JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 49, NO. 6, PP. 683–690 (2012)

Editorial

Assessment and Science Education: Our Essential New Priority?

Nancy Butler Songer[1] and Maria Araceli Ruiz-Primo[2]

[1]*School of Education, University of Michigan, Ann Arbor, Michigan*
[2]*School of Education and Human Development, University of Colorado Denver, Denver, Colorado*

Received 13 June 2012; Accepted 14 June 2012

Over 10 years ago, a National Research Council committee led by Jim Pellegrino and Robert Glaser generated the fundamental text on educational assessment titled, *Knowing What Students Know: The Science and Design of Educational Assessment* (NRC, 2001). In this document, the authors emphasize that assessment is a process of reasoning from evidence, ''a process by which educators use students' responses to specially created or naturally occurring stimuli to draw inferences about the students' knowledge and skills'' (National Research Council 2001, p. 20).

Knowing What Students Know (NRC, 2001; KWSK) was fundamental to advancing the conversation on assessment in science and other disciplines for several reasons. First, KWSK provided a set of guiding principles for the development and evaluation of educational assessments. The assessment triangle, introduced as an assessment model, was particularly important as it brought recognition to the importance of using cognitive models to drive the design of the assessment and define the empirical evidence needed to support the interpretations derived from observed performance. Second, KWSK called for a ''balanced assessment system'' of classroom and large-scale assessments that are: *comprehensive*—using multiple sources of evidence about students' learning; *coherent*—a shared learning model coordinating curriculum, instruction, and assessment; and *continuous*—longitudinal assessment of learning progress over time. This assessment system posed a model to follow in any educational system. Third, KWSK confirmed the idea that assessment should be designed with a specific purpose in mind and cannot serve multiple purposes. Fourth, KWSK emphasized the necessity for assessments to be sensitive to cultural and linguistic difference characteristics of the tested audience. Fifth, KWSK presented new advances in educational measurement, psychometrics, and technology. In all of these ways, KWSK set a high bar for quality assessments. At the

Correspondence to: N. B. Songer; E-mail: songer@umich.edu
DOI 10.1002/tea.21033
Published online 11 July 2012 in Wiley Online Library (wileyonlinelibrary.com).

same time, it established the importance of collaboration among science educators, psychometricans, and researchers towards products and research outcomes with the potential to improve and extend the quality and reach of educational assessments in science education.

In our view, the work foreshadowed by KWSK—the creation and use of science education assessments rooted in fundamental foundations in learning theories—has never been more important. In the United States, A Framework for K-12 Science Education (NRC, 2011) and the associated Next Generation Science Standards (Achieve, 2012) provide a conceptual framework and an organizational structure for national standards focusing on ''higher levels of understanding of science and engineering'' (Achieve, 2012, p. 2). These critical documents contribute a perspective on the science knowledge that should be fostered in K-12 that is deeply rooted in what we know about how individuals learn science. They draw from learning theory-driven research ideas, such as learning progressions previously highlighted in the *Journal of Research in Science Teaching* 46:6, 2009, and from an emphasis on knowledge that fuses science content with science practices and crosscutting concepts (e.g., NRC, 2011; Songer & Gotwals, 2012). On balance, since these documents do not focus on assessment, they present only cursory information about the implications of such standards on the design, evaluation, or use of assessment instruments used to gather evidence of student learning.

Why a Special Issue on Science Education and Assessment?

Eleven years after the publication of KWSK, we present a special issue of the *Journal of Research on Science Teaching* focusing on assessment. We believe this special issue comes at an extremely critical time. In the current accountability era and with an increased attention to Science, Mathematics, Engineering and Technology (STEM) education, it seems essential that we harness our collective intellectual, institutional, and technological resources towards quality assessment design and evaluation that focuses on worthy purposes. Without good assessments, our instructional effectiveness, individual achievement, and program and institutional quality are all at risk. Without a community of scholars fully invested in science education assessment work, we risk the possibility of outsiders leading a conversation that should be ours to orchestrate.

