Psicología Educativa 20 (2014) 65-77

Psicología Educativa

www.elsevier.es/psed

Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress

James W. Pellegrino

Learning Sciences Research Institute, University of Illinois at Chicago, U.S.A.

ARTICLE INFORMATION

Manuscript received: 06/07/2014
Revision received: 10/08/2014
Accepted: 14/08/2014

Keywords:
Assessment
Cognition
Measurement
Testing
Design
Learning
Instruction

ABSTRACT

This paper argues that assessment can serve as a positive influence on attaining 21st century learning goals. Section I focuses on 21st century education challenges and the types of assessments needed to support attainment of learning objectives relevant to a global society. Sections II and III discuss the purposes and contexts of educational assessment and three important conceptual frameworks: (a) assessment as a process of reasoning from evidence, (b) assessment driven by models of learning expressed as learning progressions, and (c) the use of an evidence-centered design process to develop and interpret assessments. Section IV considers the implications for design of classroom and large-scale assessment. Sections V and VI consider the elements of a balanced system of assessments and key indicators of quality we must keep at the forefront as we work towards implementing coherent assessment systems as part of the process of educational transformation in the 21st century.

La evaluación como una influencia positiva en el proceso de enseñanza-aprendizaje del siglo XXI: aplicación de un enfoque sistémico al progreso

Palabras clave:
Evaluación
Cognición
Medida
Medición con tests
Diseño
Aprendizaje
Enseñanza

RESUMEN

Este artículo plantea que la evaluación puede constituir una influencia positiva para lograr los objetivos de aprendizaje del siglo XXI. La sección I se centra en los retos educativos del siglo XXI y el tipo de evaluación que se necesita para lograr los objetivos de aprendizaje relevantes para el conjunto de la sociedad. En las secciones II y III se analizan los objetivos y contextos de la evaluación educativa y tres importantes marcos conceptuales: (a) la evaluación como un proceso de razonamiento a partir de la evidencia, (b) la evaluación realizada desde modelos de aprendizaje formulados como progresiones de aprendizaje y (c) la utilización de un diseño centrado en la evidencia para diseñar la evaluación e interpretar sus resultados. La sección IV examina sus implicaciones de cara al diseño de la evaluación en el aula y de la evaluación educativa a gran escala. En las secciones V y VI se consideran los componentes de un sistema equilibrado de evaluación y los indicadores clave de calidad que hay que tener muy presentes si se desea poner en marcha un sistema coherente de evaluación como parte del proceso de transformación educativa en el siglo XXI.

Assessment is often seen by individuals in both the educational practice and research communities as a negative influence on teaching and learning, especially when high stakes are attached to the outcomes of test scores (Kaestle, 2013; Linn, 2013). This paper argues that when assessment is properly conceived, designed, and implemented it can serve as a positive influence on attaining the learning goals we have for students in the 21st century. To make the argument I draw upon a report issued by the U.S. National Research Council (NRC) entitled "Knowing What Students Know: The Science and Design of Educational Assessment" (Pellegrino, Chudowsky, & Glaser, 2001), as well as several recent reports that elaborate on points made in the 2001 NRC report. These recent reports focus on issues of educational assessment design and use given the current context of major changes in disciplinary learning standards in the United States (e.g., Darling-Hammond et al., 2013; Gordon Commission, 2013a, 2013b; Pellegrino, Wilson, Koenig, & Beatty, 2014). While many of my arguments are illustrated by drawing upon the current U.S. educational context, they are applicable to any educational system where the uses of assessment range across levels

*Correspondence concerning this article should be addressed to James Pellegrino. M/C 1240 West Harrison Street. University of Illinois at Chicago. Chicago, Illinois 60607-7137. E-mail: pellegjw@uic.edu

DOI: http://dx.doi.org/10.1016/j.pse.2014.11.002
1135-755X/

from the classroom to district, state, national or international contexts.

In Section I, the focus is on the broader challenge of 21st century education and the types of assessments we need to support attainment of learning objectives that are relevant to a global society. The section ends with a brief discussion of five elements of assessment systems that can support the evaluation of such deeper learning. Section II introduces the purposes and contexts of educational assessment and then Section III discusses three related conceptual frameworks: (a) assessment as a process of reasoning from evidence, (b) assessment driven by models of learning expressed as learning progressions, and (c) the use of an evidence-centered design process to develop and interpret assessments. Section IV turns to the implications of the material in Section III for classroom assessment and large-scale assessment. Section V then considers the elements of a balanced system of assessments and Section VI returns to the five elements of assessment systems discussed in Section I and closes by briefly describing key indicators of quality we must keep at the forefront as we work towards implementing coherent assessment systems as part of the process of educational transformation in the 21st century.

I. The Educational Challenge Before Us

The changing nature of work and society means that the premium in today's world is not merely on students' acquiring information, but on their ability to analyze, synthesize, and apply what they have learned to address new problems, design solutions, collaborate effectively, and communicate persuasively (see e.g., Bereiter & Scardamalia, 2013; Pellegrino & Hilton, 2012). In the United States, policymakers in nearly every state have adopted new standards intended to ensure that all students graduate from high school ready for college and careers. Achieving that goal will require a transformation in teaching, learning, and assessment so that all students develop the deeper learning competencies that are necessary for postsecondary success. This transformation will require an overhaul in curriculum and assessment systems to support such deeper learning competencies. Ministries of education around the world have been redesigning curriculum and assessment systems to emphasize these skills. For example, as Singapore prepared to revamp its assessment system, then Education Minister, Tharman Shanmugaratnam, noted (Ng, 2008):

[We need] less dependence on rote learning, repetitive tests and a 'one size fits all' type of instruction, and more on engaged learning, discovery through experiences, differentiated teaching, the learning of life-long skills, and the building of character, so that students can … develop the attributes, mindsets, character and values for future success.

Reforms in Singapore, like those in New Zealand, Hong Kong, a number of Australian states and Canadian provinces, and other high-achieving jurisdictions have introduced increasingly ambitious performance assessments that require students to find, evaluate, and use information rather than just recalling facts. In addition, these assessments – which call on students to design and conduct investigations, analyze data, draw valid conclusions, and report findings – frequently call on students to demonstrate what they know in investigations that produce sophisticated written, oral, mathematical, physical, and multimedia products (Darling-Hammond & Adamson (2010) (See Appendix for examples). These assessments, along with other investments in thoughtful curriculum, high-quality teaching, and equitably funded schools, for example, appear to contribute to their high achievement (Darling-Hammond, 2010).

The United States is poised to take a major step in the direction of curriculum and assessments for this kind of deeper learning with the adoption of new Common Core State Standards (CCSSI, 2010a, 2010b) and the Next Generation Science Standards (Achieve, 2013). These standards are intended to be "fewer, higher, and deeper" than previous iterations of standards, which have been criticized for being a "mile wide and an inch deep". They aim to ensure that students are prepared for college and careers with deeper knowledge and more transferable skills in these disciplines, including the capacity to read and listen critically for understanding, to write and speak clearly and persuasively, with reference to evidence, and to calculate and communicate mathematically, reason quantitatively and scientifically, and design solutions to complex problems.

The Common Core standards in English language arts and mathematics, and the Next Generation Science Standards will require a more integrated approach to delivering content instruction across all subject areas (Pellegrino & Hilton, 2012). The Common Core standards in English language arts are written to include the development of critical reading, writing, speaking, and listening skills in history, science, mathematics, and the arts as well as in English class. The Common Core standards in mathematics are written to include the use of mathematical skills and concepts in fields like science, technology, and engineering. These standards emphasize the ways in which students should use literacy and numeracy skills across the curriculum and in life. As states seek to implement these standards, they must also examine how their assessments support and evaluate these skills and create incentives for them to be well taught.

