improved instructional practice as an effect of the professional development intervention. Instead, several tests must be developed that address the selected content domain at each grade level for every teacher in the program. Test blueprints offer a way to develop uniform testing tasks as much as possible so data can be compared across teachers' classrooms. The first step in this process is to develop the blueprints that specify the level of the content and cognitive domains, and the item type and difficulty for each item on the test. Through a rigorous process of alignment and with attention to the criteria suggested by Webb (1999) listed above, each test item is then developed. Each of these steps is covered in more detail in the sections that follow.

CONTENT DOMAIN

A hierarchy of ideas or concepts resides within any grade level content domain that constitutes all that should be known and understood by a proficient student learner at a

given grade level. Some concepts and tasks necessary for this understanding are more important than others, and should therefore receive more emphasis. The standards documents at the national and state levels typically do not prioritize concepts or even disaggregate the standards into subordinate ideas, so it is left to the teacher or curriculum specialist to unpack the standards and place emphasis where it is most needed, which can be a daunting task. Using a framework for this process is vital. The conceptual framework for the content levels we used in test blueprint construction was taken from the work of Wiggins and McTighe (2005) who have identified three levels within the content domain: (a) worth being familiar with; (b) important to know and do; and (c) big ideas or enduring understandings. This framework is presented visually by Wiggins and McTighe as three nested ovals (see p. 71) and is explained in more detail below.

Worth being familiar with. This is the largest oval and represents the knowledge that students should be familiar with from the entire field of all possible content within a given domain and grade level (see Wiggins & McTighe, 2005). Much of the knowledge at this level will be encountered by students through hearing, reading, researching, or viewing information about it such that it will be associated with the big idea as explained below and will be included in test items.

Important to know and do. This is the middle oval. Knowledge within this oval is more focused on concepts that will directly connect to the big idea and enable transfer to related knowledge and other big ideas. The knowledge at this level is prerequisite to understanding the big ideas within a unit of study. Associated with this level are the skills and performances that are also prerequisite and are directly assessed.

Big ideas or enduring understandings. This is the innermost oval. Big ideas in a content domain are unifying concepts that help the learner connect discrete subordinate ideas and promote understanding and transfer (Bransford, Brown, & Cocking, 1999; Bybee, 2002), and in the science discipline, the key ideas are those that are essential to scientific literacy (Rutherford & Ahlgren, 1989). A unifying concept or big idea could be thought of as a keystone in the construction of an archway. All the subordinate stones in the arch are held in place by the single keystone at the top. Without this keystone, the structure of the arch would fail and all that would remain would be a pile of discrete stones. Big ideas have the power to explain phenomena (Wynn & Wiggins, 1997), provide the basis for dealing with new problems (Bloom, Madaus, & Hastings, 1981), and are broad and abstract, thus requiring deep probing because their meanings and value are rarely obvious to the novice learner (Erickson, 2001; Wiggins & McTighe, 2005).

The process of organizing concepts into these three levels is an important step in constructing a test blueprint that will be described in detail in a later section. Equal in importance is an understanding of the cognitive domain.

COGNITIVE DOMAIN

The cognitive domain "deals with the recall or recognition of knowledge and the development of understandings and intellectual abilities and skills" (Reigeluth & Moore, 1999, p. 52). Bloom's work (1956) on the taxonomy of the cognitive domain has provided a common language for educators that is often used for identifying and classifying the knowledge level of educational goals. Bloom's Taxonomy identifies six levels of knowledge with recall at the lowest level and evaluation at the highest. More

recently, Bloom's Taxonomy has been revised into a two-dimensional 4-row x 6-column framework focusing on knowledge level and cognitive processes (Anderson & Krathwohl, 2001). When constructing test blueprints for small-scale assessments, subdividing domains to the *nth* degree does not yield a useful economy of effort and adds unneeded complexity. We therefore turned to the assessment framework for the *Trends in International Mathematics and Science Study* (TIMSS) (Mullis, Martin, Graham J. Rudock, O'Sullivan, & Preuschoff, 2009), which has collapsed Bloom's Taxonomy into three levels in the cognitive domain based on what students must know and do to complete the TIMSS assessment. In the TIMSS framework, each of the levels is referred to as a separate domain as follows:

> The first domain, knowing, covers science facts, procedures, and concepts students need to know, while the second domain, applying, focuses on the ability of the student to apply knowledge and conceptual understanding to a science problem. The third domain, reasoning, goes beyond the solution of routine science problems to encompass unfamiliar situations, complex contexts, and multi-step problems (p. 80).

