NARST
A worldwide organization for improving science teaching and learning through research

# Supporting the Implementation of *NGSS* through Research: **Assessment**

Maria Araceli Ruiz-Primo, University of Colorado Denver, Chair
Lou DiBello, University of Illinois, Chicago
Guillermo Solano-Flores, University of Colorado, Boulder

The *NGSS* identify three dimensions essential for providing students with a quality science education: *science and engineering practices*, *crosscutting concepts* across domains of science, and *disciplinary core ideas*. Integrating these dimensions provides a meaningful context for science content and practices—how science knowledge is acquired and understood through conceptual connections across disciplines (Achieve, 2013b). The dimensions are blended in what is called in the standards, *performance expectations—assessable components* that articulate what students should know and be able to do when instruction is effective (Achieve, 2013a).

This statement focuses on critical issues related to assessments designed to provide evidence regarding students' proficiency in meeting the *NGSS* performance expectations. Two questions guide the statement: *(1) What conceptual frameworks should be considered if these assessments are to be useful in both the classroom and large-scale assessment contexts?* and *(2) What is required to develop assessments that align with the performance expectations proposed by the NGSS?*

## Conceptual Frameworks for Assessment

Four critical assessment frameworks bear on the quality of the *NGSS* performance expectations: *systems*, *contexts*, *purposes*, and, at the center of all of them, *validity and fairness*.

### Assessment Systems

Assessment of the *NGSS* performance expectations should reflect the characteristics of seamless educational assessment systems (see Pellegrino, Chudowsky, & Glaser, 2001; Wilson & Bertenthal, 2006)—i.e., they should be coherent, comprehensive, and continuous. *Coherent* refers to both the consistency between classroom and large-scale assessment and the alignment of assessment with curriculum and instruction. *Comprehensive* refers to the need to use diverse assessment methods to inform decision making and to diagnose students' strengths and needs (e.g., performance assessments and predict-observe-explain and informal strategies in the classroom). Assessments should tap aspects of student knowledge that go beyond declarative and procedural knowledge, including assessments that focus on principled understanding (knowing why). This is true for both large-scale and classroom assessment. Both should call for synthesis of information from several sources. *Continuous* refers to the measurement of student progress over time. Progress over time should focus on student growth across years (growth models) and at the classroom level. It is important for teachers to understand that on-going assessment of student learning (knowing where they are in their level of understanding) is a critical aspect of effective teaching.

**Assessment Contexts**

The knowledge and skills students acquire in class should be (but in fact may not be) the knowledge and skills probed by large-scale assessments (e.g., Raizen et al., 1989). Two requirements must be met for coherence between external and classroom-based assessments contexts (Ruiz-Primo, 2002): (a) students, teachers, district personnel, and policy makers must share the same learning goals; and (b) models of student learning underlying external and classroom assessments must be compatible.

**Assessment Purposes**

Developing and selecting assessments focused on the *NGSS* performance expectations should be guided by an assessment purpose (Pellegrino, Chudowsky, & Glaser, 2001): *assist learning*, *measure individual achievement*, or *evaluate programs*. It is critical to understand that one type of assessment cannot fit all purposes, and trade-offs and compromises need to be clearly recognized. Uses of assessments for purposes other than those for which they were created raise validity concerns.

**Assessment Validity and Fairness**

Validity is critical to any form of assessment in both the large-scale and the classroom contexts. While there are multiple ways of viewing validity, for the purposes of this document, validity refers to the reasonableness of judgments made about students' knowledge and skills based on a given set of assessment activities. A validity argument "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2006, p. 23). These judgments should not be influenced by factors irrelevant to the knowledge and skills being measured (e.g., student proficiency in the language in which tests are administered), nor should they favor students who are more familiar with the cultural conventions or contexts used in tasks.

Evidence on validity should be gathered both *prospectively,* during assessment design, development and pilot testing, and *retrospectively,* once assessments are completed and available for evaluation, while being fully integrated within learning contexts. To be valid, assessments and assessment systems should focus on three overlapping types of evidence (see Kane, 2006; Messick, 1989; Moss et al., 2006). The forms of evidence are cognitive, instructional and analytical/interpretive:

> **Cognitive aspects of validity.** Assuming shared goals and consistent models of learning at all levels of an assessment system, evidence is needed of the extent to which assessments elicit forms of thinking and reasoning that are relevant to the knowledge and skills targeted, and are not confounded with linguistic skills or working memory load. Such evidence typically is collected through conversations with students or by having students verbalize their thinking while responding to tasks. In the classroom context, teachers should be able to identify students' understanding at a given point in instruction. In the large-scale context,

verbal protocols and cognitive interviews should be conducted with samples of pilot students to infer how they reason when they respond to items.

**Instructional aspects of validity.** Given a model of coherent, comprehensive and continuous assessment as an integrated component of classroom instruction, evidence is required on how well an assessment supports teaching practice (*formative value*), provides timely instructional information (*practical value*), and reflects what students learn in the classroom (*instructional sensitivity value*).