We hope this special issue will advance our conversation towards greater awareness of the value and importance of work focused on science education and assessment. We believe that the science education community is particularly well positioned at this time to fulfill some of the potential for quality assessments in science education that were outlined in KWSK. While the science education community struggles to characterize effective cognitive models on which assessment should be based, we nevertheless have relatively high agreement on some of the characteristics of science knowledge that a new generation of assessments need to do a better job of assessing (see Pellegrino, this volume). Through this collection of current research articles and insights from James Pellegrino, the lead author of Knowing What Students Know (NRC, 2001), we hope to not only to build on earlier assessment-related work in science education, but also to serve as a catalyst for firmly establishing the importance of assessment-related research work in science education. We hope to demonstrate both new potential and the considerable effort required to develop, evaluate and use quality assessments that provide information on what individuals know and are able to do.

Overview of Articles in this Special Issue

In 2011 a call for contributions to a Special Issue on Assessment appeared in the *Journal of Research in Science Teaching*. Twenty-four articles were submitted in response to this call, and 5 research articles were selected. These five articles and one insightful commentary

chapter are included in this volume. While the articles touch on issues across a wide spectrum of topics, they can be organized by two themes: Three of the articles focus on the development and technical evaluation of assessments, and the other two focus on the validity implications of assessments already developed. The first three articles emphasize the development and evaluation process focusing on empirical evidence to support interpretations. The fourth article examines the validity implications for diverse populations of a large-scale assessment, and the fifth article focuses on the discussion of inferences that can be drawn from the comparison of assessments designed for different purposes. A brief overview of each of the articles is presented below.

Developing and Evaluating Instructionally Sensitive Assessments in Science (Ruiz-Primo, Li, Wills, Giamellaro, Lan, Mason, and Sands)

This article focuses on a major gap in the instructional sensitivity literature: how to develop instructionally sensitive assessments. In the article, Ruiz-Primo et al. define instructionally sensitive assessments as those that provide evidence that students' scores yield valid information about: (1) whether students had the opportunity to learn certain curricular content; (2) the quality of instruction students received; and (3) which aspects of the curricular content may require further attention. The approach proposed for this type of assessment is guided by three underlying concepts: variation of item proximity to the science curriculum at hand, transfer of learning, and big ideas. The authors explain the item characteristics that were manipulated to create items at different distances (close and proximal) from what students learned in their classroom. Empirical evidence based on the students' pretest to posttest performance gains in seven classrooms is used to test the suitability of the approach and the validity of the instructionally sensitive claims. Items developed at different distances from the science module showed a pattern of pre-to-post gain consistent with their instructional sensitivity, that is, the closer the items were to the science module, the larger the observed gains and effect sizes.

Developing a Construct-Based Assessment to Examine Students' Analogical Reasoning Around Physical Models in Earth Science (Rivet and Kastens)

This article by Rivet and Kastens contributes to our knowledge about assessment of students' analogical reasoning around physical models in earth science. The article provides information about how to conceptualize, develop, and evaluate the technical qualities of an assessment developed on an articulated construct with different levels of understanding. The authors describe, based on literature on analogical mapping, a three-tiered construct depicting the increasing sophistication of students' analogical reasoning. Assessment items were developed with the help of teachers. Assessments were designed in a way that they could elicit students' reasoning that reflected differences of the quality of reasoning. The authors describe how interviews with students contributed to the design of the items. The items were evaluated with 164 students. The approach used to evaluate the assessment provides evidence about how the assessment proved to be more effective in distinguishing students at the lower level of the constructs than at the higher levels. It also allowed identification of items that were not working as expected. The article adds to understanding of how assessment can be used with a defined construct based on a cognitive model of how students develop expertise in the domain of physical models in earth science.

Cognitive Foundations for Science Assessment Design: Knowing What Students Know About Evolution (Opfer, Nehm, and Ha)

This study describes an assessment in biology developed around four cognitive design principles associated with differences between experts and novices based on the literature on

evolution change: (1) the use of core concepts to organize and facilitate long-term recall; (2) the use of causally central features in explaining evolution change; (3) the greater emphasis on scientifically normative explanations over naïve ideas in reasoning about evolutionary change, and (4) the focus on abstract versus superficial characteristics of assessment tasks. Using these principles, the authors developed four items in which superficial features are manipulated to evaluate the consistency of students' thinking patterns in response to items tapping the same construct. The article describes the scoring schema used to judge the quality of the explanations. The article provides empirical evidence to support the four cognitive principles that guided the design of the items.