Within this framework, specific examples and the specifications used to create the TIMSS items are provided, which could also be used to guide the development of learning tasks and hence, test item development. It is these examples, shown in Table 1, which we used to inform the item development for our test blueprints.

Table 1. *Skills and Abilities Specified by the TIMSS Cognitive Domain Framework*

Knowing	Applying	Reasoning
Recall/Recognize	Compare/Contrast/Classify	Analyze
Define	Use Models	Integrate/Synthesize
Describe	Relate	Hypothesize/Predict
Illustrate with Examples	Interpret Information	Design
Demonstrate Knowledge of Scientific Instruments	Find Solutions	Draw Conclusions
	Explain	Generalize
		Evaluate
		Justify

PUTTING IT ALL TOGETHER

Once teachers interacted with the learning tasks intended to increase their understanding of content and cognitive domains using various resources (i.e.., Mullis et al., 2009; Wiggins & McTighe, 2005), they selected the topic for their respective science units. The next step was for the teachers to seek all the information they could about the topic from state core standards documents and from national standards and reform documents such as *Science for All Americans* (Rutherford & Ahlgren, 1989), *National Science Education Standards* (NRC, 1996), *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993) and from district-adopted science texts, other teaching materials, and their own content knowledge. Working in teams of two or three teachers, the groups next listed all the concepts they found within the content domain relative to their chosen topic. The concepts were listed as concept statements (Victor & Kellough, 2004) as in the following partial list of concepts relative to the topic of friction (see p. 455):

1. Friction is the force that resists the movement of one material over another material. Whenever two materials rub against each other, friction is produced.
2. Friction can make it difficult to push one material across another material.
3. Friction is caused by the attraction (adhesion) of the molecules of one surface to the molecules of another as the surfaces rub together.
4. Friction does not depend on speed.
5. The direction of friction force is always in a direction opposing motion.
6. There are three major kinds of friction: sliding, rolling, and fluid.
7. Sliding or kinetic friction is the resistance produced by two surfaces sliding across each other.
8. Rolling friction is the resistance produced when a rolling body moves over a surface.
9. Fluid friction or viscosity is the resistance produced between moving fluids or between fluids and a solid.

Once the groups had completed their lists, they had identified between 15 and 25 concept statements for each unit. The next step was to organize the lists into a coherent sequence and then into the content domains using an outline format. Teachers started by first selecting the concept statements that would fit within the innermost oval, the big ideas or enduring understandings (#1 and #6), which became the highest level of the outline. Next they moved outward to the middle oval to those concepts that were important to know and do (#2, #3, #7, #8, #9), followed by those worth being familiar with (#4), which became the final two levels of the outline. Any concepts that did not fit into one of these three groupings were discarded as being outside the realm of what a proficient student should be responsible for understanding at that particular grade level (#5). With the science concepts organized in outline format showing the content progression by the levels of the content domain, it was now clear exactly what would be taught and in what order. The role of the concept statements was to provide the content goals for a given lesson.

