J Sci Teacher Educ (2013) 24:269–295
DOI 10.1007/s10972-012-9299-7

Valid and Reliable Science Content Assessments for Science Teachers

Thomas R. Tretter · Sherri L. Brown · William S. Bush · Jon C. Saderholm · Vicki-Lynn Holmes

Published online: 4 June 2012

Abstract Science teachers' content knowledge is an important influence on student learning, highlighting an ongoing need for programs, and assessments of those programs, designed to support teacher learning of science. Valid and reliable assessments of teacher science knowledge are needed for direct measurement of this crucial variable. This paper describes multiple sources of validity and reliability (Cronbach's alpha greater than 0.8) evidence for physical, life, and earth/space science assessments—part of the Diagnostic Teacher Assessments of Mathematics and Science (DTAMS) project. Validity was strengthened by systematic synthesis of relevant documents, extensive use of external reviewers, and field tests with 900 teachers during assessment development process. Subsequent results from 4,400 teachers, analyzed with Rasch IRT modeling techniques, offer construct and concurrent validity evidence.

T. R. Tretter (✉) · W. S. Bush
Department of Middle and Secondary Education, University of Louisville, Louisville, KY 40292, USA
e-mail: tom.tretter@louisville.edu

W. S. Bush
e-mail: bill.bush@louisville.edu

S. L. Brown
Department of Early Childhood and Elementary Education, University of Louisville, Louisville, KY 40292, USA
e-mail: s.brown@louisville.edu

J. C. Saderholm
Berea College, Berea, KY, USA
e-mail: jon_saderholm@berea.edu

V.-L. Holmes
Hope College, Holland, MI, USA
e-mail: holmesv@hope.edu

Keywords Science teacher content assessments · Teacher content knowledge · Valid assessments · Reliable assessments

Introduction

The No Child Left Behind (NCLB) Act of 2001 (NCLB 2001) required states to ensure that all teachers of core academic subjects in public schools be highly qualified by the end of the 2005–2006 school year. However, a shortage of certified science teachers in the United States (Ingersoll 2000) has led many states to develop alternative routes to teaching, including attempts to convert teachers from content areas of over-supply into science teachers. Such teachers may not have the science content background to be highly qualified for teaching science, generating the need for programs and interventions to enhance science content knowledge for groups of teachers. Middle school science teachers, in particular, do not usually teach in only one or two science content domains; they often teach many topics across the science domains of physical, life, and earth/space sciences. According to the US Department of Education, during the 1999–2000 school year, almost half the nation's middle school science teachers did not have a major in their subject (US Department of Education 2004).

Laczko-Kerr and Berliner (2002) highlighted the importance of science certification by comparing student achievement in regularly certified teachers' classrooms and in under-certified teachers' classrooms. They found that students of under-certified teachers scored approximately 20 % less academic growth per year when compared to students of regularly certified teachers. A report from the National Academies (2006) also linked poor science student achievement with the lack of highly qualified science teachers and recommended improving science education by providing high-quality summer professional development (PD) programs for science teachers. In order to develop and retain well-prepared science teachers (Darling-Hammond 2000), districts, science teacher educators, researchers, professional development providers, and instructors need instruments to be able to assess the effectiveness of various interventions and programs on the development of teachers' science content knowledge.

The purpose of this paper is to offer the validity and reliability evidence for teacher science content assessments developed as part of the Diagnostic Teacher Assessments of Mathematics and Science (DTAMS) project. A total of three separate assessments—focused on physical science, life science, and earth/space science—have been developed. This paper is organized around two major phases in developing and establishing the validity and reliability of these assessments: the process of assessment development which included strategies designed to strengthen the validity and reliability of score interpretations, and empirical results from approximately 4,400 teachers.

Structure of the Assessments

These assessments were developed to target the science knowledge needed by middle school teachers. The three independent science teacher assessments (physical, life, and earth/space) have several common elements: (a) structured to provide information about both depth and breadth of knowledge; (b) relatively brief—taking <1 h per assessment to complete; (c) consist of a blend of multiple-choice and open-response questions; and (d) available in multiple, statistically similar forms.

Depth and Breadth of Knowledge

Both depth and breadth of middle school teachers' science content knowledge are measured and reported as subscales by these assessments. Depth of content knowledge refers to various types of understanding. For example, on a lower level, a teacher may have memorized a definition or understands terminology; on a higher level, a teacher could explain the relationship of the concept to other phenomena or apply concepts to new situations. Breadth of knowledge refers to a reasonable spectrum of specific content knowledge within a domain that would adequately represent that domain within the constraints of keeping the assessment length to less than an hour to complete.

Depth of Knowledge

Our taxonomy of depth of teachers' science knowledge was grounded in the work of Li and Shavelson (2001) and Shulman (1986). Li and Shavelson's framework included four types of knowledge: (a) declarative knowledge (scientific definitions and facts); (b) procedural knowledge ("if–then production of rules or a sequence of steps, in the form of actions or steps that can be carried out to achieve a certain goal leading to task completion" p. 6) also called skills of doing science by Stiggins (1997); (c) schematic knowledge ("principles, schemes, and mental models" p. 5); and (d) strategic knowledge ("knowledge of when, where, and how to use certain types of knowledge in a new situation and knowledge of assembling cognitive operation" p. 7). For the second category labeled "procedural knowledge", we chose to use the term "inquiry" to reflect the science standards documents use of that term to represent the skills and habits of mind of doing science.

Shulman's (1986) definition of pedagogical content knowledge is consistent with Li and Shavelson's (2001) strategic knowledge applied to the act of teaching—knowing when, where, and how to use knowledge in teaching situations. We chose to use Shulman's label of pedagogical content knowledge for this fourth type of knowledge in our assessments' depth taxonomy. Pedagogical content knowledge in science is, of course, complex, diverse, and not standardized. Because an important aspect of competent science teaching is to recognize and correct student misconceptions, a decision was made to operationalize this knowledge type by restricting the pedagogy questions to focus on students' misconceptions and subsequent appropriate instructional decisions. See Table 1 for descriptions of each knowledge type.

In addition to the four knowledge types, an additional type of knowledge emerged from the assessment development process. Science education reform documents (American Association for the Advancement of Science (AAAS) 1993;

Table 1 Depth of knowledge descriptions

Label	Description
Declarative	Declarative knowledge describes knowledge that is primarily recall in nature, including reproduction of scientific definitions and facts
Inquiry	Inquiry knowledge represents knowledge of scientific inquiry and procedures, knowing how to generate and evaluate scientific evidence
Schematic	Schematic knowledge represents a deep understanding of science concepts, procedures, laws, theories, principles, and rules. This type of knowledge includes understanding of the connections and relationships among scientific phenomena
Pedagogical	Pedagogical knowledge represents strategic knowledge for teaching, knowing when, where, and how to apply domain-specific knowledge and strategies (pedagogical and scientific) to provide students with learning experiences that will foster effective learning of science

National Research Council (NRC) 1996) included an emphasis on science, technology, and society (STS) knowledge. The STS knowledge type inherently overlapped both of the other dimensions (depth of knowledge and breadth of knowledge) because it applied to all science topics as well as various depths of knowledge. Therefore, STS was conceptualized not as a knowledge depth parallel to the other four, but rather as a third dimension overlaid on top of the two-dimensional grid defined by content categories (breadth) and knowledge types (depth).