"I Never Thought of it as Freezing" How Students Answer Questions on Large-Scale Science Tests and What They Know About Science (Noble, Suarez, Rosebery, O'Connor, Warren, Hudicourt-Barnes)

This study by Noble and coworkers draws from a sociocultural perspective on assessment to examine interview data associated with students' interpretation of standardized test items. The aim was to gain insights into possible factors contributing to persistent achievement gaps between students from historically non-dominant communities and students from dominant communities. The authors matched data on students' understanding of target knowledge with data focused on state science assessment items intended to evaluate that target knowledge for students from three populations: students from low-income households, middle-class, native English speakers, and middle class students who were English language learners. Although the number of test items was relatively small, the results were intriguing: a higher percentages of false negatives were observed among low income and ELL students as compared to native English speakers, and a higher percentage of false positives were observed among native English speakers. The work clearly speaks to the need for more research that can help us examine not just the validity of high-stakes assessment instruments, but also the validity of assessment instruments relative to different target populations, such as those historically underrepresented in science.

The Consequence of School Improvement: Examining the Association Between Two Standardized Assessments Measuring School Improvement and Student Science Achievement (Maltese and Hochbein)

This article focuses on exploring possible relationships between performance on standardized assessments used to evaluate school improvement and other measures of performance, such as student performance on college entrance examinations in mathematics and science. To address questions of this kind, Maltese and Hochbein examined patterns of performance on ACT science and mathematics exams as compared to patterns of performance on standardized measures in English and mathematics used to determine school improvement over 3 years. Results demonstrated no consistent relationship between student performance on college entrance examinations in science and mathematics and standardized measures of school improvement in literacy and mathematics. The authors provide several interesting alternative explanations for these findings. One explanation focused on the possibility that college-bound students demonstrate consistently strong scores and resiliency regardless of whether or not their school is demonstrating improvements, while another explanation focused on a more dismal view highlighting the lack of relationship between school-wide improvements in English and mathematics and corresponding improvement on college-bound tests in STEM areas. The article concludes with valuable insights about the importance of making careful choices about the content areas we prioritize through testing and instructional time, and the value of science and STEM content within these prioritized areas.

Assessment of Science Learning: Living in Interesting Times (Pellegrino)

Framed by a quote from Robert Kennedy outlining our chronological positioning as living in ''interesting times,'' Jim Pellegrino's insightful commentary concludes our special issue. This commentary offers wisdom drawn from his distinguished career that includes important work in science education and assessment. This chapter not only skillfully highlights principles from the canon of quality assessment (e.g., the assessment triangle and the need for evidence associated with the validity of assumptions for any assessment use). In addition, Pellegrino emphasizes new and important opportunities available within the area of science education and assessment, such as the recent convergence on teaching, learning, and assessing science knowledge that fuses content, science practices, and crosscutting concepts (e.g., NRC, 2011). The chapter closes with a call for systems of assessment that meet standards of comprehensiveness, coherence, and continuity—a challenge that once again echoes themes introduced in KWSK as it suggests new opportunities, particularly through new partnerships and advances in technology and psychometrics that we are encouraged to embrace.

Continuing Challenges

Despite advances in the assessment field, we see ongoing challenges in science education and assessment as set forth in KWSK (NRC, 2001). Developing and evaluating high quality assessments persists as a critical problem in education for many reasons, a few of which we discuss here. We often forget that assessment development goes hand-in-hand with the technical evaluation of the assessment that involves not only reliability, but also, and more importantly, with validity. Furthermore, problems in understanding what it takes to evaluate the validity of assessments are not solved despite the advances in approaches to validity (e.g., Kane, 2006). Researchers, including those in science education, still often view empirical evidence to support score interpretation more as a checklist rather than as a chain of reasoning based on confirming and disconfirming evidence. Further, we still have not learned how to include and meaningfully consider the diversity of the populations that are being tested during the entire process of assessment development and evaluation (Solano-Flores, 2008).