To establish the more global learning objectives, teachers could look at the structure of the content by sequence and level of content domain and construct the objectives. For example, using #6 - #9 the objective might be: *Students will understand the three types of friction by using a variety of material to investigate, compare and contrast the differences in amount of resistance produced in each of the three types.* The teachers reviewed each other's concept statement lists and learning objectives, comparing the ideas to the scope of the various standards and reference documents, as did the content experts on our instructional team. The teachers in concert with the instructional team decided that ideal specifications for the test blueprints might be 50% of the points on the test that addressed the big ideas, 30% that addressed important to know and do, and 20% that addressed worth being familiar with. A test with 25 points possible would then have 12 points addressing big ideas, 8 points addressing important to know and do, and 5 points addressing worth being familiar with. With this process of tight alignment concluded, the next step was to design the test items, taking into consideration the cognitive domain and "how well the sample of assessment task represents the domain of tasks to be measured and how it emphasizes the most important content" (Miller et al., 2008, p. 74) in order to provide strong evidence for validity.

Cognitive domain. Using the target percentages specified in the TIMSS devoted to each of the cognitive domains (Mullis et al., 2009), 35 % of the possible points on our

tests would be devoted to the knowing domain, 35% to the applying domain, and 30% to the reasoning domain. Therefore, the target test of 25 possible points for each unit would include 9 points dedicated to knowing, 9 points to applying, and 7 to reasoning.

Test items. In order for test items to contribute to a high degree of test validity, they must be adequately challenging in terms of depth of content knowledge and cognitive expectations as appropriate to grade level and avoid a host of threats to validity. Metzenberg (Metzenberg, 2004, n.d.) provides an extensive comparison of test items from science and mathematics state tests that illustrate the inherent threats to validity. The greatest threat he discusses is the depth of content knowledge, finding that many items at the high school level test knowledge that should be mastered at the middle school level. Although the state or national standard to which a given item is aligned may represent challenging expectations, the actual item often falls short of that challenge. Other threats he discusses include technical defects that artificially boost student performance such as cueing by including the same word in both the stem and the correct answer in multiple-choice items, or by using a different writing style for the correct answer choice. Just-in-time teaching is another threat to item validity that teaches the content in the stem of the item before the question is asked and answer choices provided.

Once teachers were facilitated in learning and understanding these common threats to item validity, the test blueprints were finalized and item construction began. The test blueprints developed by the teachers in each program are provided in the next section.

EXAMPLES

To illustrate the process of using test blueprints and to show the resulting possibilities for data disaggregation and analysis, two examples are provided below. Within each of the programs, the teachers participated in what is described by Wiggins and McTighe (2005) as backward design wherein the learning goals (concept statements and objectives) are established first, assessments are developed next and are designed to measure the learning goals, and finally, the learning activities are selected or developed. The development of test blueprints occurs within the second step of backward design.

The first professional development program example explained below, funded by the National Science Foundation (#0341925), was a pilot study designed to facilitate middle school teachers in the integration of engineering into their science curriculum. The second example, also with middle school science teachers, was funded by the office of technology within the state department of education in a western state. This program facilitated teachers in the integration of technology into their science curriculum.

SCIENCE AND ENGINEERING INTEGRATION

The 8 teachers in this pilot program were taught engineering content and principles along with the engineering design process as an instructional intervention. They then collaborated in small groups to develop an integrated unit with an end of unit test constructed using a test blueprint, which they administered to their students. A total of three units, each addressing a different topic, were developed. Prior to developing the unit tests, the teachers were provided instruction and information about the content and cognitive domains and test item validity as described above. Discussion and development of the test blueprint was the next step. Although the level of cognitive

domain is nearly synonymous with the knowledge learning degree of difficulty, a range of difficulty level is expected for test items for each domain (Mullis et al., 2009), so the teachers included both on their blueprint, which is shown in Table 2.

The next task for the teachers was to develop the items according to the specifications on the test blueprint. The focus for the units was on science content knowledge using engineering design pedagogy to teach the science ideas, so the test items were developed exclusively from the science content domain. During the item development process, the teachers decided to adjust the test blueprint according to their needs, so the finalized blueprint deviated slightly from the ideal percentages they originally specified. The test items were reviewed by teacher peers and by content and assessment experts from the instructional team. Once teachers taught their units, they administered the tests to their students and returned the test protocols to the instructional team for data entry and analysis.