Breadth of Knowledge

In order to design assessments to sample across the breadth of science knowledge needed by middle school science teachers, the extent of the relevant knowledge base needed to be identified. The science knowledge base most critical for middle school teachers to know was identified from a synthesis across 8 documents: 3 standards, 2 student assessment frameworks, and 3 teacher content knowledge recommendations. The standards documents included the *National Science Education Standards* (NSES) (National Research Council 1996); *Benchmarks for Science Literacy* (American Association for the Advancement of Science (AAAS) 1993); and Mid-continent Research for Education and Learning's [McREL's] science portion of *Content knowledge: A compendium of standards and benchmarks for K-12 education* (Kendall and Marzano 2004). Student assessment frameworks used to synthesize content recommendations included the *National Assessment for Educational Progress* (NAEP) (National Assessment Governing Board (NAGB) 2004) and *Trends in International Mathematics and Science Study* (TIMSS) (Mullis et al. 2003). Sources of teacher content knowledge recommendations included National Science Teachers Association (NSTA 2003); Interstate New Teacher Assessment and Support Consortium (INTASC) Science Standards Drafting Committee (INTASC 2002); and *The Praxis Series: Professional Assessments for Beginning Teachers* (2004). Saderholm and Tretter (2008) described in detail this process for determining the breadth of knowledge for the physical science assessment, including the process to determine relative weightings of each major content category within a domain.

To establish a consensus of content topics across all these documents, teams of five to seven science educators (from university departments of education), scientists (from university science departments), and middle school science teachers were convened to review the documents in each subject area. Project staff further analyzed the content of the documents to synthesize the consensus topics into major categories and subcategories.

Once the depth of knowledge and breadth of knowledge dimensions had been identified, a two-dimensional grid was generated for each of the three content area assessments. In addition to the major content categories identified for each science assessment, subcategories within each content category were also identified from the synthesis of literature so that items could be generated to intentionally sample across the spectrum of content within each category. See Table 2 for the two-dimensional item specification grid for each of the three assessments.

Item Distribution

Individual items for each assessment were generated to intentionally sample across both dimensions (breadth and depth of knowledge) of the grids in Table 2. Along with being distributed across the three or four major content categories for each assessment, items also were distributed within content categories across subcategories. A quick scan of Table 2 illustrates that the physical science assessment is intentionally designed (see Saderholm and Tretter 2008) to include an unequal number of items per content category (i.e., there are more items in the "energy" category based on the literature synthesis).

In order for assessments to be completed by testtakers in approximately an hour or less, a multiple-choice format was used for items 1–20 (scored 1 point each) measuring all knowledge types except pedagogy. Open-response questions 21–25 each had two parts: part (a) required the identification of the correct science concepts (scored 0 or 1) necessary to address a misconception held by students as expressed in the question stem; part (b) required a detailed pedagogical description of how the teacher would address the misconception using best instructional practices (scored 0, 1, or 2). This resulted in each complete assessment consisting of 20 multiple-choice items and five two-part open-response items with a maximum possible raw score of 35.

Validity from Assessment Development Process

The Standards for Educational and Psychological Testing (American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (US) 1999) described validity as the development of an argument to support the proposed interpretation of test scores. Various types of evidence may be applied to evaluate the validity. Evidence may be based on test

Table 2 Depth and breadth item specification grid showing targeted cells for individual assessment items

Physical Science (all versions)				
	Depth of knowledge			
Breadth of Knowledge (25 content points; 10 PCK points)	Declarative (5 pts)	Inquiry (5 pts)	Schematic (15 pts)	PCK[a] (10 pts)
1. Properties and changes of properties in matter (6 content category points possible)				
i. Physical properties (e.g. density, boiling point), mixtures, physical vs. chemical change	8			
ii. Chemical reactions, conservation of mass, chemical families, compounds		5	23a	23b
iii. Elements and atomic structure			9	
iv. States of matter, kinetic theory, gas laws			13, 16[b]	
2. Motion and forces (6 content points possible)				
i. Position & direction of moving things, speed & velocity, graphical representation	7			
ii. Forces & acceleration (includes friction, Newton's 2nd law, gravity), addition of forces balanced & unbalanced forces, momentum & impulse (includes Newton's third law)	19	12	4, 20[b]	
iii. Newton's 1st law (inertia)			21a	21b
3. Energy (13 content points possible)				
i. Energy as ability to do work/cause change, mechanical energy & transfer, simple machines, kinetic & potential energy, systems & conservation of energy	18		10[b], 24a	24b
ii. Waves, sound, light, color and vision, electromagnetic spectrum, sunlight		17	3, 25a	25b
iii. Static electricity, electric current and circuits, magnetism, electromagnetism		14	11[b], 22a	22b
iv. Heat, temperature, temperature scales, thermodynamics		15	1, 2[b]	
v. Chemical energy, nuclear energy (radioactivity, fusion, fission)	6			

Table 2 continued

Life Science (all versions)				
	Depth of knowledge			
Breadth of Knowledge (25 content points; 10 PCK points)	Declarative (7 pts)	Inquiry (7 pts)	Schematic (11 pts)	PCK[a] (10 pts)
1. Structure/Function of Living Systems (6 content points possible)				
i. Complimentary nature of structure and function, increase in complexity: cell-tissue-organ	7			
ii. Cells		3		5[b]
iii. Animal organ systems	13		11	
iv. Plant organ systems			21a	21b
2. Internal Regulation and Behavior (6 content points possible)				
i. Organisms convert energy (photosynthesis, respiration, metabolism)	1	2		
ii. Homeostasis; behavior from cellular to organ levels			22a	22b
iii. Cellular communication (hormones), immune system, and disease fighting	23a		14[b], 16	23b
3. Reproduction & Heredity, Diversity and Adaptation of Organisms (7 content points possible)				
i. Plant and animal reproduction (sexual & asexual)		8, 10[b]		
ii. Genetics	12	9[b]		
iii. Fitness & survival (nature vs. nurture), adaptations, evolution (change over time), extinction	25a		15	25b
iv. Taxonomy		4		
4. Interdependence of Organisms (Ecology, Populations) (6 content points possible)				
i. Organisms obtain and use resources (consumers, producers, parasitism, mutualism, etc.)	17			
ii. Cycling of nature			6[b], 19, 24a	24b
iii. Populations/ biomes/ ecosystems		20	18	

Table 2 continued

Earth/Space Science (all versions)

Breadth of Knowledge (25 content points; 10 PCK points)	Depth of Knowledge			
	Declarative (8 pts)	Inquiry (6 pts)	Schematic (11 pts)	PCK[a] (10 pts)
1. Atmosphere, Hydrosphere (8 content points possible)				
i. Atmosphere composition & properties	12b	13	11	
ii. Weather/weather hazards/cloud formation	18	3[b], 5		
iii. Climate, oceans & water cycle			14, 22[a]	22[b]
2. Lithosphere (11 content points possible)				
i. Rock/mineral/soil properties	6, 9[b]	20		
ii. Rock/mineral cycle & processes (weathering, erosion, glaciers)	1		19	
iii. Movements of plates/ earthquakes/volcanoes & Earth's layers	2, 7[b]		23[a]	23[b]
iv. History of the Earth: geologic time/uniformitarianism, fossils as evidence		17	15, 21[a]	21[b]
3. Earth in the Solar System (6 content points possible)				
i. Solar system	16			
ii. Moon phases, motion and eclipses		10	25a	25b
iii. Sun as energy source and reason for seasons			4, 8[b]	
iv. Gravity and tides			24a	24b

Note Numbers in the cells are the assessment item numbers on the assessments

[a] Open response (#21b–25b) pedagogical questions part *b* are each scored on 2-point scale

[b] STS item points are already incorporated into the 35-point total for breadth & depth subscales; these 5 items are simply pulled out to report an additional STS subscale

content, response processes, internal structure, and relations to other variables (Goodwin and Leech 2003). Several of these sources of validity evidence are used for establishing validity in the interpretation of scores from these assessments.