In the United States, two consortia of states – the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) – have been formed to develop next generation assessments of these standards. These are part of multiple initiatives to rethink assessments that accompany the disciplinary standards-driven educational reforms. Thus, it is timely to consider what the features of high-quality assessment systems that meet these new goals should include. The 2013 report of the Gordon Commission, written by many leading experts in curriculum, teaching, and assessment, described the most critical objectives this way:

To be helpful in achieving the learning goals laid out in the Common Core, assessments must fully represent the competencies that the increasingly complex and changing world demands. The best assessments can accelerate the acquisition of these competencies if they guide the actions of teachers and enable students to gauge their progress. To do so, the tasks and activities in the assessments must be models worthy of the attention and energy of teachers and students. The Commission calls on policy makers at all levels to actively promote this badly needed transformation in current assessment practice... [T]he assessment systems [must] be robust enough to drive the instructional changes required to meet the standards... and provide evidence of student learning useful to teachers.

New assessments must advance competencies that are matched to the era in which we live. Contemporary students must be able to evaluate the validity and relevance of disparate pieces of information and draw conclusions from them. They need to use what they know to make conjectures and seek evidence to test them, come up with new ideas, and contribute productively to their networks, whether on the job or in their communities. As the world grows increasingly complex and interconnected, people need to be able to recognize patterns, make comparisons, resolve contradictions, and understand causes and effects. They need to learn to be comfortable with ambiguity and recognize that perspective shapes information and the meanings we draw from it. At the most general level, the emphasis in our educational systems needs to be on helping individuals make sense out of the world and how to operate effectively within it. Finally, it is also important that assessments do more than document what students are capable of and what they know. To be as useful as possible, assessments should provide clues

as to why students think the way they do and how they are learning as well as the reasons for misunderstandings (Gordon Commission, 2013b).

No single assessment can evaluate all of the kinds of learning we value for students; nor can a single instrument meet all of the goals held by parents, practitioners, and policymakers. As argued below, it is important to envision a coordinated system of assessments, in which different tools are used for different purposes – for example, formative and summative, diagnostic vs. large-scale reporting. Within such systems, however, all assessments should faithfully represent the Standards, and all should model good teaching and learning practice.

At least five major features define the elements of assessment systems than can fully measure high quality standards such as the Common Core State Standards and the Next Generation Science Standards and support the evaluation of deeper learning (see Darling-Hammond et al. (2013) for an elaboration of the relevance, meaning and salient features of each of these five criteria):

(1) Assessment of Higher-Order Cognitive Skills: Most of the tasks students encounter should tap the kinds of cognitive skills that have been characterized as "higher-level" – skills that support transferable learning, rather than emphasizing only skills that tap rote learning and the use of basic procedures. While there is a necessary place for basic skills and procedural knowledge, it must be balanced with attention to critical thinking and applications of knowledge to new contexts.

(2) High-Fidelity Assessment of Critical Abilities: In addition to key subject matter concepts, assessments should include the critical abilities articulated in the standards, such as communication (speaking, reading, writing, and listening in multi-media forms), collaboration, modeling, complex problem solving, and research. Tasks should measure these abilities directly as they will be used in the real world, rather than through a remote proxy.

(3) Standards that are Internationally Benchmarked: In terms of content and performance standards, the assessments should be as rigorous as those of the leading education countries, in terms of the kind of content and tasks they present as well as the level of performance they expect.

(4) Use of Items that are Instructionally Sensitive and Educationally Valuable: The tasks should be designed so that the underlying concepts can be taught and learned, distinguishing between students who have been well- or badly-taught, rather than reflecting students' differential access to outside-of-school experiences (frequently associated with their socioeconomic status or cultural context) or depending on tricky interpretations that mostly reflect test-taking skills. Preparing for (and sometimes engaging in) the assessments should engage students in instructionally valuable activities, and results from the tests should provide instructionally useful information.

(5) Assessments that are Valid, Reliable, and Fair: In order to be truly valid for a wide range of learners, assessments should measure well what they purport to measure, be accurate in evaluating students' abilities and do so reliably across testing contexts and scorers. They should also be unbiased and accessible and used in ways that support positive outcomes for students and instructional quality.

One major challenge then is determining a way forward in which we can create systems of assessments that meet the goals we have for the educational system and that match up with the criteria outlined above. In what follows, we consider the contexts of educational assessment, the conceptual underpinnings of assessment, and the principled processes of design that are foundational to achieving the systems of assessment that meet the criteria outlined above. These include assessments designed to support classroom teaching and learning as well as those designed for monitoring progress in educational systems.

II. Educational Assessment in Context

Assessment Purposes and Contexts

From teachers' classroom quizzes, mid-term, or final exams to nationally and internationally-administered standardized tests, assessments of students' knowledge and skills have become a ubiquitous part of the educational landscape. Assessments of school learning provide information to help educators, administrators, policy makers, students, parents, and researchers judge the state of student learning and make decisions about implications and actions. The specific purposes for which an assessment will be used are an important consideration in all phases of its design. For example, assessments used by instructors in classrooms to assist or monitor learning typically need to provide more detailed information than assessments whose results will be used by policy makers or accrediting agencies. One of the central points of the Knowing What Students Know report was that assessments are developed for specific purposes and the nature of their design is very much constrained by their intended interpretive use.

Assessment to assist learning. In the classroom context, instructors use various forms of assessment to inform day-to-day and month-to-month decisions about next steps for instruction, to give students feedback about their progress, and to motivate students. One familiar type of classroom assessment is a teacher-made quiz, but assessment also includes more informal methods for determining how students are progressing in their learning, such as classroom projects, feedback from computer-assisted instruction, classroom observation, written work, homework, and conversations with and among students – all interpreted by the teacher in light of additional information about the students, the schooling context, and the content being studied.

These situations are referred to as assessments to assist learning, or the formative use of assessment (see e.g., Black & Wiliam, 1998; Wiliam, 2007). These assessments provide specific information about students' strengths and difficulties with learning. For example, statistics teachers need to know more than the fact that a student does not understand probability; they need to know the details of this misunderstanding, such as the student's tendency to confuse conditional and compound probability. Teachers can use information from these types of assessment to adapt their instruction to meet students' needs, which may be difficult to anticipate and are likely to vary from one student to another. Students can use this information to determine which skills and knowledge they need to study further and what adjustments in their thinking they need to make.

Assessment of individual achievement. Another type of assessment used to make decisions about individuals is that conducted to help determine whether a student has attained a certain level of competency after completing a particular phase of education, be it a two-week curricular unit, a semester-long course, or 12 years of schooling. This is referred to as assessment of individual achievement, or the summative use of assessment. Some of the most familiar forms of summative assessment are those used by classroom instructors, such as end-of-unit or end-of-course tests, which often are used to assign letter grades when a course is finished. Large scale assessments – which are administered at the direction of users external to the classroom – also provide information about the attainment of individual students, as well as comparative information about how one individual performs relative to others. Because large-scale assessments are typically given only once a year and involve a time lag between testing and availability of results, the results seldom provide information that can be used to help teachers or students make day-to-day or month-to-month decisions about teaching and learning.

Assessment to evaluate programs. Another common purpose of assessment is to help administrators, policy makers or researchers

formulate judgments about the quality and effectiveness of educational programs and institutions. Instructional evaluation can be considered formative in nature when used to improve the effectiveness of instruction. Summative uses of assessment for evaluation are incorporated increasingly in making high-stakes decisions not only about individuals, but also about programs and institutions (e.g., Linn, 2013). For instance, public reporting of state assessment results by school and district can influence the judgments of parents and taxpayers about the quality and efficacy of their schools and affect decisions about resource allocations. Just as with individuals, the quality of the measure is of critical importance in the validity of these decisions.