Table 2. *Science and Engineering Integrated Module Test Blueprint*

Item #	Type	Content Level	Cognitive Level	Difficulty	Points
1	Multiple choice	Important	Knowing	Easy	1
2	Multiple choice	Worth	Knowing	Easy	1
3	Multiple choice	Big Idea	Knowing	Easy	1
4	Multiple choice	Important	Applying	Moderate	1
5	Multiple choice	Important	Applying	Moderate	1
6	Multiple choice	Worth	Applying	Moderate	1
7	Short Answer	Important	Knowing	Easy	1
8	Short Answer	Important	Knowing	Easy	1
9	Short Answer	Important	Knowing	Easy	1
10	Short Answer	Worth	Applying	Moderate	1
11	Short Answer	Worth	Applying	Moderate	1
12	Short Answer	Important	Applying	Moderate	1
13	True/False	Big Idea	Knowing	Easy	1
14	True/False	Big Idea	Applying	Easy	1
15	True/False	Worth	Reasoning	Easy	1
16	True/False	Worth	Applying	Moderate	1
17	True/False	Worth	Applying	Moderate	1
18	True/False	Big Idea	Reasoning	Moderate	1
19	Academic Prompt	Big Idea	Reasoning	Difficult	7
Point Totals		Worth = 7 (28%) Important = 7 (28%) Big Idea = 11 (44%)	Knowing = 7 (28%) Applying = 9 (36%) Reasoning = 10 (40%)		

Scores for each item on the test were entered into the data set along with demographic variables for each of the 441 students in the study. Individual test items were then summed into six subscales, three for the content domain (important, worth, and big idea) and three for the cognitive domain (knowing, applying and reasoning). Within each of these subscales, the data could be further disaggregated by the levels of the demographic variables (Table 3) and used for a variety of different statistical analysis procedures. The results would provide much more in-depth information about

students' performance than simply a single test score. For this study, student performance was compared to statewide performance scores on the 8th grade science CRT by demographic variables .

Table 3. *Test Results Disaggregated by Content and Cognitive Domains*

	Content Domain					
	Worth Being Familiar With		Important to Know and Do		Big Ideas	
Independent Variables	*M*	*SD*	*M*	*SD*	*M*	*SD*
Gender						
Males (n=243)	4.60	1.55	4.56	1.64	7.00	2.69
Females (n=171)	4.38	1.69	4.44	1.71	6.73	2.70
Qualified for Special Education						
Yes (n=42)	4.38	1.50	3.79	1.66	4.98	2.83
No (n=327)	4.53	1.62	4.59	1.65	7.10	2.60
Low Socioeconomic Status						
Yes (n=99)	4.06	1.81	3.83	1.68	6.41	3.08
No (n=314)	4.66	1.52	4.72	1.61	7.03	2.56
Ethnicity						
Black (n=11)	4.00	1.79	3.18	1.66	7.27	3.98
Hispanic (n=55)	4.22	1.51	3.96	1.64	6.56	2.92
Asian/Pacific (n=20)	5.20	2.31	4.95	1.90	7.65	2.60
American Indian (n=19)	3.84	1.68	3.32	1.34	5.74	2.64
White (n=308)	4.58	1.54	4.74	1.60	6.96	2.60
	Cognitive Domain					
	Knowing		Applying		Reasoning	
Gender						
Males (n=243)	4.44	1.56	6.13	1.90	5.59	2.51
Females (n=171)	4.52	1.79	5.79	2.02	5.24	2.45
Qualified for Special Education						
Yes (n=42)	3.59	1.64	5.69	1.66	3.86	2.45
No (n=327)	4.57	1.63	6.02	1.99	5.63	2.43
Low Socioeconomic Status						
Yes (n=99)	3.96	1.58	5.26	2.08	5.08	2.82
No (n=314)	4.64	1.65	6.22	1.86	5.56	2.37
Ethnicity						
Black (n=11)	3.64	1.21	5.00	2.23	5.82	3.31
Hispanic (n=55)	4.18	1.66	5.49	1.84	5.07	2.80
Asian/Pacific (n=20)	5.30	1.34	6.60	2.74	5.90	2.61
American Indian (n=19)	3.58	1.64	4.68	1.57	4.63	2.34
White (n=308)	4.56	1.66	6.16	1.88	5.54	2.39