Validity of Breadth of Knowledge Categories

Validly establishing the breadth of knowledge categories teachers should know was done by teams of 5–7 science educators, scientists, and middle school science teachers who synthesized recommended science content knowledge for middle school across the 8 standards and frameworks listed above. These standards and frameworks were established by experts in the field and have been used to establish state and national science teaching standards (National Science Teachers Association 2003). This synthesis process was not designed to articulate the full extent of content knowledge that would be desirable for middle school teachers to possess, but rather represented a minimum of science content knowledge by identifying content consensus among these documents.

Validity of Depth of Knowledge Categories

The content analysis process resulted in a consensus list of content understandings that were initially synthesized into three science understandings categories: (a) conceptual understandings of scientific phenomena and concepts; (b) abilities and understandings related to doing science or scientific inquiry; and (c) applications of scientific inquiry and concepts to human society, that is, STS. These science understandings categories were integrated with the knowledge-type framework synthesized from Li and Shavelson (2001) and Shulman (1986) described earlier.

Additional validity evidence for the misconceptions used to operationalize the pedagogical content knowledge questions came from the research literature, including journals, ERIC documents, research reference books, and symposia and conference proceedings. Misconceptions are prevalent in teachers as well as students. Preservice and practicing teachers sometimes hold unstable, incomplete, and disconnected structures of knowledge (Gess-Newsome and Lederman 1993; Hoz et al. 1990; Tamir et al. 1981). Research on science teachers' understanding and misconceptions reveals that they tend to struggle with a number of science concepts, such as electrochemistry (De Jong et al. 1995), conservation, mole, and atomic mass (Haidar 1997), biology/geology (Hoz et al. 1990), evolution (Lee et al. 1997; Rutledge and Warden 1999), force and motion (Preece 1997), and energy (Trumper 1997).

Generation and Validation of Assessment Items

Generation of Assessment Items

For each content domain (physical, life, and earth/space science), an expert team of science educators, middle school science teachers, and university science faculty was assembled to generate items. Team members from these different populations were included to enhance item validity characteristics by providing multiple perspectives on the science content as well as appropriateness for middle school teachers. Each content team reviewed their assessment's item specification grid, clarified definitions of the various knowledge types and content subcategories, and developed a common understanding of the item generation task. Each team generated sample items that spanned the spectra of knowledge types and content subcategories. Items were targeted to fit the criteria, according to the writer, for specific cells in the item specification grid (see Table 2). Following the group meetings, team members individually generated items intended to fit particular cells.

To strengthen the validity of categorizing items in the item specification grid, items were initially evaluated via writing team discussions. Sample items were either confirmed in their intended placement, modified to better fit, or discarded from the question set. After this initial round of item generation and discussion, team members repeated the cycle of new item generation and group discussion in order to generate sufficient prototype items to be distributed nationwide for external review.

Nationwide External Review of Items to Strengthen Validity

Upon completion of item generation, item content validity was strengthened through evaluation by independent external reviewers. Independent reviewers from the same three categories of professionals as the item generation teams (scientists, middle science teachers, science educators) were solicited nationwide. Each item was reviewed by members from all three categories of professional external reviewers.

Along with the items themselves, reviewers were provided with types of knowledge descriptors in Table 1 and content categories and subcategories in Table 2. Items were sent along with a review form that solicited: (a) the correct answer to the multiple-choice items, (b) categorization of each item into a content category and subcategory, (c) categorization of each item into a knowledge-type category, (d) rating the item as STS or not, and (e) a rating of the appropriateness of the item for middle school teachers. For each item, there were a total of 27–31 reviewers in life science, 29–33 in physical science, and 20–22 in earth/space science.

Data from the reviewers were used to identify items that met criteria project personnel established for each of the categories reviewed. If an item met all three of those criteria (content category and subcategory classification, knowledge-type classification, and appropriateness), it was retained for field testing. If it met two of the criteria, the item was reviewed to see whether it could be clarified or improved. Significantly revised items received a second external review, whereas those with more minor revisions were included on the field tests and subjected to further analysis based on field test results.

Initial Field Tests and Assessment Revisions

The initial round of field testing included a total of 891 teachers (268 physical science, 355 life science, and 268 earth/space science) located at sites in eight states.

The six versions of each of the three assessments were assigned the label “iteration 1” (and labeled with “.1” after the version number) to distinguish them from later revisions.

The initial field test results were the basis for assessment revision which generated iteration 2 for each of the assessments. Individual items that did not intercorrelate strongly with others in the same subscale were identified for review and possible edits (e.g., wording, or content, or clearly focusing on one concept rather than multiple concepts interwoven in an item). Patterns of distractors chosen by large percentages of testtakers were also identified for review and possible edits.

Data for the results reported in this study are from the revised, iteration 2, assessments that have been administered to 4,432 teachers at 304 sites distributed across 21 states. Although some sites administered different versions of these assessments to their teachers more than once (e.g., in a pretest–posttest design with their PD institute as the intervention), the data presented here are based only on pretest results and so do not reflect the impact of any interventions.

Reliability

Across all three content areas, 4,432 (1,759 life; 1,460 physical; 1,213 earth) pretest assessments completed by teachers were included in these analyses. Although designed specifically for middle school teachers, many professional development sites where the coordinator chose to use these assessments included teachers with a variety of grade-level certifications, and so substantial numbers of elementary (usually grades 4 and 5) and high school-certified teachers are included in our sample.

These results were used to compute two types of reliability of scores from these assessments. First, interrater reliability was established between scorers of the open-response items by using percent of agreements among two to three scorers. From the collaborative efforts in selecting benchmark answers from initial field tests, project personnel refined scoring rubrics for the open-response items. Secondly, internal reliability was determined by computing Cronbach’s alpha.

Interrater Reliability

There are five open-response items on each assessment as described earlier. Project personnel initially collaborated to develop rubrics and score the field tests. Scorers independently scored each test and in doing so would note particular benchmark answers. Scorers then met as a group to compare values of each open-response question’s part (a) and (b); this process resulted in lengthy discussions regarding each scorer’s rationale for selecting a particular point value. After multiple meetings of groups of scorers, a scoring rubric was created from the process. These detailed rubrics included benchmark answers for each point value for each question. Using these rubrics, scorers achieved the following interrater reliability: 86 % for life science; 83 % for physical science; and 84 % for earth science.

Internal Reliability

Cronbach's alpha is one of the most commonly reported reliability measures of internal consistency of items in a scale (Henson 2001). Table 3 provides the reliability coefficients for all versions of each of the three content assessments.

As shown in Table 3, for all versions of each physical and life science assessment, Cronbach's alpha was at least 0.80. The earth/space science assessment likewise had most versions with Cronbach's alpha at least 0.80 with two exceptions (0.78 and 0.71).

In addition to reporting Cronbach's alpha testwide, because many users may be interested in subscales in addition to the overall score, we report in Table 4 alpha by subscale.

As would be expected for scales that are composed of relatively few items, these values of alpha are not as robust as the ones testwide. This suggests that any interpretations to be made at the subscale level should be done with appropriate caution.