Further Considerations of Purposes, Levels and Timescales

As noted above, assessment occurs in multiple contexts, has a variety of formal and informal uses, and is conducted to meet different purposes. The purpose of an assessment determines priorities, and the context of use imposes constraints on the design. Thus, it is essential to recognize that one type of assessment does not fit all purposes or contexts of use. In general, the more purposes a single assessment aims to serve, the more each purpose will be compromised and the overall product will represent a sub-optimal design for each intended use. A persistent mistake is to assume that an assessment is appropriate and interpretable for a particular context of use without determining if there is evidence regarding the validity of such assumptions within that context. The one-size-fits-all fallacy is especially frequent and problematic since it produces inappropriate choices of assessments for instructional or research purposes that in turn can lead to invalid conclusions regarding persons, programs, and/or institutions.

Although assessments are currently used for many purposes in the educational system, a premise of the Knowing What Students Know report is that their effectiveness and utility must ultimately be judged by the extent to which they promote student learning. The aim of assessment should be "to educate and improve student performance, not merely to audit it" (Wiggins, 1998, p.7). Because assessments are developed for specific purposes, the nature of their design is very much constrained by their intended use. While it may seem reasonable to dichotomize between internal classroom assessments, administered by instructors, and external tests, administered by districts, states, or nations or other agencies, such a dichotomy is an oversimplification of a continuum that reflects the proximity of an assessment to the enactment of specific instructional and learning activities. Ruiz-Primo, Shavelson, Hamilton, & Klein (2002) defined five discrete points on a continuum of assessment distance: immediate (e.g., observations or artifacts from the enactment of a specific instructional activity), close (e.g., embedded assessments and semiformal quizzes of learning from one or more activities), proximal (e.g., formal classroom exams of learning from a specific curriculum), distal (e.g., criterion-referenced achievement tests such as required by the federal No Child Left Behind legislation), and remote (broader outcomes measured over time, including norm-referenced achievement tests and some national and international achievement measures). Different assessments should be understood as different points on this continuum if they are to be effectively aligned with each other and with curriculum and instruction. In essence, an assessment is a test of transfer and it can be near or far transfer depending on where the assessment falls along the continuum noted above.

The level at which an assessment is intended to function, which involves varying distance in "space and time" from the enactment of instruction and learning, has implications for how and how well it can fulfill various functions of assessment, be they formative, summative, or program evaluation (NRC, 2003). As argued elsewhere (Hickey & Pellegrino, 2005; Pellegrino & Hickey, 2006), it is also the case that the different levels and functions of assessment can have varying degrees of match with theoretical stances about the nature of knowing and learning.

Although assessments used in various contexts, for differing purposes, and at different timescales often look quite different, they share certain common principles. One such principle is that assessment is always a process of reasoning from evidence. By its very nature, moreover, assessment is imprecise to some degree. Assessment results are only estimates of what a person knows and can do. We elaborate on both of these issues in the following two sections.

III. Conceptual Frameworks

Assessment as a Process of Evidentiary Reasoning: The Assessment Triangle

Educators assess students to learn about what they know and can do, but assessments do not offer a direct pipeline into a student's mind. Assessing educational outcomes is not as straightforward as measuring height or weight; the attributes to be measured are mental representations and processes that are not outwardly visible. Thus, an assessment is a tool designed to observe students' behavior and produce data that can be used to draw reasonable inferences about what students know. Deciding what to assess and how to do so is not as simple as it might appear.

The process of collecting evidence to support inferences about what students know represents a chain of reasoning from evidence about student learning that characterizes all assessments, from classroom quizzes and standardized achievement tests, to computerized tutoring programs, to the conversation a student has with her teacher as they work through a math problem or discuss the meaning of a text. People reason from evidence every day about any number of decisions, small and large. When leaving the house in the morning, for example, one does not know with certainty that it is going to rain, but may reasonably decide to take an umbrella on the basis of such evidence as the morning weather report and the threatening clouds in the sky.

The first question in the assessment reasoning process is "evidence about what?" Data become evidence in an analytic problem only when one has established their relevance to a conjecture being considered (Schum, 1987, p. 16). Data do not provide their own meaning; their value as evidence can arise only through some interpretational framework. What a person perceives visually, for example, depends not only on the data she receives as photons of light striking her retinas, but also on what she thinks she might see. In the present context, educational assessments provide data such as written essays, marks on answer sheets, presentations of projects, or students' explanations of their problem solutions. These data become evidence only with respect to conjectures about how students acquire knowledge and skill.

In the Knowing What Students Know report the process of reasoning from evidence was portrayed as a triad of three interconnected elements: the assessment triangle. The vertices of the assessment triangle (see Figure 1) represent the three key elements underlying any assessment: a model of student cognition and learning in the domain of the assessment; a set of assumptions and principles about the kinds of observations that will provide evidence of students' competencies; and an interpretation process for making sense of the evidence in light of the assessment purpose and student understanding. These three elements may be explicit or implicit, but an assessment cannot be designed and implemented, or evaluated, without consideration of each. The three are represented as vertices of a triangle because each is connected to and dependent on the other two. A major tenet of the Knowing What Students Know report is that for an assessment to be effective and valid, the three

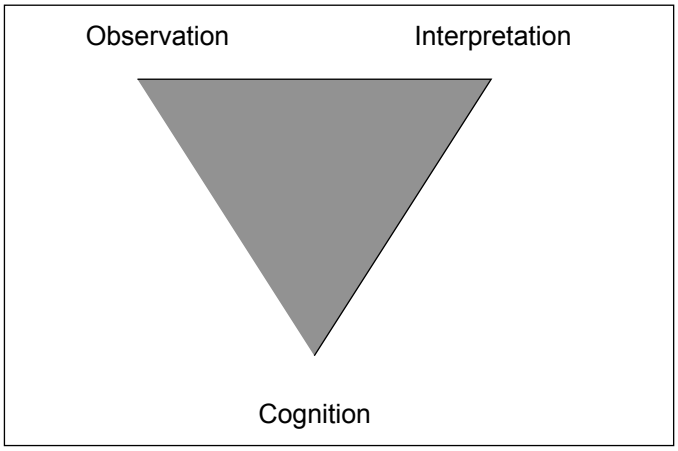

Figure 1. The Assessment Triangle

elements must be in synchrony. The assessment triangle provides a useful framework for analyzing the underpinnings of current assessments to determine how well they accomplish the goals we have in mind, as well as for designing future assessments and establishing validity (e.g., see Marion & Pellegrino, 2006).

The cognition corner of the triangle refers to theory, data, and a set of assumptions about how students represent knowledge and develop competence in a subject matter domain (e.g., fractions, Newton's laws, thermodynamics). In any particular assessment application, a theory of learning in the domain is needed to identify the set of knowledge and skills that is important to measure for the intended context of use, whether that be to characterize the competencies students have acquired at some point in time to make a summative judgment, or to make formative judgments to guide subsequent instruction so as to maximize learning. A central premise is that the cognitive theory should represent the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in a domain.

Every assessment is also based on a set of assumptions and principles about the kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates important knowledge and skills. The tasks to which students are asked to respond on an assessment are not arbitrary. They must be carefully designed to provide evidence that is linked to the cognitive model of learning and to support the kinds of inferences and decisions that will be made on the basis of the assessment results. The observation vertex of the assessment triangle represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students. In assessment, one has the opportunity to structure some small corner of the world to make observations. The assessment designer can use this capability to maximize the value of the data collected, as seen through the lens of the underlying assumptions about how students learn in the domain.

Every assessment is also based on certain assumptions and models for interpreting the evidence collected from observations. The interpretation vertex of the triangle encompasses all the methods and tools used to reason from fallible observations. It expresses how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed. In the context of large-scale assessment, the interpretation method is usually a statistical model, which is a characterization or summarization of patterns one would expect to see in the data given varying levels of student competency. In the context of classroom assessment, the interpretation is often made less formally by the teacher, and is often based on an intuitive or qualitative model rather than a formal statistical one. Even informally teachers make coordinated judgments about what aspects of students' understanding and learning are relevant, how a student has performed one or more tasks, and what the performances mean about the student's knowledge and understanding.