**Analytical and interpretive aspects of validity.** Assessment purposes should guide inferences made about students' achievement based on assessment outcomes. These inferences should be based on analyses and interpretations of students' performances using the shared cognitive model as a guide. In the classroom, these analyses and interpretations refer to the accuracy of the teachers' judgments of their students' understanding relative to learning goals. In the large-scale context, evidence should be gathered on the extent to which assessments' content is aligned with the *NGSS* and on the psychometric properties, including measures and indices of score reliability, estimation accuracy and model-data fit. Critical to this form of validity is the quality and sufficiency of information relevant to the intended interpretive purpose.

These overlapping aspects of validity interact with and complement one another and drive the assembly of linked evidence for testing hypotheses about validity and for constructing a comprehensive validity argument that supports: (a) instructional decision-making, (b) improvement of student learning, and (c) projections of student performance on external summative assessments.

## Assessment Development

In large-scale assessment, the process of assessment development is multidisciplinary. In the classroom context, it should be a collegial effort in which teachers support each other improving their ideas on assessment activities and tasks that inform instruction.  A realistic vision must include provision for adequate teacher professional development activities for promoting quality assessment activities and should be informed by the research and development literature. In addition to content, assessment developers should address the cognitive, structural, and linguistic factors involved in student performance. This process should pay attention to the intimate relationship between validity and fairness at all its stages. Also, it should ensure that the performance expectations of assessments aligned with the *NGSS* are valid and fair, and have the intended instructional impacts. Development issues can be grouped into three equally important categories: assessment targets, fairness and inclusion, and integration into classroom practice.

### Assessment Targets

**Clear identification of the target core ideas and performance Expectations.** The *NGSS* contain a wealth of core disciplinary ideas and sub-ideas—a sizable increase in the amount of content specified in the 1996 National Science Education Standards (Coffey & Alberts, 2013). To be able to make reasonable claims about students' scientific knowledge and practices at each grade level, we need to ask: (1) *What are the most important core ideas or performance expectations to assess and to report?* and (2) *What set of performance expectations and ideas are critical to ensuring classroom interpretability and usability of assessments?* A coherent assessment system demands a clear identification of the most relevant core ideas and expectations to be assessed.

**Use of models of learning as unifying elements.** Learning models are key to ensuring cohesion among curriculum, instruction, and assessment (Pellegrino, Chudowsky, & Glaser, 2001). They go beyond simple learning progressions, logical taxonomies, or specifications of content, and encompass cognitive theories of the construct representations that underlie student performance. These models should help to formalize a framework of learning and instruction to guide assessment design that is coherent across grades. As long as the underlying models of learning are consistent, assessments will be aligned along the vertical dimension of the assessment system (across grades) and the horizontal dimension such that external assessments complement classroom assessments (Huff & Goodman, 2007).

**Use of a cognitive diagnostic approach.** Task development should be guided by cognitive models of how students engage with tasks and represent knowledge. A cognitive approach identifies the processes, strategies, and knowledge underlying student performance on a given assessment task, and can connect the learning models with diagnostic psychometric models in ways that enable the reporting of informative profiles to teachers and students (see Roussos et al., 2007). This approach must be sensitive to students' strengths and weaknesses in a subject and allow the identification of causes of difficulty (Huff & Goodman, 2007). Assessment tasks best inform teaching and learning when they are situated within an aligned, integrated system of curriculum, instruction, and assessment (Nichols, 1993).

## Fairness and Inclusion

**Consideration of language issues at all stages of the development process.** Language is critical to learning and demonstrating science knowledge. To a great extent, assessment development is about ensuring that the linguistic and other representational forms used in assessment tasks are understood by students as intended. Issues of language should not be addressed as an afterthought. Because English Language Learners (ELLs)—students who are developing English as a second language while continuing to develop their native language—are included in large-scale assessments provided in English, so should they also be included at all stages of assessment development and validation. Contrary to common beliefs, the majority of ELL students have basic communication skills in English and can participate in pilot tests and cognitive interviews *in English* thus providing valuable information for improving the linguistic and representational features of assessment tasks.

**Proper representation of linguistic groups and the language professions.** Fairness can be achieved only when the process of assessment development takes into account different ways in which students may make sense of science tasks. Because first languages are powerful influences in students' epistemologies, the process of assessment development should include students with different first language backgrounds. The size and makeup of samples of students included in the process of assessment development (e.g., in pilot studies or cognitive interviews) should represent the linguistic mosaic of student linguistic backgrounds in the nation or in the school district or state in which assessments are to be administered (Solano-Flores, 2008). Also, linguists and specialists in bilingual development should be included at all stages in the process of assessment development, as these professionals provide expertise on the subtle ways in which linguistic features can bias items against linguistic minorities.