Science and Technology Integration. Thirty-eight middle school teachers and their students from five rural school districts in a western state participated in this study. Each teacher was given a technology package that included a laptop computer, an LCD projector, an electronic microscope, a flash drive, and a one-year high speed Internet connection at home. A significant portion of the professional development time was

spent with engaging teachers in learning experiences to increase their knowledge and skills relative to the use of computer-based technology and Internet-based learning objects that could be integrated into their science instruction.

A total of 22 lesson were developed by the teachers working in teams of three or more, each from a different school in order to maximize the number of school settings across which data on the same lesson could be collected, and to capitalize on the potential power of social interaction among the teachers (Vygotsky, 1978). Teachers selected topics for the lessons from their regular curriculum. The same rigorous process was used for test blueprint construction (see Table 4) and for lesson development as described in the previous example. Based on the technology instruction provided during the course, teachers designed an integrated technology component for each lesson. Teachers had full control over the type and the content of the technology component, with the only parameter being that the technology component must function as a tool for teaching and learning the science content and not for the technology to *be* the content. Once lessons and assessments were completed, they were made available online so that individual teachers could select and download the two

Table 4. *Science and Technology Integrated Unit Test Blueprint*

Item #	Type	Content Level	Cognitive Level	Difficulty	Points
1	Multiple choice	Worth	Knowing	Easy	1
2	Multiple choice	Worth	Applying	Easy	1
3	Multiple choice	Important	Knowing	Easy	1
4	Multiple choice	Worth	Applying	Moderate	1
5	Multiple choice	Big Idea	Knowing	Easy	1
6	Multiple choice	Important	Applying	Moderate	1
7	True/False	Important	Knowing	Easy	1
8	True/False	Important	Applying	Easy	1
9	True/False	Worth	Knowing	Moderate	1
10	True/False	Big Idea	Applying	Moderate	1
11	True/False	Big Idea	Knowing	Easy	1
12	True/False	Important	Applying	Moderate	1
13	Short Answer	Important	Reasoning	Moderate	1
14	Short Answer	Important	Applying	Easy	1
15	Short Answer	Important	Knowing	Difficult	1
16	Short Answer	Worth	Applying	Moderate	1
17	Short Answer	Big Idea	Knowing	Moderate	1
18	Short Answer	Big Idea	Reasoning	Moderate	1
19	Academic Prompt	Big Idea	Reasoning	Difficult	7
Point Totals		Worth = 5 (20%) Important = 8 (32%) Big Idea = 12 (48%)	Knowing = 8 (32%) Applying = 8 (32%) Reasoning = 9 (36%)		

lessons they would teach during the school year. Half the teachers' class periods were randomly selected for the control group and half for the treatment group. Both groups

were taught the same lesson; however, the treatment group's lesson contained the integrated technology component. The teachers taught two lessons during the year of the study and for the second lesson, the treatment and control groups were reversed. This methodology yielded a total of more than 10,000 student tests that were entered into the data set and analyzed. Tables 5 and 6 show the descriptive statistics disaggregated by levels of independent variables and by levels of the content and cognitive domains.