Results

Results from all version numbers of a science assessment were aggregated. Concurrent validity of content scores was explored by comparing them to other

Table 3 Internal reliability for entire assessment (30 items = 35 points)

Version	Cronbach's alpha	95 % Confidence interval for alpha	*N*
Physical science (alpha = 0.85 for all versions combined)			
1	0.894	0.878, 0.908	397
2	0.850	0.822, 0.875	262
3	0.862	0.841, 0.882	365
4	0.798	0.762, 0.831	275
5	0.813	0.746, 0.869	74
6	0.807	0.744, 0.861	87
Life science (alpha = 0.83 for all versions combined)			
1	0.816	0.788, 0.842	367
2	0.841	0.819, 0.861	465
3	0.809	0.778, 0.837	330
4	0.848	0.747, 0.922	25
5	0.801	0.769, 0.831	321
6	0.804	0.767, 0.837	251
Earth/space science (alpha = 0.80 for all versions combined)			
1	0.714	0.656, 0.766	215
2	0.839	0.810, 0.865	271
3	0.809	0.752, 0.859	102
4	0.775	0.726, 0.819	184
5	0.814	0.782, 0.844	288
6	0.856	0.821, 0.887	151

Table 4 Subscale internal reliability

Subscale	Number of points	Cronbach's alpha	95 % Confidence interval for alpha
Physical science (n = 1,460)			
Declarative	5	0.365	0.312, 0.415
Inquiry	5	0.380	0.328, 0.429
Schematic	15	0.735	0.715, 0.755
Pedagogy	10	0.760	0.740, 0.779
Matter	6	0.549	0.513, 0.584
Motion forces	6	0.440	0.394, 0.483
Energy	13	0.639	0.611, 0.666
STS	5	0.263	0.201, 0.321
Life science (n = 1,759)			
Declarative	7	0.492	0.455, 0.528
Inquiry	7	0.585	0.554, 0.614
Schematic	11	0.642	0.617, 0.667
Pedagogy	10	0.757	0.739, 0.775
Struct-Fnct	6	0.458	0.418, 0.497
Int. Regulation	6	0.466	0.426, 0.504
Heredity-Div	7	0.583	0.553, 0.612
Interdependence	6	0.533	0.499, 0.566
STS	5	0.440	0.397, 0.480
Earth/space science (n = 1,213)			
Declarative	8	0.589	0.553, 0.623
Inquiry	6	0.289	0.225, 0.349
Schematic	11	0.619	0.586, 0.650
Pedagogy	10	0.662	0.631, 0.691
Atmo–Hydro	8	0.450	0.402, 0.496
Litho	11	0.616	0.583, 0.648
Solar	6	0.350	0.292, 0.405
STS	5	0.371	0.314, 0.426

measures often used as proxies for teacher content knowledge when direct measures of content knowledge are unavailable: teacher characteristics such as number of college science courses taken and certification areas. This section explores the characteristics of the sample and then explores the relationships between science assessment scores and grade level or certified content area.

Groupings of Teachers

Grade-Level Certification

One category of grouping was based on teachers' self-reported certification grade level. These teachers were certified throughout the nation in a number of different states, and the grade bands associated with certifications varied from state to state. The numerous variations on reported grade-level certifications were collapsed into four levels: elementary (any range of grades pre-K to sixth grade); middle (any range of grades 5 or 6–8 or 9, including K-8 certifications); high (grades 6–12 or higher grade spans); and other (e.g., art teachers or media specialists certified K-12). Results reported here exclude the "other" category.

Science Certification

A second grouping variable generated from self-reports was the science-specific certification. There were four levels of this grouping variable: no science, general science, other (non-target) science, and target science. Some teachers were certified as generalists (e.g., many elementary teachers do not have a science-specific certification) and were classified as having no science-specific certification. Others had a general science certification (e.g., many middle school teachers) that did not identify a specific science discipline.

Many teachers held certifications for specific sciences such as biology, chemistry, or physics. If the content of the assessment they took matched their science-specific certification (e.g., physics-certified teachers taking the physical science assessment), these teachers were categorized into the "target science" group. Those who held a specific science certification that did not match the content area of the assessment they completed (e.g., a biology-certified teacher taking the earth/space science assessment) were categorized into the "other science" group. For those teachers who held multiple specific science certifications, they were categorized as "target" if any of their certifications matched the assessment.

Interaction Between Grade-Level Certification and Science Certification

One might expect a significant correlation between these two grouping variables. For example, elementary teachers tended to be generalists without a science-specific certification, whereas middle school teachers often held a general science certification. Table 5 displays the interaction between these two grouping variables.

As anticipated, the Spearman rho showed a significant correlation between grade-level certification and science specialty certification for all assessments. In general, the higher the grade-level certification, the more specialized the certified science area. Most elementary-certified teachers taking these assessments had no specific science certification, and many middle school teachers who took these assessments had a general science certification. However, even greater numbers of middle-certified teachers in this sample had no science certification, suggesting that many sites choosing to use these science assessments may be working with middle-certified teachers to add on a science certification. By contrast, many high school teachers had either certification in the same science domain as the content focus of the assessment they took or they had certification in another specific science domain.

Table 5 Interaction between grade-level certification and science certification

	No science	General sci.	Other sci. [a]	Target sci. [b]
Physical science assessment ($\rho = .644^{***}$)				
Elementary	176	16	3	0
Middle	420	166	54	11
High	35	81	139	108
Life science assessment ($\rho = .657^{***}$)				
Elementary	294	20	6	2
Middle	465	204	28	70
High	23	66	33	249
Earth/space science assessment ($\rho = .579^{***}$)				
Elementary	173	8	1	2
Middle	352	166	46	26
High	18	44	83	49

Data indicate numbers of teachers in each cell

*** $p < .001$ for Spearman rho correlation

[a] "Other science" is any discipline-specific science certification that does NOT match the content area of the assessment being taken

[b] "Target science" indicates certification in the same content area as the assessment being taken

Years Experience and Number of Content Courses

Other data self-reported by teachers included years teaching experience and the number of college physical, life, and earth science courses they have completed. The amount of teaching experience has been identified as the major source of pedagogical content knowledge (PCK; van Driel et al. 1998), and any differences between groups in years experience may contribute to measured differences in the PCK subscale score. Similarly, the number of college content courses may contribute to the outcome measure of content knowledge.

Table 6 compares the three grade-level groups of teachers on their years experience and numbers of content courses. In each case, a Tukey highly significant difference (HSD) post hoc test evaluated homogenous subsets for those ANOVA tests which were significant.

Across all three content assessments in Table 6, the effect size for any significant differences in number of years experience was small, suggesting that even in cases where the ANOVA resulted in a significant difference, there is little practical difference in years experience. Inspection of the mean number of years for each group confirms that since the total mean years experience ranged from 8.4 to 11.9, it is unlikely that any differences in pedagogical content knowledge or content knowledge would be attributable to differing years of experience.