**Attention to the relationship between cognition and language.** Native language is a major factor shaping cognitive processes (Bialystok, 2001). Each type of assessment task is sensitive to different aspects of knowledge (e.g., recall, critical thinking, problem solving) and poses different forms of linguistic demands (e.g., recognizing terms, expressing ideas). This issue, which is critical to valid and fair assessment of *all* students, is even more important for ELLs. Thus, assessments should comprise a wide variety of task types (e.g., multiple-choice, completion, essay, hands-on, computer simulations), as each provides a unique set of opportunities to for students to demonstrate multiple aspects of knowledge.

**Testing accommodations for students with special needs.** Testing accommodations for ELL students and students with disabilities should be developed and verified throughout the entire process of assessment development. When determining if a given form of accommodation is appropriate, four criteria are especially important: (1) How sensitive is the accommodation to the needs of individual test takers? (2) How readily can the accommodation be implemented with fidelity? (3) How much will students benefit from the accommodation without prior familiarity? and (4) What unintended consequences may occur if test takers who do not need the accommodation are nonetheless provided with it?

**Use of technological advances in testing.** Innovative assessment formats can be designed that take advantage of multiple forms of information display and the interactive capabilities of computers. Along with these possibilities comes the challenge of ensuring that new forms of assessment meet assessment targets and do not put students with limited access to computers at disadvantage.

### Integration into Classroom Practice

**Increased efforts to help integrate assessment results into classroom practice.** Large-scale assessments can only reach their full potential to provide the best sources of information to advance student learning if they are better designed and provide sufficient information about students' strengths and weaknesses that teachers can use to inform their instructional practice. At the classroom level, teachers need to be better equipped to improve their classroom assessment practices (Notar et al., 2004; Stiggins, 2001). Teacher preparation programs should equip teachers with the skills required to develop, select and implement quality assessment tools, including differentiation of instruction based on assessment outcomes. Finally, curriculum developers must provide teachers with models of learning-based classroom assessment tools that require high cognitive demands of students and provide accurate outcomes. Student work examples at various levels of performance should be made available to users, stakeholders, and the public.

## References

Achieve (2013a. *How to read the next generation science standards*. Retrieved fromhttp://www.nextgenscience.org/sites/ngss/files/How%20to%20Read%20Next%20NGSS%20%20PUBLIC%20January%20Draft%20-%20FINAL.pdf

Achieve (2013b). *Important information about the second public draft of the next generation science standards*. Retrieved from http://www.nextgenscience.org/sites/ngss/files/January%20Public%20Release%20NGSS%20Front%20Matter.pdf

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition.* Cambridge, UK: Cambridge University Press.

Coffey, J. & Alberts, B. (2013). Improving education standards. *Science*, *339*(6119), 489.

Huff, K., & Goodman, D. P. (2007). The demand of cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.). *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). Cambridge, NY: Cambridge University Press.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, *30*, 109–162.

Nichols, P. D. (1993). *A framework for developing assessments that aid instructional decisions*. (ACT Research Report Series 93-1). Iowa City, IA: American College Testing.

Notar, C. E., Zuelke, D. C., Wilson, J. D., & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. *Journal of Instructional Psychology*, *31*(2), 115-129.

O'Neil, T., Sireci, S. G., & Huff, K. L. (2004). Evaluating the content validity of a state-mandated sceicne assessment across two successive administrations of a state-mandated science assessment. *Educational Assessment and Evaluation*, *9*(3-4), 129-151.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001) (Eds.). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pellegrino, J. W., DiBello, L. V., & Brophy, S. P. (in press). The Science and Design of Assessment in Engineering Education. In A. Johri & B. M. Olds (Eds.), *Cambridge handbook of engineering education research*. Cambridge University Press: New York, NY.

Pellegrino, J. W., DiBello, L. V., James, K., Jorion, N., & Schroeder, L. (2011). Concept inventories as aids for instruction: A validity framework with examples of application. In W. Hernandez (Ed.) *Book of Proceedings of the Research in Engineering Education Symposium, Madrid, October 2011.* Retrieved from http://rees2009.pbworks.com/w/file/fetch/63149087/REES%202011%20proceedings.pdf.

Raizen, S. A., Baron, J. B., Champagne, A. B., Haertel, E., Mullis, I. V. S., & Oakes, J. (1989). *Assessment in elementary school science education*. Andover, Massachusetts: The Network, Inc.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and practice* (pp. 275–318). Cambridge: Cambridge University Press.

Ruiz-Primo, M. A. (2002, February). *On a seamless assessment system*. Paper presented at the annual meeting of the American Association of the Advancement of Science, Boston, MA.

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.

Stiggins, R. (2001). The unfilled promise of classroom assessment. *Educational Measurement: Issues and Practice*, *20*(3), 5-15.

Wilson, M., & Bertenthal, M. W. (2006). *Systems for state science assessments*. Washington, DC: The National Academies Press.