Table 5. *Science and Technology Test Results Disaggregated by Content Domain and Group*

		Content Domain					
		Worth Being Familiar With		Important to Know and Do		Big Ideas	
Demographic Variables		*M*	*SD*	*M*	*SD*	*M*	*SD*
Gender							
Males	n=2518	3.45	1.22	5.48	1.76	7.20	3.13
	n=2531	(3.37)*	(1.24)	(5.26)	(1.82)	(7.11)	(3.15)
Females	n=2220	3.43	1.21	5.43	1.76	7.33	1.71
	n=2537	(3.40)	(1.22)	(5.32)	(1.82)	(7.20)	(1.70)
Qualified for Special Education							
Yes	n=642	2.86	1.32	4.64	1.81	5.70	3.11
	n=644	(2.75)	(1.24)	(4.26)	(1.79)	(5.41)	(2.96)
No	n=4222	3.51	1.17	5.57	1.69	7.54	2.99
	n=4324	(3.49)	(1.20)	(5.42)	(1.76)	(7.49)	(3.06)
Low Socioeconomic Status							
Yes	n=1650	3.17	1.23	5.01	1.77	6.43	3.09
	n=1680	(3.09)	(1.25)	(4.81)	(1.84)	(6.29)	(3.11)
No	n=3056	3.59	1.19	5.65	1.75	7.70	2.97
	n=3115	(3.54)	(1.20)	(5.65)	(1.68)	(7.64)	(3.05)
Ethnicity							
Black	n=118	3.25	1.24	5.24	1.79	6.70	3.50
	n=119	(3.23)	(1.11)	(4.83)	(1.77)	(7.26)	(3.26)
Hispanic	n=780	3.30	1.20	5.13	1.77	6.88	3.10
	n=829	(3.16)	(1.26)	(4.95)	(1.89)	(6.57)	(3.25)
Asian/Pacific	n=119	3.38	1.24	5.55	1.75	7.69	2.87
	n=115	(3.45)	(1.36)	(5.75)	(1.83)	(7.93)	(3.18)
Am. Indian	n=155	3.13	1.24	4.95	1.68	6.55	2.93
	n=162	(3.02)	(1.22)	(4.56)	(1.78)	(6.27)	(3.20)
White	n=3682	3.50	1.21	5.53	1.72	7.43	3.04
	n=3725	(3.46)	(1.22)	(5.38)	(1.77)	(7.39)	(3.07)

*Control group results are in parentheses.

Table 6. *Science and Technology Test Results Disaggregated by Cognitive Domain and Group*

		Cognitive Domain					
		Knowing		Applying		Reasoning	
Independent Variables		*M*	*SD*	*M*	*SD*	*M*	*SD*
Gender							
Males	n=2518	5.66	1.70	5.40	1.74	5.07	2.81
	n=2531	(5.56)*	(1.77)	(5.21)	(1.76)	(4.94)	(2.81)
Females	n=2220	5.64	1.67	5.39	1.68	5.19	2.71
	n=2537	(5.62)	(1.71)	(5.25)	(1.76)	(5.16)	(2.79)
Qualified for Special Education							
Yes	n=642	4.80	1.85	4.57	1.77	3.81	2.74
	n=644	(4.64)	(1.80)	(4.25)	(1.69)	(3.53)	(2.71)
No	n=4222	5.78	1.61	5.52	1.67	5.35	2.71
	n=4324	(5.72)	(1.69)	(5.37)	(1.72)	(5.28)	(2.74)
Low Socioeconomic Status							
Yes	n=1650	5.24	1.72	4.98	1.70	4.39	2.77
	n=1680	(5.10)	(1.80)	(4.81)	(1.78)	(4.25)	(2.79)
No	n=3056	5.84	1.64	5.61	1.69	5.49	2.69
	n=3115	(5.82)	(1.66)	(5.45)	(1.72)	(5.42)	(2.73)
Ethnicity							
Black	n=118	5.30	1.88	5.13	1.74	4.77	3.01
	n=119	(5.43)	(1.95)	(4.91)	(1.56)	(4.97)	(2.84)
Hispanic	n=780	5.43	1.71	5.09	1.69	4.79	2.80
	n=829	(5.33)	(1.80)	(4.85)	(1.80)	(4.50)	(2.88)
Asian/Pacific	n=119	5.60	1.80	5.55	1.72	5.48	2.53
	n=115	(5.70)	(1.77)	(5.63)	(1.95)	(5.77)	(2.77)
Am. Indian	n=155	5.21	1.64	4.85	1.73	4.55	2.64
	n=162	(5.00)	(1.78)	(4.61)	(1.73)	(4.24)	(2.82)
White	n=3682	5.73	1.66	5.48	1.70	5.24	2.75
	n=3725	(5.67)	(1.70)	(5.34)	(1.73)	(5.20)	(2.75)

*Control group results are in parentheses.