A crucial point is that each of the three elements of the assessment triangle not only must make sense on its own, but also must connect to each of the other two elements in a meaningful way to lead to an effective assessment and sound inferences. Thus, to have an effective assessment, all three vertices of the triangle must work together in synchrony. Central to this entire process, however, are theories and data on how students learn and what students know as they develop competence for important aspects of the curriculum.

Domain Specific Learning: The Concept of Learning Progressions

As argued above, the targets of inference for any given assessment should be largely determined by models of cognition and learning that describe how people represent knowledge and develop competence in the domain of interest (the cognition element of the assessment triangle) and what are the important elements of such competence such as how knowledge is organized, etc. Starting with a model of learning is one of the main features that distinguishes the proposed approach to assessment design from typical current approaches. The model suggests the most important aspects of student achievement about which one would want to draw inferences, and provides clues about the types of assessment tasks that will elicit evidence to support those inferences (see also Pellegrino et al., 2001; Pellegrino, Baxter, & Glaser, 1999).

Consistent with these ideas, there has been a recent spurt of interest in the topic of "learning progressions" (see Duschl, Schweingruber, & Shouse, 2007; National Research Council, 2012; Wilson & Bertenthal, 2006). A variety of definitions of learning progressions (also called learning trajectories) now exist in the literature, with substantial differences in focus and intent (see e.g., Alonzo & Gotwals, 2012; Corcoran, Mosher, & Rogat, 2009; Daro, Mosher, Corcoran, Barrett, & Consortium for Policy Research in Education, 2011; Duncan & Hmelo-Silver, 2009). Learning progressions are empirically-grounded and testable hypotheses about how students' understanding of, and ability to use, core concepts and explanations and related disciplinary practices grow and become more sophisticated over time, with appropriate instruction (Duschl et al., 2007). These hypotheses describe the pathways students are likely to follow as they master core concepts. The hypothesized learning trajectories are tested empirically to ensure their construct validity (Does the hypothesized sequence describe a path most students actually experience given appropriate instruction?) and ultimately to assess their consequential validity (Does instruction based on the learning progression produce better results for most students?). The reliance on empirical evidence differentiates learning trajectories from traditional topical scope and sequence specification. Topical scope and sequence descriptions are typically based only on logical analysis of current disciplinary knowledge and on personal experiences in teaching.

Any hypothesized learning progression has implications for assessment, because effective assessments should be aligned with an empirically grounded cognitive model. A model of a learning progression should contain at least the following elements:

(1) Target performances or learning goals which are the end points of a learning progression and are defined by societal expectations, analysis of the discipline, and/or requirements for entry into the next level of education.

(2) Progress variables that are the dimensions of understanding, application, and practice that are being developed and tracked over time. These may be core concepts in the discipline or practices central to literary, scientific or mathematical work.

(3) Levels of achievement that are intermediate steps in the developmental pathway(s) traced by a learning progression. These

levels may reflect levels of integration or common stages that characterize the development of student thinking. There may be intermediate steps that are non-canonical but are stepping stones to canonical ideas:

(4) Learning performances that are the kinds of tasks students at a particular level of achievement would be capable of performing. They provide specifications for the development of assessments by which students would demonstrate their knowledge and understanding; and

(5) Assessments, which are the specific measures used to track student development along the hypothesized progression. Learning progressions include an approach to assessment, as assessments are integral to their development, validation, and use.

Research on cognition and learning has produced a rich set of descriptions of domain-specific learning and performance that can serve to guide assessment design, particularly for certain areas of reading, mathematics, and science (e.g., American Association for the Advancement of Science, 2001; Bransford, Brown, cocking, Donovan, & Pellegrino, 2000; Duschl et al, 2007; Kilpatrick, Swafford, & Findell, 2001; Snow, Burns, & Griifin, 1998; Wilson & Bertenthal, 2006). That said, there is much left to do in mapping out learning progressions for multiple areas of the curriculum in ways that can effectively guide the design of instruction and assessment. Nevertheless, there is a good bit known about student cognition and learning that we can make use of right now to guide how we design systems of assessments, especially those that attempt to cover the progress of learning within and across grades. The paper by Deane and Song (2014) in this issue provides an excellent example of the application of the learning progressions framework, as well as the evidence centered design process discussed in the next section, as part of development of the CBAL assessment program in areas of the English language arts.

Assessment Development: Evidence Centered Design

While it is especially useful to conceptualize assessment as a process of reasoning from evidence, the design of an actual assessment is a challenging endeavor that needs to be guided by theory and research about cognition as well as practical prescriptions regarding the processes that lead to a productive and potentially valid assessment for a particular context of use. As in any design activity, scientific knowledge provides direction and constrains the set of possibilities, but it does not prescribe the exact nature of the design, nor does it preclude ingenuity to achieve a final product. Design is always a complex process that applies theory and research to achieve near-optimal solutions under a series of multiple constraints, some of which are outside the realm of science. In the case of educational assessment, the design is influenced in important ways by variables such as its purpose (e.g., to assist learning, to measure individual attainment, or to evaluate a program), the context in which it will be used (classroom or large-scale), and practical constraints (e.g., resources and time).

The tendency in assessment design is to work from a somewhat "loose" description of what it is that students are supposed to know and be able to do (e.g., standards or a curriculum framework) to the development of tasks or problems for them to answer. Given the complexities of the assessment design process, it is unlikely that such a loose process can lead to generation of a quality assessment without a great deal of artistry, luck, and trial and error. As a consequence, many assessments are insufficient on a number of dimensions including representation of the cognitive constructs and content to be covered and uncertainty about the scope of the inferences that can be drawn from task performance.

Recognizing that assessment is an evidentiary reasoning process, it has proven useful to be more systematic in framing the process of assessment design as an Evidence Centered Design process (e.g., Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006). For an extensive discussion of the logic and multiple components of ECD as applied to test development, the reader is referred to the paper by Zieky (2014) in this issue. For present purposes, Figure 2 suffices to capture three essential components of the overall process. As shown in the figure, the process starts by defining as precisely as possible the claims that one wants to be able to make about student knowledge and the ways in which students are supposed to know and understand some particular aspect of a content domain. Examples might include aspects of algebraic thinking, ratio and proportion, force and motion, heat and temperature etc. The most critical aspect of defining the claims one wants to make for purposes of assessment is to be as precise as possible about the elements that matter and express these in the form of verbs of cognition that are much more precise and less vague than high level cognitive superordinate verbs such as know and understand. Example verbs might include compare, describe, analyze, compute, elaborate, explain, predict, justify, etc. Guiding this process of specifying the claims is theory and research on the nature of domain-specific knowing and learning.

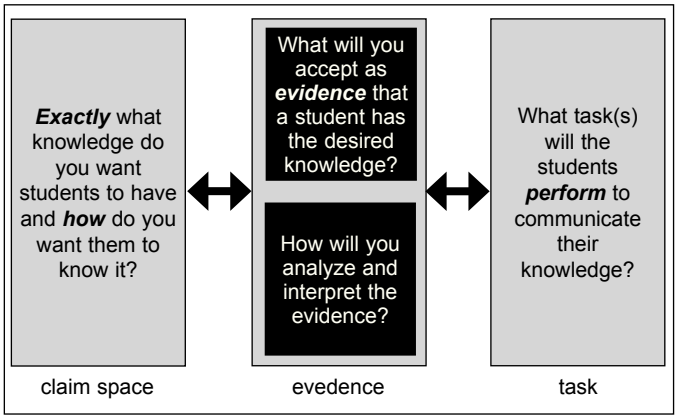

Figure 2. Simplified representation of three critical components of the evidence centered design process and their reciprocal relationships.

While the claims one wishes to make or verify are about the student, they are linked to the forms of evidence that would provide support for those claims – the warrants in support of each claim. The evidence statements associated with given sets of claims capture the features of work products or performances that would give substance to the claims. This includes which features need to be present and how they are weighted in any evidentiary scheme – i.e., what matters most and what matters least or not at all. For example, if the evidence in support of a claim about a student's knowledge of the laws of motion is that the student can analyze a physical situation in terms of the forces acting on all the bodies, then the evidence might be a free body diagram that is drawn with all the forces labeled including their magnitudes and directions.