From Table 6, the high school teachers had on average substantially more physical science courses than either the middle school or elementary teachers. With the exception of those taking the life science assessment where the middle school teachers had more physical science courses than elementary teachers, those taking

Table 6 Mean (standard deviation) of years experience and number of content courses by teacher certification grade level

	Years experience	Num. physical courses	Num. life courses	Num. earth courses
Physical Science Assessment[a]				
Elementary (n=171-193)	9.5(8.0)	2.0(2.9)	2.3(3.4)	2.1(2.6)
Middle (n=588-653)	10.3(8.5)	2.5(3.0)	3.6(4.3)	2.5(3.1)
High (n=335-365)	11.9(9.7)	7.2(6.0)	9.1(6.8)	3.0(4.1)
ANOVA results	p=.003	p<.001	p<.001	p=.01
Partial η^2 effect size	.01	.222	.224	.008
Life Science Assessment[a]				
Elementary (n=272-316)	8.4(7.7)	1.4(1.8)	1.8(2.3)	1.5(1.7)
Middle (n=676-747)	10.9(8.8)	2.6(3.1)	3.8(4.2)	2.6(3.0)
High (n=323-365)	11.4(9.4)	5.9(5.0)	10.5(6.4)	2.9(3.8)
ANOVA results	p<.001	p<.001	p<.001	p<.001
Partial η^2 effect size	.017	.184	.335	.026
Earth/Space Science Assessment[a]				
Elementary (n=168-181)	9.1(8.1)	1.7(2.5)	2.0(2.6)	2.1(2.8)
Middle (n=534-577)	10.0(8.7)	2.3(2.9)	3.4(3.9)	2.7(3.0)
High (n=185-189)	11.3(10.1)	5.7(5.1)	8.3(6.3)	4.4(5.4)
ANOVA results	p=.057	p<.001	p<.001	p<.001
Partial η^2 effect size		.161	.206	.046

Classification into elementary, middle, or high is based on certification status. The number of teachers for each measure varies slightly depending on how many provided data

[a] Brackets indicate Tukey HSD post hoc homogenous subsets for cases where the ANOVA showed a significant difference among the three groups

the other assessments showed similar numbers of physical science courses for both middle- and elementary-certified teachers. High school-certified teachers took more life science courses than either middle or elementary teachers, and in turn, middle school teachers took more life science courses than elementary teachers.

As expected, these results overall showed that among teachers who took the physical science and life science assessments, high school teachers had more relevant college science courses than middle school teachers, who in turn usually had more science courses than elementary teachers. Among teachers who took the earth/space science assessment, the pattern is similar but less distinct, as indicated by overlapping homogenous subsets and lower effect sizes. If results show there are higher science content knowledge scores for high school teachers who took more relevant college science courses, this would serve as concurrent validity evidence with course-taking. However, rather than analyses with teacher raw content scores which would not have interval-level characteristics, Rasch modeling approaches were used to explore item and assessment quality and to transform the data into interval-level measures for further analyses.

Rasch Modeling

Rasch analysis is an item response theory–based modeling approach that transforms data such as count data (e.g., number right) into interval-level data along a scale of interest (Bond and Fox 2007). Teachers' raw assessment total scores were converted via a logarithmic transformation of the odds of success (log-odds transformation) into a log-odds unit (a logit); the partial credit Rasch model was used because of the 3-point scoring structure of part (b) of the open-response questions. The teachers' logit scores are then on an interval scale, and the same process applied to item difficulty scores puts item difficulty on the same invariant logit scale which permits person ability and item difficulty to be mathematically combined for computation. Rasch theory models outcomes of teachers' performances on specific assessment items as a probabilistic function of the difference between teachers' ability and each item's difficulty (Bond and Fox 2007; Boone et al. 2011; Liu 2010). An important assumption of Rasch theory (that was tested with this study's data and reported below) is unidimensionality; in the case of this study, items on a given assessment should all be measuring the construct of teachers' knowledge of physical (life, earth/space) science.

After the raw teacher scores were used to generate teacher and item logit scores, item fit and person fit (see Table 7) were investigated to determine whether the data fit the Rasch model (i.e., if these items did indeed tap a unidimensional construct—a contributor to construct validity if supported). Fit was evaluated with the mean square statistics of infit (more sensitive to items and people which match more closely in difficulty and ability) and outfit (more sensitive to outliers). Linacre (2011) suggested that infit and outfit mean squares between 0.5 and 1.5 are "productive for measurement" (p. 600), which will be the evaluative criteria used here. The standardized z-statistic was not appropriate for evaluating fit because of its dependence on sample size, and according to Linacre (2011), any sample above about 300 makes this statistic not useful for fit analysis. Bond and Fox (2007) suggested that it is appropriate to "routinely pay more attention to infit values than to outfit values" (p. 57) because of the strong influence of outliers on outfit.

Table 7 Person and item quality measures fitting the Rasch Model

Item fit[a]		Person fit[a]		Item reliability	Person reliability
MnSq Infit	MnSq Outfit	MnSq Infit[b]	MnSq Outfit[b]		
Physical science (n = 1,460)					
0.73–1.23	0.68–1.44	95.3 %	91.8 %	0.99	0.84
Life science (n = 1,759)					
0.89–1.17	0.86–1.38	93.4 %	71 %	0.99	0.81
Earth/space science (n = 1,213)					
0.86–1.17	0.83–1.23	94.6%	90.9 %	0.99	0.79

[a] Mean square (MnSq) fit indices between 0.5 and 1.5 are "productive for measurement" (Linacre 2011, p. 600)

[b] Person fit data are reported as percentage of the sample falling between the 0.5–1.5 parameters

Table 7 shows that all 30 items on all three assessments had good fit, indicating that these three sets of items did each tap a unidimensional construct. The person fit statistics showed that, with one exception, over 90 % of the sample in this study showed good fit, indicating that in general this sample of teachers were appropriate for being assessed with these measures. The exception was the relatively lower 71 % of sample that met the outfit guidelines on the life science assessment. Data to be presented shortly will show that overall the scores on the life science assessment were somewhat higher than on the other two assessments (see Table 6 which indicates that both the middle- and high school-certified teachers had taken more life science courses than physical or earth). The effect of these higher life science scores was to have more high-scoring outliers and simultaneously leave more of the lower-scoring people as outliers on the other end of the scale. The scores of these more numerous outliers disproportionately contributed to poor outfit; by contrast, the infit percentage was adequately strong.

Person and item reliabilities were also investigated. Person reliability indicates the extent to which this assessment would reliably replicate the placement of teachers on the logit scale (this is the most typical interpretation of reliability), and item reliability indicates the reliability of replicating the order of item difficulties with a different sample of teachers drawn from the same population. Table 7 shows that item reliability is quite strong and person reliability is adequate for reasonable measurement.

Having confirmed with the fit and reliability statistics that these items do reliably map onto a unidimensional construct for this sample, a person-item Wright map was used to explore how well these items measured the spectrum of ability levels within this sample (see Fig. 1).

For all three assessments, Fig. 1 shows that each set of items matches the ability levels of the teacher samples quite well. The means of each distribution (teacher ability and item difficulty) are similar to each other on the logit scale, and the range of item difficulties encompasses the vast majority of the teacher abilities in each case without leaving any substantial measurement gaps in abilities. More items clustered at spots where many teacher abilities are clustered strengthen the ability of this assessment to measure this group.

Note that for all three assessments, the open-response items (21a&b–25a&b) are mostly at the most difficult end of the scale, as might be expected since there is no effect of guessing correct answers as is likely partly true for the multiple-choice items 1–20. If those open-response items were not present on the assessment, there would be substantial numbers of high-ability teachers who would not be adequately assessed with the remaining set of mostly easier items, and these assessments would be much less useful to measure teacher science content knowledge of these strong teachers. In general, the part (b) of open-response items 21–25, which are the items asking teachers to express their pedagogical content knowledge (PCK) through explanation of instructional decision-making and practices within a specific scenario, lie at the most difficult end of the spectrum. This is evidence that strong PCK knowledge is generally more challenging for teachers than straightforward science content knowledge, which is logical given that strong content knowledge is a necessary but not sufficient condition to possess strong PCK.