Many in science education and more broadly still use a single assessment for multiple and different purposes despite our knowledge that assessments designed to be used for one purpose are not well suited for a different purpose. In an accountability context, assessments results are used not only to judge students' learning and achievement, but also to evaluate teacher effectiveness and to make serious decisions such as promoting or firing teachers. Assessments are also used to judge the quality of schools and the quality of education at the state level. In each of these cases, do we consistently ask about the validity of these assessments for such diverse purposes? There is still a lack of consensus and understanding within national, local or regional (e.g., states in the U.S.) accountability systems in their struggle to accurately evaluate ''educational treatments'' with appropriate assessments (e.g., curriculum choices, teacher quality; Briggs & Wiley, 2008).

We continue to debate how high quality assessments can signal worthy goals to pursue (Pellegrino & Goldman, 2008) and how assessment in the classroom context can be used to improve students' learning. Despite the enchantment for formative assessment, the research and practice community still has little understanding of what it really takes to implement it properly. For example, many see formative assessment as all about instruments implemented during instruction, as opposed to the ongoing processes of adjusting instruction based on students' needs indentified using the information collected through diverse forms of

assessments. This misunderstanding has led to the creation of benchmarks and interim assessments, distracting from the authentic purposes of formative assessment. We have much to learn about how to develop assessments that, by definition, support the learning process.

One area that we view as particularly needing to be addressed is the lack of recognition that courses specifically about assessment are essential in both teacher education and graduate level programs. We suggest that part of the reason assessment remains so mysterious and challenging for many teachers and researchers is that they mistakenly assume that the work is only for measurement experts. We ignore the importance of creating teams with diverse expertise (e.g., teachers, content experts, methodologist, and experts in educational assessment and measurement) to develop assessments that can help improve the quality of science assessments. We believe that efforts to extend assessment conversations to larger audiences, as this special issue attempts to do, will help all of us to realize the importance of collaborative partnerships.

A common premise about assessment is that it is a positive force in improving students' learning. Assessments can play critical roles in the educational system by focusing on different purposes: supporting learning, assessing individual achievement, and evaluating programs and institutions (NRC, 2001). The quality of assessments is critical whether in the classroom or in a large-scale context and whether they are designed for formative or summative purposes. However, despite major advances in the cognitive and measurement sciences, development of high quality assessments continues to be a struggle. This is certainly true in science education. We still have a long way to go in improving assessment and assessment practices. What is measured, why and how it is measured, whose achievement is measured, what type of inferences are made, and the kinds of evidence supporting such inferences—all remain enduring issues in assessment development in science education.

Concluding Remarks

The five articles and the other submissions are representative of the work currently being conducted in the field of science education to respond to the challenges previously described. However, it is important not to overlook the focus areas that were not represented in the pool of articles. None of the submitted articles addressed the challenges of creating balanced assessments systems with potential for building coherence among state, district, and classroom formative and summative assessment. Nor did we receive articles on assessment development that took into account the diversity of the population being tested. Articles focusing on issues related to developing assessment for science teachers (whether of science content or practices) were also absent. These are all highly complex and difficult issues that require the attention of the whole community aligned with a strong research agenda. In what follows we suggest several priority areas for a research agenda that we believe would do much to help the field make significant progress in some of the most critical areas.

Research on Assessment for English Language Learners (ELLs)

The needs of ELLs and the goals for valid and fair assessment for these students are far from being properly addressed. Those at the center of shaping assessment policy and practice are somehow oblivious to knowledge from the fields of bilingualism and language development. The evidence is considerable: inappropriate definitions of ELL populations and English proficiency; failure to include these students in the entire process of assessment development; limited participation of language specialists in the process of item writing; and the use in large-scale assessment programs of testing accommodations that are linguistically ineffective. This evidence shows how assessment policy and practice is yet to benefit from knowledge

from the language sciences—knowledge that has been available for decades. Future efforts to develop more valid and fair assessments for ELL populations should also continue to keep as a primary priority the promotion of a view of assessment as a multidisciplinary endeavor among professionals in the field of educational measurement.