The precision that comes from elaborating the claims and evidence statements associated with a domain of knowledge and skill pays off when one turns to the design of tasks or situations that can provide the requisite evidence. In essence, tasks are not designed or selected until it is clear what forms of evidence are needed to support the range of claims associated with a given assessment situation. The tasks need to provide all the necessary evidence and they should allow students to "show what they know" in a way that is as unambiguous as possible with respect to what the task performance implies about student knowledge and skill – i.e., the inferences about student cognition that are permissible and sustainable from a given set of assessment tasks or items. Interesting applications of the ECD approach can be found in the large-scale

assessment programs under development and validation by the two large consortia of states developing assessments aligned to the new Common Core State Standards in mathematics and English language arts in the United States (see PARCC, 2014; SBAC, 2014).

It is beyond the scope of this paper to also consider issues of measurement and statistical inference with regard to student performance on a given assessment. Nevertheless, it is important to note that the interpretation component of the Assessment Triangle, as well as application of an ECD framework for assessment design, often relies upon application of a formal measurement model. A variety of such models are available for use in contexts ranging from classroom assessment to large-scale standardized tests of the types used in national and international assessment programs (see e.g., Pellegrino et al. 2001; Pellegrino, DiBello, & Brophy, 2014). The paper by de la Torre and Minchen (2014) in this issue provides an excellent discussion of the benefits of a particular class of such models, known as Diagnostic Classification Models, when the goal of the assessment design is to obtain interpretive information closely tied to a detailed cognitive model of student knowledge and skills. In such a case, there is a close coupling among the elements of the assessment triangle that is manifest in details of the assessment design that includes rules for making inferences from the evidence obtained across a set of carefully designed tasks. Often, the goal of obtaining such detailed diagnostic information is its use as part of a classroom formative assessment process.

IV. Implications For Assessment Design

The Design and Use of Classroom Assessment

Learning scientists generally argue that classroom assessment practices need to change to better support learning (also see Shepard, 2000). The content and character of assessments need to be significantly improved to reflect the latest empirical research on learning and, given what we now know about learning progressions, the gathering and use of assessment information and insights should become a part of the ongoing learning process. This latter point further suggests that teacher education programs should provide teachers with a deep understanding of how to use assessment in their instruction. Many educational assessment experts believe that if assessment, curriculum, and instruction were more integrally connected, student learning would improve (e.g., Pellegrino et al., 1999; Stiggins, 1997).

According to Sadler (1989), three elements are required if teachers are to successfully use assessment to promote learning:

(1) A clear view of the learning goals (derived from the curriculum)

(2) Information about the present state of the learner (derived from assessment)

(3) Action to close the gap (taken through instruction)

Each of these three elements informs the other. For instance, formulating assessment procedures for classroom use can spur a teacher to think more specifically about learning goals, thus leading to modification of curriculum and instruction. These modifications can, in turn, lead to refined assessment procedures, and so on. The mere existence of classroom assessment along the lines discussed here will not ensure effective learning. The clarity and appropriateness of the curriculum goals, the validity of the assessments in relationship to these goals, the interpretation of the assessment evidence, and the relevance and quality of the instruction that ensues are all critical determinants of the outcome.

Effective teaching must start with a model of cognition and learning in the domain. For most teachers, the ultimate goals for learning are established by the curriculum, which is usually mandated externally (e.g., by state curriculum standards). But the externally mandated curriculum does not specify the empirically based cognition and learning outcomes that are necessary for assessment to be effective. As a result, teachers (and others responsible for designing curriculum, instruction, and assessment) must fashion intermediate goals that can serve as an effective route to achieving the externally mandated goals and, to do so effectively, they must have an understanding of how students represent knowledge and develop competence in the domain. Formative assessment should be based in cognitive theories about how people learn particular subject matter to ensure that instruction centers on what is most important for the next stage of learning, given a learner's current state of understanding.

Pre-service and professional development are needed to help teachers formulate models of learning progressions so they can identify students' naïve or initial sense-making strategies and build on those to move students toward more sophisticated understandings. This will increase teachers' diagnostic expertise so they can make informed decisions about next steps for student learning. Several cognitively-based approaches to instruction and assessment have been shown to have a positive impact on student learning, including the Cognitively Guided Instruction program (Carpenter, Fennema, & Franke, 1996) and others (Cobb et al., 1991; Griffin & Case, 1997).

The Design and Use of Large-Scale Assessment

Large-scale assessments are further removed from instruction but can still benefit learning if well designed and properly used. If the principles of design identified above were applied, substantially more valid, useful, and fair information would be gained from large-scale assessments. However, before schools, districts, states, or nations can fully capitalize on contemporary theory and research, they may need to substantially change how they approach large-scale assessment. Specifically, they must relax some of the constraints that currently drive many large-scale assessment practices, as follows.

Large-scale summative assessments should focus on the most critical and central aspects of learning in a domain – as identified by curriculum standards and informed by cognitive research and theory. Large-scale assessments typically are based on models of learning that are less detailed than classroom assessments. For summative purposes, one might need to know whether a student has mastered the more complex aspects of multicolumn subtraction, including borrowing from and across zero, whereas a teacher needs to know exactly which procedural errors lead to mistakes. Although policymakers and parents may not need all the diagnostic detail that would be useful to a teacher and student during the course of instruction, large-scale summative assessments should be based on a model of learning that is compatible with and derived from the same set of knowledge and assumptions about learning as classroom assessment.

Research on cognition and learning suggests a broad range of competencies that should be assessed when measuring student achievement, many of which are essentially untapped by current assessments. Examples are knowledge organization, problem representation, strategy use, metacognition, and participatory activities (e.g., formulating questions, constructing and evaluating arguments, contributing to group problem-solving). These are important elements of contemporary theory and research on the acquisition of competence and expertise and are discussed and illustrated in detail in the various references mentioned earlier in the section on learning progressions. Large-scale assessments should not ignore these aspects of competency and should provide information about these aspects of the nature of student understanding, rather than simply ranking students according to general proficiency estimates. If tests are based on a research-grounded theory of cognition and learning, those tests can provide positive direction for instruction, making "teaching to the test" productive for learning rather than destructive (this point is discussed further below).

Unfortunately, given current constraints of standardized test administration, only limited improvements in large-scale assessments are possible. These constraints include the need to provide reliable and comparable scores for individuals as well as groups, the need to sample a broad set of curriculum standards within a limited testing time per student, and the need to offer cost-efficiency in terms of development, scoring, and administration. To meet these kinds of demands, designers typically create assessments that are given at a specified time, with all students being given the same (or parallel) tests under strictly standardized conditions (often referred to as on-demand assessment). Tasks are generally of the kind that can be presented in paper-and-pencil format that students can respond to quickly, and that can be scored reliably and efficiently. As a result, learning outcomes that lend themselves to being assessed in these ways are assessed, but aspects of learning that cannot be observed under such constrained conditions are not. Designing new assessments that capture the complexity of cognition and learning will require examining the assumptions and values that currently drive assessment design choices and breaking out of the current paradigm to explore alternative approaches to large-scale assessment, including innovative uses of technology (see e.g., Quellmalz & Pellegrino, 2009; Pellegrino et al., 2014).

V. Balanced Assessment Systems

Many different assessments are used in schools, with each serving varying needs and different audiences. Perhaps the biggest divide is between external, large-scale assessments for purposes of summative evaluation and comparison by policy makers, and classroom assessments designed to assist teachers in their instructional work. One result of this variety is that users can become frustrated when different assessments have conflicting achievement goals and results. Sometimes such discrepancies can be meaningful and useful, such as when assessments are explicitly aimed at measuring different school outcomes. More often, however, conflicting assessment goals and feedback cause much confusion for educators, students, and parents. In this section we describe a vision for coordinated systems of multiple assessments that work together, along with curriculum and instruction, to promote learning.