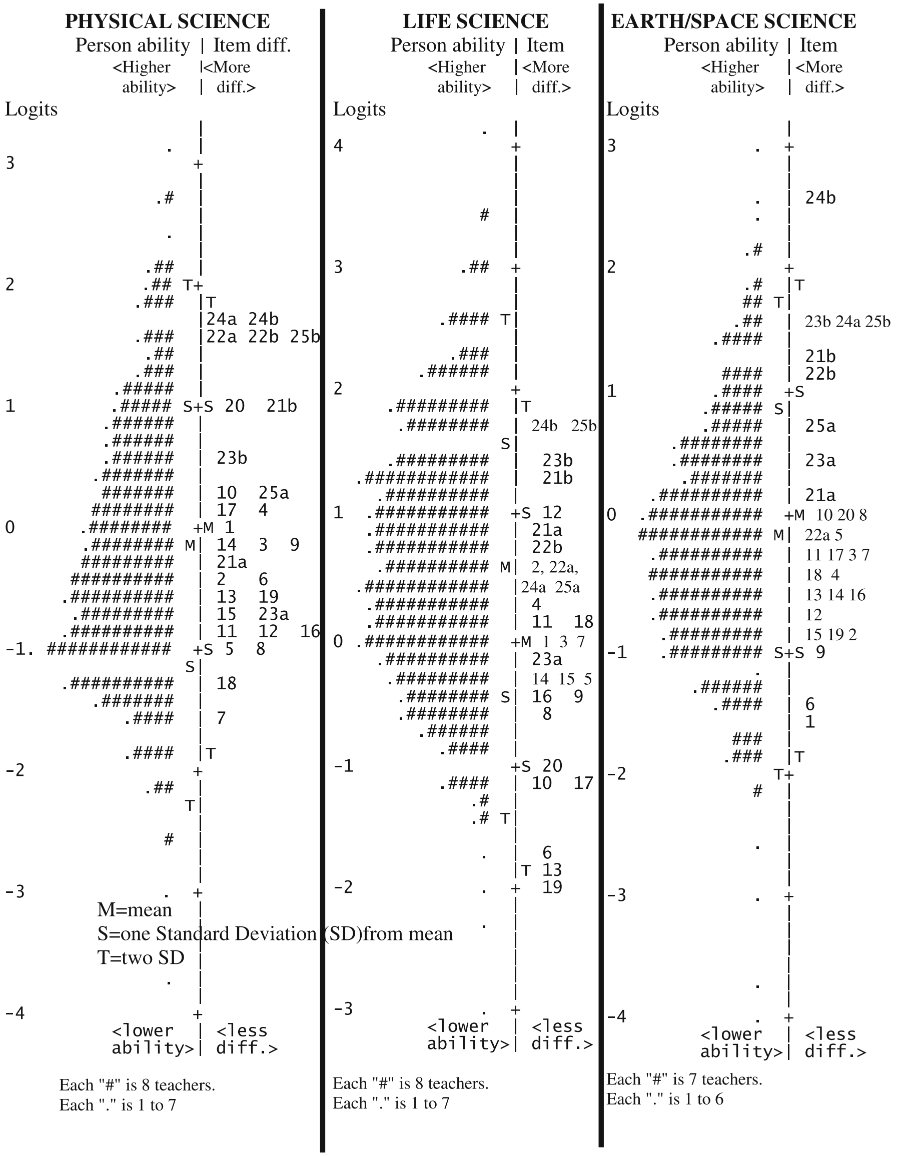

Fig. 1 Weight person-item map for each science assessment

Content Knowledge Scores

Total content knowledge scores were explored for each of the three content assessments by grade-level groups (elementary, middle, high) and by science certification categories (no science, general science, other science, target science). When the omnibus test indicated a statistically significant difference existed, a post hoc HSD analysis was used to examine homogenous subsets.

In order to compare scores of these groups, interval-level data were needed, and so the logit scores of teachers were used in these analyses rather than raw scores, which are ordinal data capturing frequency of ‘right answers.’ Table 8 provides the

Table 8 Converting raw scores to logits (interval scale) or percentiles

Raw score	Physical science		Life science		Earth/space science	
	Logit	percentile	Logit	percentile	Logit	percentile
0	−4.95	0	−5.09	1	−4.95	1
1	−3.71	1	−3.84	1	−3.71	1
2	−2.96	1	−3.07	1	−2.97	1
3	−2.50	1	−2.60	1	−2.51	1
4	−2.16	2	−2.24	1	−2.17	1
5	−1.88	3	−1.95	1	−1.89	2
6	−1.63	6	−1.70	1	−1.65	4
7	−1.42	9	−1.47	1	−1.44	6
8	−1.22	14	−1.27	2	−1.24	10
9	−1.04	20	−1.08	3	−1.06	14
10	−0.87	27	−0.91	5	−0.89	20
11	−0.70	32	−0.74	8	−0.73	26
12	−0.54	38	−0.59	11	−0.58	32
13	−0.39	43	−0.43	15	−0.42	38
14	−0.24	48	−0.29	19	−0.28	44
15	−0.10	53	−0.14	24	−0.13	51
16	0.04	58	0.00	29	0.01	58
17	0.18	63	0.14	34	0.16	64
18	0.32	67	0.28	39	0.30	69
19	0.46	71	0.42	45	0.44	74
20	0.60	75	0.56	50	0.59	79
21	0.73	78	0.71	55	0.74	83
22	0.87	82	0.85	60	0.89	86
23	1.01	85	1.00	66	1.04	89
24	1.16	88	1.16	71	1.20	92
25	1.30	90	1.32	76	1.36	94
26	1.46	92	1.49	81	1.53	96
27	1.62	93	1.67	85	1.71	97
28	1.79	95	1.86	89	1.91	98
29	1.97	97	2.07	93	2.12	99
30	2.18	98	2.31	95	2.36	99
31	2.43	99	2.59	97	2.64	99
32	2.73	99	2.94	99	2.97	99
33	3.15	99	3.40	99	3.42	100
34	3.84	100	4.15	99	4.15	100
35	5.04	100	5.39	99	5.37	100

conversion of raw scores to logits for these samples of teachers and also offers the percentile within which each raw score would fall to facilitate interpretation.

Content Knowledge by Grade Level

Total content knowledge scores were investigated by grade level for each of the three content area assessments (see Table 9).

A one-way ANOVA among these three grade-level groups for each of the three science content assessments, followed by a Tukey HSD post hoc test of homogenous subgroups, showed that for each of the three assessments, the high school teachers scored significantly higher than middle school teachers, who in turn scored significantly higher than elementary teachers. The effect sizes were largest for the physical science assessment (partial $\eta^2 = 23$ %), followed by life science (15 %) and then earth/space science (7 %).