DISCUSSION AND IMPLICATIONS

Using test blueprints for professional development studies has several advantages. First, although teachers involved in a given program may be teaching different grade levels and different topics, developing the unit test using a blueprint means that item #10, for instance, as listed on the blueprint, may specify a short answer of moderate difficulty drawn from the worth being familiar with level of the content domain and from the applying level of the cognitive domain regardless of grade level or content topic. As each teacher group adheres to the specification for this item when developing the unit test, the assumption is that any student taking any of the unit tests prepared using the blueprint would confront the same cognitive load relative to that grade level to answer the question correctly. This uniformity of item specifications then allows the item response to be compared to the same item response on all the tests regardless of the content topic with a greater degree of confidence in the validity of the interpretation of the scores.

Another advantage is the opportunity for data disaggregation. Not only can the test scores be disaggregated as usual by the levels of the demographic variables, but researchers can further examine how students representing each of these levels perform within the levels of both the content and cognitive domains. This capability may bring to light much more subtle but important differences. For instance, in one of our data sets, male students outscored female students on the overall mean test score, but when disaggregated data were analyzed it was noted that female students outscored males in understanding the big ideas that were tested. This information is important for both researchers in terms of more accurately interpreting data analysis results, and for the classroom teacher who now knows how classroom instruction might be adjusted to address performance gaps relative to specific student populations.

Another advantage is the long-term effect that learning how to construct test blueprints can have on teachers and by extension, on their students. The process of reviewing content and examining alignment with state and national standards and frameworks in preparation for writing test items provides the teachers with a solid content review that may enhance their grade level content knowledge. Many teachers who have learned this skill of test blueprint development in our professional development programs have reported that they have revisited previous tests and improved them using the test blueprint methodology. Among those teachers who have shared their experiences about test blueprints with us, most also comment that the rigor of their tests has improved and their expectations for student learning have increased. While this result is gratifying, it is also important to note that the teachers rarely enter the test data item by item because of the time factor, and as a result, they look only at the overall test scores.

For researchers, there are several issues to consider. Our study in the second example above yielded over 10,000 student tests, each item of which had to be entered by hand by our research assistants, which was a daunting task. The total data set contains over 300,000 data points. With today's technology, having students take the tests on computers would reduce the necessity of hand entry except for written responses requiring hand grading, but the rural schools in our study were not equipped with adequate computer labs or broadband internet connections that allowed students to do so.

Extra time during professional development sessions must also be set aside to engage teachers in learning experiences relative to the content and cognitive domains and test item validity, and the process of developing the test blueprint can be tedious. But we feel this extra time and effort is an investment that can pay positive dividends in terms of teacher thinking, beliefs, and ultimately in improved instruction, which has a direct impact on student learning (Desimone, 2009).

There are indeed many ways to measure student performance as an impact of teacher professional development (Douglas, 2009). This article has reviewed a systematic approach for facilitating teachers in developing clear and concise test blueprints for end of unit tests that measure learners' achievement relative to science content tightly aligned with local and national content standards and frameworks. These types of tests are very useful at the researcher level in measuring the impact of teacher professional development on student learning and at the teacher level for improving and adjusting instruction to better meet the needs of individual student populations.

REFERENCES

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

American Educational Research Association. (2009). Using multiple methods with randomized controlled trials: National forum on multiple methods yields new online product. *Educational Researcher, 38*(3), 228.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18-20.

Bloom, B. (Ed.). (1956). *Taxonomy of educational objectives: Classification of educational goals. Handbook I: Cognitive domain*. New York: Longman, Green & Co.