In many education systems worldwide, assessment is focused on classroom activities designed to provide information about the progress of learning and external, large-scale standardized assessments play a relatively minor or secondary role in the educational system (see National Research Council, 2003). In the United States, however, the resources invested in producing and using large-scale tests – in terms of money, instructional time, research, and development – far outweigh the investment in the design and use of effective classroom assessment (see e.g., Kaestle, 2013). And unfortunately, there is ample evidence that the large-scale assessments in use today in the U.S. and elsewhere negatively impact classroom instruction and assessment. For instance, as discussed earlier, teachers feel pressure to teach to the test, which (given the focus of today's assessments on disconnected facts and skills) results in a narrowing of instruction. This would not necessarily be a problem if the assessments found on such tests were of higher quality and represented the full range of levels of thinking and reasoning that we desire for students to attain. Then we would have tests worth teaching towards and the tasks would be much closer to those that are useful in the context of classroom instruction to promote student learning and engagement. They would be tasks and performances that merit the time and attention of teachers and students. If that was true, then we would not have the problem that exists now because teachers model their own classroom tests after the highly limiting and less-than-ideal tasks found on typical standardized tests (Koretz, 2009; Linn, 2000; Shepard, 2000). Given that they will engage in such a modeling exercise when the external tests matter for purposes such as accountability, it would be far better if what they were modeling constituted high quality and valid assessments of student achievement. So, in addition to the need to strike a better balance between classroom and large-scale assessment, we also need to coordinate systems of assessments that collectively support a common set of learning and teaching goals, rather than work at cross-purposes. To this end, an assessment system should exhibit three properties: comprehensiveness, coherence, and continuity.

By *comprehensiveness*, I mean that a range of measurement approaches should be used to provide a variety of evidence to support educational decision-making. No single test score can be considered a definitive measure of a student's competence. Multiple measures enhance the validity and fairness of the inferences drawn by giving students various ways and opportunities to demonstrate their competence. Multiple measures can also be used to provide evidence that improvements in test scores represent real gains in learning, as opposed to score inflation due to teaching narrowly to one particular test (e.g., Koretz, 2009).

By *coherence*, I mean that the models of student learning underlying the various external and classroom assessments within a system should be compatible. While a large-scale assessment might be based on a model of learning that is coarser than that underlying the assessments used in classrooms, the conceptual base for the large-scale assessment should be a broader version of one that makes sense at the finer-grained level (Mislevy, 1996). In this way, the external assessment results will be consistent with the more detailed understanding of learning underlying classroom instruction and assessment. As one moves up and down the levels of the system, from the classroom through the school, district, and state, assessments along this vertical dimension should align. As long as the underlying models of learning are consistent, the assessments will complement each other rather than present conflicting goals for learning.

Finally, an ideal assessment system would be designed to be *continuous*. That is, assessments should measure student progress over time, akin more to a videotape record rather than to the snapshots provided by most current tests. To provide such pictures of progress, multiple sets of observations over time must be linked conceptually so that change can be observed and interpreted. Models of student progress in learning should underlie the assessment system, and tests should be designed to provide information that maps back to the progression. Figure 3 provides a graphical illustration of what an assessment system might look and some of the factors that would serve to achieve balance and support these three principles.

Figure 3 demonstrates that such a system would be (a) coordinated across levels, (b) unified by common learning goals, and (c) synchronized by unifying progress variables. No existing system of assessments has these design features and meets all three criteria of comprehensiveness, coherence, and continuity, but there are examples of assessments that represent steps toward these goals. For instance, Australia's Developmental Assessment program (Forster & Masters, 2001; Masters & Forster, 1996) and the BEAR assessment system (Wilson, Draney, & Kennedy, 2001; Wilson & Sloane, 2000) show how progress maps can be used to achieve coherence between formative and summative assessment, as well as among curriculum, instruction, and assessments. Progress maps also enable the measurement of growth (thus meeting the continuity criterion). The Australian Council for Educational Research has produced an excellent set of resource materials for teachers to support their use of a wide range of assessment strategies – from written tests to portfolios to projects at the classroom level – that can all be designed to link back to the progress maps (thus meeting the criterion of comprehensiveness).

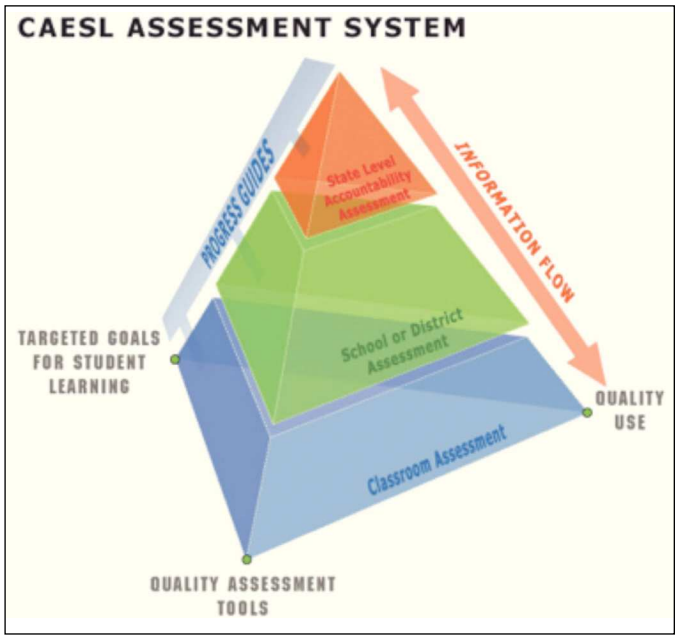

Figure 3. Center for Assessment and Evaluation of Student Learning (CAESL) representation of a coordinated, multilevel assessment system (from Herman, Wilson, Shavelson, Timms, & Schneider, 2005, reprinted with permission of the authors).

VI. Moving Forward: Necessity and Opportunity

Because assessments are tied to claims we would like to substantiate about students' competencies, new approaches to assessment must be treated as a process of gathering evidence to confirm or disconfirm particular claims (Gorin, 2013). That evidence, which in a system of assessments can come from multiple sources, can be used to improve both how they are taught and how and what students are learning. The evidence might include a range of activities ranging from simple to complex performance tasks pursued within classrooms as well as assessments external to regular classroom activities (Bennett, 2013). Pellegrino et al. (2014) have described in some detail such a systems approach for science assessment. The description they provide is designed to promote the vision of science learning and teaching associated with the U.S. National Research Council's Framework for K-12 science education (National Research Council, 2012) and the derivative Next Generation Science Standards (Achieve, 2013).

Digital technologies hold great promise for helping to bring about many of the changes in assessment that many believe are necessary. Technologies available today and innovations on the immediate horizon can be used to access information, create simulations and scenarios, allow students to engage in learning games and other activities, and enable collaboration among students. Such activities make it possible to observe, document, and assess students' work as they are engaged in natural activities – perhaps reducing the need to separate formal assessment for accountability from learning in the moment (e.g., Behrens & DiCerbo, 2013). Technologies will certainly make possible the greater use of formative assessment that in turn has been shown to significantly impact student achievement. Digital activities may also provide information about non-cognitive abilities, such as persistence, creativity, and teamwork that current testing approaches cannot. Juxtaposed with the promise is the need for considerable work to be done on issues of scoring and interpretation of evidence before such embedded assessment can be useful for these varied purposes.

Many issues, including some alluded to above, have been discussed and debated among educators and assessment experts for many years. As part of those discussions it is now widely recognized that large-scale standardized testing has exerted a greater and greater influence over American schooling (Kaestle, 2013; Linn, 2013). At the same time, it has been shown repeatedly that teachers have the largest impact on education of any in-school factor. And it is what teachers do and what they teach and how they assess in classrooms that give teachers that influence. If teachers and schools are to enable the kind of transferable learning required of young people in contemporary society, assessments will need to support curriculum and teaching focused on such learning, along with traditional basic skills. New assessment systems, grounded in new standards, should include the features described earlier in this paper.