These results provide concurrent (with college course-taking) validity support for these assessments in measuring teacher content knowledge. As expected, high school teachers, who in general are required to have more extensive university preparation in science for science certification and who, among those who took these tests, had the greatest number of college science courses (see Table 6) in physical science, life science, and earth/space science, also had the highest mean scores. The stronger effect size for physical and life science assessments also aligns with the stronger effect sizes in number of courses taken by high school teachers in physical and life sciences. In the case of earth/space science, the effect size for high school teachers' higher number of college earth/space science courses was relatively small (5 % in Table 6), which aligned with the small effect size (7 % in Table 9) for their earth/space science knowledge scores. These results confirm that these science

Table 9 Total assessment logit score mean (standard deviation) by certification grade level [Raw score mean (SD)]

	Physical Science [a]	Life Science [a]	Earth/Space Science [a]
Elementary (logits)	-0.81(.77)	0.10(.88)	-0.39(.73)
[raw scores]	[10.8(4.7)]	[16.7(5.5)]	[13.4(4.7)]
	(n=221)	(n=344)	(n=219)
Middle (logits)	-0.30(.95)	0.57(1.02)	-0.05(.96)
[raw scores]	[14.0(6.0)]	[19.9(6.0)]	[15.8(5.9)]
	(n=709)	(n=803)	(n=628)
High (logits)	0.64(1.02)	1.27(.96)	0.42(.94)
[raw scores]	[20.2(6.5)]	[24.2(5.1)]	[18.7(5.9)]
	(n=392)	(n=385)	(n=205)
ANOVA results	p<.001	p<.001	p<.001
Partial η^2 effect size	.232	.151	.074

Logit scores are used for the analysis because they are interval data. Raw scores are reported in small square brackets below logits for convenience of interpretation. Maximum raw score on all assessments is 35

[a] Large outside brackets indicate Tukey HSD post hoc homogenous subsets

assessments produced scores that are consistent with expectations from the number of relevant college-level science courses taken.

Content Knowledge by Science Certification Category

Total content knowledge scores were investigated by science certification category (see Table 8).

The pattern of results in Table 10 showed that teachers with a specific science certification that matches the content of the assessment ("target science") tended to score significantly higher than other teachers and that those with no science certification scored lower than all other groups. These results also offer support for concurrent (with certification) validity of these science assessments. Those whose certification matched the content of the assessment scored highest, while those with no science certification scored the lowest.

Discussion

This discussion summarizes validity and reliability arguments and outlines potential uses of these assessments.

Table 10 Total assessment score mean (standard deviation) by science certification [raw score mean (SD)]

	Physical Science [a]	Life Science [a]	Earth/Space Science [a]
No Science (logits) [raw scores]	-0.57(.86) [12.3(5.4)] (n=757)	0.30(.91) [18.0(5.7)] (n=923)	-0.29(.81) [14.1(5.0)] (n=659)
General Science (logits) [raw scores]	0.19(.99) [17.3(6.4)] (n=307)	0.80(1.15) [21.4(6.2)] (n=322)	0.19(1.0) [17.2(6.3)] (n=209)
Other Science [b] (logits) [raw scores]	0.41(.91) [18.7(6.0)] (n=219)	0.71(.88) [20.8(5.6)] (n=85)	0.17(1.0) [17.1(6.2)] (n=150)
Target Science [c] (logits) [raw scores]	0.96(1.09) [22.2(6.8)] (n=136)	1.26(.94) [24.3(5.0)] (n=361)	0.59(1.02) [19.9(6.1)] (n=124)
ANOVA results	p<.001	p<.001	p<.001
Partial η^2 effect size	.251	.140	.106

Logit scores are used for the analysis because they are interval data. Raw scores are reported in square brackets below logits for convenience of interpretation. Maximum raw score on all assessments is 35

[a] Brackets indicate Tukey HSD post hoc homogenous subsets

[b] "Other science" indicates certification in a science other than the content area targeted by the assessment

[c] "Target science" indicates certification in the same content area as the assessment being taken

Validity Evidence from Assessment Development Process

These assessments were designed to strategically sample across both a depth of knowledge and a breadth of knowledge dimension. Validity for the depth of knowledge dimension was strengthened because this dimension was grounded in taxonomies described in the literature. The breadth of knowledge for each science domain was determined by a systematic and collaborative team process of synthesizing content from a suite of relevant national and international documents (see Saderholm and Tretter 2008 for a detailed description of this process).

Assessment items were then generated to strategically, simultaneously, proportionally sample across both the depth and breadth dimensions. Items were initially generated and revised through several cycles by teams consisting of middle school science teachers, postsecondary science educators, and scientists with specialties in the domains being assessed. These items then underwent a structured external review by individuals representing the same three sets of expertise. This process enhanced both the content validity of individual items and the construct validity of systematically sampling across depth and breadth dimensions. Items that survived the external review process were then collated into field tests, and results were used to revise the assessments, enhancing validity through testing individuals in the target population.

Validity and Reliability Evidence from Empirical Results

Reliability from Empirical Results

Reliability of scores of open-response items by trained scorers showed reasonable interrater reliability. Internal consistency reliability for the assessments as a whole indicated reasonable Cronbach's alphas of at least 0.80 or higher for nearly all versions of all 3 science domains (see Table 3). Because of anticipated interest in users wishing to explore interpretations at the subscale level, Cronbach's alphas for subscales (Table 4) were also presented, but as expected due to fewer numbers of items contributing to a subscale, many of these relatively low Cronbach's alphas suggest that interpretations of subscale scores should be done with great caution. Strong item and person reliabilities (Table 7) suggested that these assessments would likely replicate the ordinality of item difficulties if the sample were to repeat the testing process and also that they would return scores that would replicate the ordinality of person scores if the sample were to retest.

Validity from Empirical Results

Good item fit (Table 7) and strong alignment of the item difficulty spectrum with the teacher ability spectrum (Fig. 1) suggested that these sets of items can appropriately measure teachers' science knowledge. Concurrent validity evidence emerged from comparisons by grade-level certification (Table 9) which aligned with number of college-level science courses taken, and by comparisons of scores by certification (Table 10).

Potential Uses of DTAMS Science Assessments

These science assessments offer evidence of both validity of potential score interpretations and reliability of those scores from multiple sources of evidence. Valid and reliable assessments of teacher science content knowledge provide access to direct measurement of a crucial variable of interest to educational researchers, professional development providers, and science teacher educators. These assessments could be used to determine the impact of workshops, courses, or other experiences on teachers' knowledge. However, because deep and well-integrated science knowledge takes substantial time and effort to acquire, it may not be reasonable to expect a substantial gain to be measured by these assessments if the intervening experiences are relatively short (e.g., a few weeks) or otherwise unable to address the spectrum of the sampled science domain in a substantial manner.

Another consideration is the science content domain intended to be measured. Because the grain size of these assessments was designed to sample across all of the physical, life, or earth/space science domains (to fit the science breadth demands of middle school teachers), it would be important for the user to judge whether their needs reasonably match this grain size.

There are a few cautions for potential users to be aware of as they consider which instruments might best serve their measurement purposes. These three science assessments are not intended to be commeasurable. It would not be appropriate to compare a score on a physical science assessment with a life science score, for example, since any particular score will not have identical meanings across assessment science domains. Also, interpretations based solely on percent correct are not meaningful. For example, a raw score of 50 % correct may or may not be a strong score. These assessments were designed to sample science knowledge both broadly and deeply, and care was taken to craft an assessment that would avoid a ceiling effect. These results (Fig. 1) suggest that this has been achieved, which implies that only a few of the very strongest teachers may achieve scores close to the maximum.

These assessments are designed around content knowledge needed to teach middle school science, but are not limited to a depth of knowledge appropriate for middle school students. This is particularly true for the pedagogy items, but is also true for many of the other items as well. Thus, using these assessments with middle school students is inappropriate since they are not likely to yield useful information for that population. Likewise, care should be taken if users choose to administer these assessments to non-middle school science teachers. As can be seen from the data presented, there are many who have chosen to administer these assessments to both elementary teachers and high school teachers. It may be reasonable to consider administering these assessments to teachers of upper elementary grades (e.g., grades 4–5) or to some high school teachers (e.g., those teaching a general 9th grade science that either spans content or serves as an introductory course to more specialized high school science courses). Likewise, high school-certified teachers in one science area who are adding on a second science certification may be an appropriate audience for these assessments.