Bloom, B., Madaus, G., & Hastings, J. T. (1981). *Evaluation to imporve learning*. New York: McGraw-Hill.

Bransford, J., Brown, A., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washigton, DC: National Academy Press.

Bridge, P. D., Musial, J., Frank, R., Roe, T., & Sawilowsky, S. (2003). Measurement practices: Methods for developing content-valid student examinations. *Medical Teacher, 25*(4), 414-421.

Bybee, R. (2002). *Learning science and the science of learning*. Washington, DC: National Science Teachers Association.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better coceptualizations and measures. *Educational Researcher, 38*(3), 181-199.

Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational Researcher, 38*(7), 518-521.

Erickson, L. (2001). *Stirring the head, heart and soul: Redefining curriculum and instruction* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8), 4-14.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment" Functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247.

Henson, R. K., Hull, D. M., & Williams, C. S. (2010). Methodology in our education research culture: Toward a stronger collective quantitative proficiency. *Educational Researcher, 39*(3), 229-240.

Howe, K. (2008). Isolating science from the humanities: The third dogma of educational research. In M. Giardina & N. Denzin (Eds.), *Qualitative research and the politics of evidence* (pp. 565-579). Walnun Creek, CA: Left Coast Press.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Joint Committee on Standards for Educational and Psychological Testing of the AERA the APA and the NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Martineau, J., Paek, P., Keene, J., & Hirsch, T. (2007). Integrated, comprehensive alignment as a foundation for measuring student progress. *Educational Measurement: Issues and Trends, 26*(1), 28-35.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (2nd ed.). Phoenix, AZ: American Council on Education and Oryx Press.

Metzenberg, S. (2004). Science and mathematics testing: What's right and wrong with the NAEP and TIMSS. In W. M. Evers & H. J. Walberg (Eds.), *Testing student learning, evaluating teacher effectiveness*. Stanford, CA: Hoover Institution Press.

Metzenberg, S. (n.d.). Improving state science assessments. http://escience.ws/stm/StateScienceAssess.pdf

Miller, D. M., Linn, R. L., & Gronlund, N. E. (2008). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Prentice Hall.

Mullis, I. V. S., Martin, M. O., Graham J. Rudock, O'Sullivan, C. Y., & Preuschoff, C. (2009). TIMSS 2011 Assessment Framework. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement.

National Council of Teachers of Mathematics. (2000). *Curriculum and evaluation standards for school mathematics* (3rd ed.). Reston, VA: National Council for Teachers of Mathematics.

National Research Council. (1996). *National science education standards*. Washington, D.C.: National Academy Press.

NRC. (1996). *National science education standards*. Washington, D.C.: National Academy Press.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2007). TIMSS Technical Report 2007. Boston: International Study Center Boston College.

Reigeluth, C. M., & Moore, J. (1999). Cognitive education and the cognitive domain. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory* (Vol. II). Manwah, NJ: Lawrence-Erlbaum Assoiates, Inc., Publishers.

Rutherford, F. J., & Ahlgren, A. (1989). *Science for all Americans*. New York: Oxford University Press.

Slavin, R. E. (2008). Perspectives on evidence-based research in education: What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14.

Tashakkori, A., & Dreswell, J. W. (2008). Editorial: Envisioning the future stewards of the social-behavioral research enterprise. *Journal of Mixed Methods Research, 2*(4), 291.

Victor, E., & Kellough, R. D. (2004). *Science k-8: An integrated approach* (10th ed.). Saddle River, NJ: Pearson.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Camebridge, MA: Harvard University Press.

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Madison, WI: National Institute for Science Education; Council of Chief State School Officers Washington, DC.

Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. Madison, WI: National Institute for Science Education; Council of Chief State School Officers Washington, DC.

Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wise, L. (2004). Vertically articulated content standards. Retrieved from http://www.nciea.org/publications/RILS_LW04.pdf

Wynn, C. M., & Wiggins, A. W. (1997). *The five biggest ideas in science*. New York: John Wiley & Sons.