Criteria for such assessment systems should be rigorous and ambitious, while taking account in the near-term of what is achievable financially, logistically, technologically, and scientifically. The path to reaching more ambitious education goals is likely to traverse distinct phases rather than occurring in one giant leap. Given where we are today and what should be feasible in the near-term, the following set of indicators has been suggested for use in evaluating whether assessment systems and their components meet the five criteria discussed in Section I (see Darling-Hammond et al., 2013 for additional details).

Indicators of Quality in a System of Next Generation Assessments

(1) Assessment of higher-order cognitive skills

✓ A large majority of items and tasks (at least two-thirds) evaluate the conceptual knowledge and applied abilities that support transfer (e.g., depth of knowledge levels 2, 3, or 4 in Webb's (1997) taxonomy or the equivalent).

✓ At least one-third of the assessment content in mathematics and at least one-half in English language arts should evaluate higher-order skills that allow students to become independent thinkers and learners (DOK levels 3 or 4).

(2) High-fidelity assessment of critical abilities

Critical abilities outlined in the standards are evaluated using high-fidelity tasks that use the skills in authentic applications:

✓ Research, including analysis and synthesis of Information

✓ Experimentation and evaluation

✓ Oral communications – speaking and listening

✓ Written communications – reading and writing

✓ Use of technology for accessing, analyzing, and communicating information

✓ Collaboration

✓ Modeling, design, and problem solving using quantitative tools

(3) Standards that are internationally benchmarked

✓ Calibration to PISA, international baccalaureate, or other internationally comparable assessment (based on evaluation of content comparability, performance standards, and analysis of student performance on embedded items).

(4) Items that are instructionally sensitive and educationally valuable

✓ Research that confirms instructional sensitivity

✓ Rich feedback on student learning and performance

✓ Tasks that reflect and can guide valuable instructional activities

(5) Assessments that are valid, reliable, and fair

✓ Evidence that the intended knowledge and skills are well measured

✓ Evidence that scores are related to the abilities they are meant to predict

✓ Evidence that the assessments are well-designed and valid for each intended use – and that uses are appropriate to the test purposes and validity evidence.

✓ Evidence that the assessments are unbiased and fairly measure the knowledge and skills of students from different language, cultural, and income backgrounds, as well as students with learning differences.

✓ Evidence that the assessments measure students learning accurately along a continuum of achievement, consistent with the purposes the assessments are intended to serve.

Educational entities – nations, states, provinces etc. – should evaluate the sets of assessments they currently have and/or develop against these criteria, and they should use their assessments in ways for which they have been appropriately validated. Doing so will help ensure positive consequences of assessment for instruction and student learning. To return to a quote from the Gordon Commission (2013b) mentioned earlier in this paper:

"The best assessments can accelerate the acquisition of 21st century knowledge and competencies if they guide the actions of teachers and enable students to gauge their progress. To do so, the tasks and activities in the assessments must be models worthy of the attention and energy of teachers and students."

Transforming educational assessment in the ways proposed depends on a systems approach that includes multiple factors. Among these are advances in cognitive theory and research and applications of technology combined with investments in teacher knowledge and accompanying changes in educational policies. Policy makers at all levels need to actively promote this much-needed transformation of current assessment practice. An open question is whether such a systems approach is attainable across the levels of educational policy and practice that are typically operative and at scales ranging from local districts, to states, nations, and even at the international assessment level.

Resumen ampliado

La evaluación es considerada a menudo como una influencia negativa en la enseñanza-aprendizaje por buena parte de la comunidad educativa –tanto en el ámbito aplicado como en el de la investigación–, especialmente cuando los resultados de la evaluación tienen importantes consecuencias. Este artículo plantea que si la evaluación es adecuadamente concebida, diseñada e implementada puede influir positivamente en la consecución de los objetivos de aprendizaje de los estudiantes del siglo XXI. Para defender esta tesis, se consideran tanto los pilares conceptuales de la evaluación como los principios fundamentales del diseño que constituyen la base de ese argumento, así como ejemplos de evaluaciones que cumplen esos criterios, entre los que se incluyen evaluaciones diseñadas para apoyar el proceso de enseñanza-aprendizaje en el aula junto a otras diseñadas para dar cuenta del progreso del sistema educativo.

La sección I se centra en los grandes retos de la educación del siglo XXI y en el tipo de evaluación que se necesita para poder lograr los objetivos de aprendizaje que son relevantes para el conjunto de la sociedad. La sección finaliza con una breve discusión de las cinco características principales que definen los componentes de un sistema de evaluación capaz de medir por completo objetivos o estándares de alta calidad y de promover la evaluación de un aprendizaje más profundo: (1) la evaluación de habilidades cognitivas de orden superior, (2) una evaluación de la capacidad crítica de alta fidelidad, (3) estándares con puntos de referencia internacionales, (4) la utilización de preguntas que sean sensibles a la instrucción y valiosas desde el punto de vista educativo y (5) evaluaciones que sean fiables, válidas y justas. En la sección VI del artículo se vuelve a estas cinco características y criterios para valorar lo conseguido. Determinar el camino que nos permita crear sistemas de evaluación para lograr las metas establecidas en el sistema educativo y que cumplan con los criterios anteriores constituye un auténtico desafío.

Las secciones II y III abordan algunas de las cuestiones fundamentales y los marcos conceptuales que son necesarios para avanzar en ese camino. La sección II analiza los objetivos y contextos de la evaluación educativa con el fin de proporcionar un marco que permita entender por qué se necesitan varios tipos de evaluación y cuáles son sus funciones en el sistema educativo. Una cuestión central es que una única evaluación no puede servir para todo y, por tanto, el diseño de una evaluación debe tener en cuenta la función que ha de realizar (e.g., formativa, sumativa, evaluación de programas) y el contexto de su utilización (e.g., clases individuales frente a distritos escolares, regiones o países). A continuación, la sección III examina tres marcos conceptuales relacionados entre sí y que son fundamentales en la conceptualización y el diseño de cualquier evaluación: (a) la evaluación como un proceso de razonamiento a partir de la evidencia, (b) la evaluación realizada desde modelos de aprendizaje formulados como progresiones de aprendizaje y (c) la utilización de un diseño centrado en la evidencia para diseñar la evaluación e interpretar sus resultados. Un aspecto clave de estos tres marcos es que el diseño y la utilización de la evaluación debe emanar de una concepción clara de qué significa la competencia en un determinado dominio curricular y cómo cambia con el tiempo esa competencia en base al proceso de enseñanza-aprendizaje. Lo que ha de guiar el diseño y uso de la evaluación del rendimiento de los estudiantes son las mejores teorías, modelos y datos empíricos acerca de la naturaleza del conocimiento y del aprendizaje.

La sección IV vuelve a las implicaciones del material cubierto en la sección anterior para el diseño de evaluaciones en el aula y también a gran escala. Se señala que los estudiosos del aprendizaje habitualmente plantean que es necesario cambiar las prácticas de evaluación en el aula para favorecer el aprendizaje. Por ejemplo, hay que mejorar significativamente el contenido y el tipo o naturaleza de las evaluaciones para que reflejen los últimos avances de la investigación sobre aprendizaje; por otro lado, dado lo que ahora se sabe acerca de las progresiones de aprendizaje, este conocimiento así como la recogida y utilización de información procedente de la evaluación deberían formar parte del proceso de formación continua. Esta última cuestión sugiere además que los programas diseñados tanto para profesores en prácticas como ya en activo deberían ayudar a ambos colectivos a conocer a fondo cómo utilizar la evaluación en el proceso de instrucción. Por lo que respecta a los programas de evaluación a gran escala, a menudo son innecesariamente restrictivos y miden solo lo que es fácil de evaluar, con formatos diseñados para mejorar la eficiencia en la recogida de datos y en el ahorro de costes en relación a la corrección de las respuestas a las preguntas de las pruebas administradas. Por el contrario, se defiende que la evaluación a gran escala debería centrarse en los aspectos más importantes y críticos del aprendizaje en un dominio de conocimiento, tal como han sido identificados en los objetivos curriculares y refrendados por la teoría y la investigación cognitiva. Diseñar nuevas evaluaciones que capturen la complejidad de la cognición y el aprendizaje va a requerir examinar muchos de los supuestos y valores que en la actualidad guían la elección del diseño de evaluación y también romper con el paradigma actual en el diseño de evaluaciones a gran escala para explorar vías alternativas, incluyendo un uso innovador de la tecnología.