These assessments were not designed to align with any particular professional development program or research question, and thus, they may not be sensitive to the actual content of a particular program. Many workshops, courses, or other professional development experiences often focus on more narrow bands of science knowledge than is represented by any one of these assessments. For example, a particular workshop may focus only on force and motion, rather than on all of physical science. In such cases, care should be taken if interpreting only the relevant subscale instead of the entire test, especially given the not-surprising lower values of Cronbach's alpha (see Table 4) for subscales which are composed of fewer items than the assessment as a whole. In general, if the content foci of a particular workshop or professional development session(s) were not strongly aligned with a particular assessment, use of one of these assessments to make claims about workshop impact may be limited. Programs that focus only on narrow bands of the middle school science curriculum will not likely see positive results on overall assessment scores. Because these assessments were designed and evaluated as intact units, using only a portion of them would weaken the validity and reliability evidence on which interpretations can be grounded, and appropriate caution would be warranted.

Do ensure that testing conditions support valid inferences. If, for example, a group of teachers were given one of these assessments in the very last hour of an intense, energy-draining 3-week summer workshop and told that they can leave when they are finished, such a condition might encourage some individuals to not accord adequate time or thought to responses, which would invalidate any results that were obtained.

These assessments are best suited for measuring impacts of programs that intend to broadly improve middle school teachers' science knowledge in one of the three content domains. This implies a best match for valid assessment in programs that include significant, sustained efforts. These assessments may be used in a comparative manner. For example, in a research design that wishes to control for science content knowledge, having logit scores for all teacher participants as a pretest measure may serve as a useful covariate. Likewise, using them in a pretest–posttest research design can be informative when other conditions support valid inferences of scores. Potential users may also choose to use these assessments to explore science content knowledge of teachers as a predictor variable in a regression-based design such as a hierarchical linear model or multiple regression design. Access to direct measures of teacher science content knowledge offers researchers, science teacher educators, and others an opportunity to incorporate this critical variable in a number of situations of interest. Readers interested in acquiring a sample assessment or seeking further details on how to acquire these for their use are encouraged to visit the project website at http://louisville.edu/education/research/centers/crmstd/diag_sci_assess_middle_teachers.html for more information.

References

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

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological

Testing (U.S.). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge, Taylor & Francis Group.

Boone, W. J., Townsend, J. S., & Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education, 95*(2), 258–280.

Darling-Hammond, L. (2000). How teacher education matters. *Journal of Teacher Education, 51*, 166–173.

De Jong, O., Acampo, J., & Verdonk, A. H. (1995). Problems in teaching the topic of redox reactions. *Journal of Research in Science Teaching, 32*(10), 1097–1110.

Gess-Newsome, J., & Lederman, N. G. (1993). Preservice biology teachers' knowledge structures as a function of professional teacher education: A year-long assessment. *Science Education, 77*(1), 25–45.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new "Standards for Educational and Psychological Testing": Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181–191.

Haidar, A. (1997). Prospective chemistry teachers' conceptions of the conservation of matter and related concepts. *Journal of Research in Science Teaching, 34*(2), 181–197.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177–189.

Hoz, R., Tomer, Y., & Tamir, P. (1990). The relations between disciplinary and pedagogical knowledge and the length of teaching experience of biology and geology teachers. *Journal of Research in Science Teaching, 27*(10), 973–985.

Ingersoll, R. (2000). *Challenges to finding and keeping teachers*. Retrieved August 27, 2004 from http://www.chemistry.org/portal/a/c/s/1/acsdisplay.html?DOC=government\scproject\scp_teaching.html.

Interstate New Teacher Assessment and Support Consortium (INTASC) Science Standards Drafting Committee. (2002). *Model standards in science for beginning teacher licensing and development: A resource for state dialogue*. Council of Chief State School Officers. Retrieved January 14, 2004 from http://www.ccsso.org/content/pdfs/ScienceStandards.pdf.

Kendall, J. S., & Marzano, R. J. (2004). *Content knowledge: A compendium of standards and benchmarks for K-12 education*. Aurora, CO: Mid-continent Research for Education and Learning. Online database: http://www.mcrel.org/standards-benchmarks/.

Laczko-Kerr, I., & Berliner, D. (2002). The effectiveness of "Teach for America" and other under-certified teachers on student academic achievement: A case of harmful public policy. *Education Policy Analysis Archives, 10*(37), 1–53. Retrieved March 1, 2012 from http://epaa.asu.edu/epaa/v10n37/.

Lee, Y. J., Izard, J., & Yeoh, O. C. (1997). *Teacher knowledge of biological evolution from the perspectives of classical test and item response theory*. Paper presented at the Education Research Association Annual Conference, Singapore.

Li, M., & Shavelson, R. J. (2001). *Examining the links between science achievement and assessment*. Paper presented at the AERA Annual Meeting, Seattle, WA.

Linacre, J. M. (2011). *A user's guide to winsteps*. Retrieved March 13, 2012 from http://www.winsteps.com/a/winsteps-manual.pdf, p. 600.

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch Modeling approach*. Charlotte, NC: Information Age Publishing, Inc.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., et al. (2003). *TIMSS assessment frameworks and specifications 2003* (2nd ed.). Retrieved January 14, 2004 from http://isc.bc.edu/timss2003i/PDF/t03_AF_sci.pdf.

National Academies, Committee on Prospering in the Global Economy of the 21st Century. (2006). *Rising above the gathering storm: Energizing and employing America for a brighter future*. Washington, DC: National Academies Press.

National Assessment Governing Board. (2004). *Science framework for the 2005 national assessment for educational progress*. Washington, D.C: Author.

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

National Science Teachers Association (NSTA). (2003). *Standards for science teacher preparation. NSTA*. Retrieved January 14, 2004 from http://www.nsta.org/main/pdfs/NSTAstandards2003.pdf.

No Child Left Behind Act. (2001). Public Law 107-110. Retrieved January 14, 2004 from http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf.

Praxis Series: Professional Assessments for Beginning Teachers. (2004). *General Science Content Knowledge Part 2 (0432)*. Educational Testing Service (ETS). Retrieved January 14, 2004 from http://www.ets.org/Media/Tests/PRAXIS/pdf/0432.pdf.

Preece, P. F. W. (1997). Force and motion: Pre-service and practicing secondary science teachers' language and understanding. *Research in Science and Technological Education, 15*(1), 123–128.

Rutledge, M. L., & Warden, M. A. (1999). The development and validation of the measure of acceptance of the theory of evolution instrument. *School Science and Mathematics, 99*, 13–18.

Saderholm, J., & Tretter, T. R. (2008). Identification of the most critical content knowledge base for middle school science teachers. *Journal of Science Teacher Education, 19*(3), 269–283.

Shulman, L. (1986). Those who understand, knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Stiggins, R. J. (1997). Performance assessment of skill and product outcomes. In K. M. Davis (Ed.), *Student-centered classroom assessment* (2nd ed., pp. 261–302). Columbus, OH: Prentice-Hall, Inc.

Tamir, P., Gal-Choppin, R., & Nussinovitz, R. (1981). How do intermediate and junior high students conceptualize living and nonliving? *Journal of Research in Science Teaching, 18*, 241–248.

Trumper, R. (1997). A survey of conceptions of energy of Israeli pre-service high school biology teachers. *International Journal of Science Education, 19*(1), 31–46.

US Department of Education, National Center for Education Statistics. (2004). *Qualifications of the public school teacher workforce: Prevalence of out-of-field teaching*, 1987–88 to 1999–2000. NCES 2002-603 Revised. Washington, D.C.

van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching, 35*(6), 673–695.