Research on Assessment of the 21st Century Skills

How can we develop assessments in science that tap adaptability, complex communication and social skills, non-routine problem solving, self-management, and system thinking? Given the importance of these skills in what has been name "knowledge work," how can we embed these skills in the design and development of assessments in science? It seems that we know more about how to assess problem solving within a content area than about how to assess adaptability, social skills, or self-management. Issues related to how to define these constructs and what type of tasks are needed in the context of science education are just few areas that by themselves require a full research agenda (Ruiz-Primo, 2009).

Research Evaluating the Role of Technology in Formative and Summative Science Education Assessment

While we have seen emerging technological resources that can rapidly process big data sets, we still need more research studies that can apply and evaluate the role and benefits of these resources for science education assessment. Advances in gathering and processing information using new technologies provide tremendous promise for more customized, frequent and useful feedback and evaluation information. As mentioned in Jim Pellegrino's chapter (this volume), we encourage more articles in the *Journal* that can help us advance our collective understanding of the many ways in which technological advances can enhance science education assessment.

Research Exploring How Different Nations Address the Same Critical Issues

We have been well aware of some important research in science education and assessment that is occurring outside the United States, both in individual countries and in large-scale international tests (e.g., PISA, OECD, 2012; TIMSS, 2008). We were therefore disappointed that this collection of articles could not reflect this work. If only because of the future challenges we face in science education and assessment, we encourage the development and writing of more research articles in science education assessment from outside the United States, which would allow the *Journal* to continue to deepen and expand insights and dialogue that represent all of us.

Assessment Design and Research on Learning Science Within Informal Science Education Contexts

While we appreciate the growing area of research focused on learning in informal science education contexts and the work that studies connections between informal and formal learning environments, we see a need for more research studies examining assessment in these contexts. Therefore, we encourage articles in future issues of the *Journal* that focus on the assessment of science education in informal contexts and between formal and informal settings.

Interdisciplinary Partnerships

As we are well aware, achieving productive interdisciplinary partnerships is no simple task (NRC, 2004). Such partnerships are essential if we are to participate fully in our own

''interesting times'' (Pellegrino, this volume) and take advantage of opportunities for acquiring major new insights and creating necessary changes ahead in the field of assessment and science education. Science educators and others have struggled for many years to understand the principles and best practices for forming fruitful collaborations between experts in various disciplines. But we have no doubt that the work that lies ahead in science education and assessment will continue to require collaborative partnerships between science educators and measurement experts. We thus look forward to future articles that articulate best practices for interdisciplinary collaborations between science educators, researchers, and experts in educational measurement. These collaborations will result in a body of strong research studies and outcomes that, in turn, will make a positive difference for instruction and learning in the most diverse schools and classrooms.

References

Achieve. (2012). Next Generation Science Standards May 2012 Public Draft. Retrieved from: http://www.nextgenscience.org on 5/14/12.

Briggs, D. C., & Wiley, E. W. (2008). Causes and effects. In K. E. Ryan & L. A. Shepard (Eds.), The future if test-based educational accountability (pp. 171–190). New York, NY: Routledge.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: Praeger Publishers.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Washington, D.C.: National Academy Press.

National Research Council. (2004). Facilitating interdisciplinary research. Washington, D.C.: National Academy Press.

National Research Council. (2011). A Framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, D.C.: The National Academies Press.

OECD. (2012). *PISA 2009*. Technical Report, PISA, OECD Publishing. DOI: 10.1787/9789264167872-en

Pellegrino, J. W., & Goldman, S. R. (2008). Beyond rhetoric: Realities and complexities of integrating assessment into classroom teaching and learning. In C. A. Dwyer (Ed.), The future of assessment (pp. 7–52). New York, NY: Routledge.

Ruiz-Primo, M. A. (2009). *Towards a framework for assessing 21st century science skills*. White Paper. The National Academies. Washington, D.C. http://www7.nationalacademies.org/bose/Ruiz_Primo_21st_Century_Paper.pdf

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. Educational Researcher, 37(4), 189–199.

Songer, N. B., & Gotwals, A. (2012). Guiding explanation construction by children at the entry points of learning progressions. Journal of Research in Science Teaching, 49(2), 141–165.

TIMSS. (2008). *TIMSS 2007 Achievement 4th 8th Grade.pdf*. Downloaded from http://timss.bc.edu/timss2007/release.html on 5 January 2012.