La sección V considera los componentes de un sistema equilibrado de evaluación que incluya la evaluación en el aula junto a las evaluaciones utilizadas por los distritos escolares, regiones y países con el fin de supervisar. Se argumenta que en países como Estados Unidos es necesario conseguir un mayor equilibrio entre la evaluación en el aula y a gran escala: en lugar de contar con distintos programas de evaluación que sirvan a objetivos dispares, se necesita coordinar sistemas de evaluación que trabajen al unísono para conseguir un conjunto común de objetivos de enseñanza y aprendizaje. Para ello, dicho sistema de evaluación debería mostrar tres propiedades, que se describen brevemente: amplia cobertura, coherencia y continuidad. Por amplia cobertura se entiende que se utiliza toda una gama de métodos de medida para obtener evidencia que contribuya a tomar decisiones en el ámbito educativo. Coherencia significa que dentro del sistema de evaluación han de ser compatibles los modelos de aprendizaje del estudiante que subyacen a las evaluaciones en el aula y a distintas evaluaciones externas. Continuidad significa que las evaluaciones deberían medir el progreso de los estudiantes a lo largo del tiempo, más en línea con la metáfora de una cinta de video que con la foto fija que ofrecen la mayoría de los tests.

La sección VI vuelve a los cinco componentes del sistema de evaluación planteado en la sección I y concluye describiendo brevemente indicadores clave de calidad que hay que tener muy presentes si se desea poner en marcha un sistema coherente de evaluación como parte del proceso de transformación educativa en el siglo XXI. Las correspondientes instancias educativas a nivel de país, región, provincia, etc. deberían examinar en relación a esos criterios los programas de evaluación que tienen actualmente en marcha o que proyectan diseñar. Asimismo, deberían asegurarse de utilizar los resultados de sus evaluaciones para fines que hayan sido adecuadamente validados. Esta forma de proceder puede contribuir a que la evaluación tenga consecuencias positivas en la enseñanza y el aprendizaje de los estudiantes.

Transformar la evaluación educativa del modo propuesto requiere una aproximación sistémica que incluye muchos factores, entre ellos los avances en la teoría e investigación cognitiva y las aplicaciones de la tecnología combinadas con inversión en la formación docente y cambios concomitantes en las políticas educativas. Las autoridades educativas en cualquier nivel (regional, nacional,…) tienen que promover esta transformación tan necesaria de la práctica actual de evaluación. Una pregunta que queda en el aire para todas ellas es que consideren si tal aproximación sistémica se puede lograr a nivel nacional e internacional.

Conflict of Interest

The author of this article declares no conflict of interest.

References

Achieve (2013). *Next Generation Science Standards*. Retrieved from http://www.nextgenscience.org/

Alonzo, A. C., & Gotwals, A. W. (2012). *Learning progression in science: Current challenges and future directions*. Rotterdam, Netherlands: Sense Publishers.

American Association for the Advancement of Science. (2001). *Atlas of science literacy*. Washington, DC: Author.

Behrens, J. T., & DiCerbo, K. E. (2013). *Technological implications for assessment ecosystems: Opportunities for digital technology to advance assessment*. Published by the Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/publications_reports/assessment_psychometric.html

Bennett, R. E. (2013). *Preparing for the future: What educational assessment must do*. Published by the Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/publications_reports/assessment_paradigms.html.

Bereiter, C., & Scardamalia, M. (2013). *What will it mean to be an educated person in the mid-21st century?* Published by the Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/publications_reports/assessment_paradigms.html

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7-73.

Bransford, J. D., Brown, A. L., Cocking, R. R., Donovan, M. S., & Pellegrino, J. W. (Eds.) (2000). *How people learn: Brain, mind, experience, and school* (expanded ed.). Washington, DC: National Academies Press.

Carpenter, T., Fennema, E., & Franke, M. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *Elementary School Journal*, 97(1), 3-20.

Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., & Perlwitz, M. (1991). Assessment of a problem-centered second-grade mathematics project. *Journal for Research in Mathematics Education, 22*(1), 3-29.

Common Core State Standards Initiative (2010a). *English language arts standards*. Washington, DC: National Governors Association and Council of Chief State School Officers. Retrieved from http://www.corestandards.org/the-standards/english-language-artsstandards.pdf

Common Core State Standards Initiativa (2010b). *Mathematics standards*. Washington, DC: National Governors Association and Council of Chief State School Officers. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. New York, NY: Columbia University, Teachers College, Consortium for Policy Research in Education, Center on Continuous Instructional Improvement.

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. NY: Teachers College Press.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., Herman, J., Pellegrino, J. W., Abedi, J., Aber, J. L., Baker, E., … Steel, C. M. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved from: http://edpolicy.stanford.edu/publications/pubs/847

Daro, P., Mosher, F. A., Corcoran, T., Barrett, J., & Consortium for Policy Research in Education. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction*. Philadelphia, PA: Consortium for Policy Research in Education.

Deane, P., & Song, Y. (2014). A case study in principled assessment design: Designing assessments to measure and support the development of argumentative reading and writing skills. *Psicología Educativa, 20*, 99-108.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*, 89-97.

Duncan, R. G., & Hmelo-Silver, C. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal for Research in Science Teaching, 46*, 606-609.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.) (2007). *Taking science to school: Learning and teaching science in grade K-8*. Washington DC: The National Academies Press.

Forster, M., & Masters, G. (2001). *Progress maps*. Victoria, Australia: Australian Council for Educational Research.

Gordon Commission on the Future of Assessment in Education (2013b). *Policy report*. Retrieved from http://www.gordoncommission.org/publications_reports.html

Gordon Commission on the Future of Assessment in Education (2013a). *Technical report*. Retrieved from http://www.gordoncommission.org/publications_reports.html

Gorin, J. S. (2013). *Assessment as evidential reasoning*. Published by the Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/publications_reports/assessment_psychometric.html

Griffin, S., & Case, R. (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education, 3*(1), 1-49.

Herman, J. L., Wilson, M. R., Shavelson, R., Timms, M., and Schneider, S. (2005, April). *The CAESL assessment model*. Paper presented at American Educational Research Association annual conference, Montreal, Canada.

Hickey, D., & Pellegrino, J. W. (2005). Theory, level, and function: Three dimensions for understanding transfer and student assessment. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 251-293). Greenwich, CO: Information Age Publishing.

Kaestle, C. (2013). *Testing policy in the United States: A historical perspective*. Published by the Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/publications_reports/assessment_education.html

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.) (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

Koretz, D. (2009). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R. L. (2013). *Test-based accountability*. Published by the Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/publications_reports/assessment_education.html

Marion, S., & Pellegrino, J. W. (2006, Winter). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, 47-57.

Masters, G., & Forster, M. (1996). *Progress maps. Assessment resource kit*. Victoria, Australia: Commonwealth of Australia

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379-416.

Mislevy, R. J., & Haertel, G. (2006). Implications of evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25*, 6-20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Committee on a Conceptual Framework for New K-12 Science Education Standards, Board on Science Education. Washington, DC: National Academies Press.

National Research Council (2003). *Assessment in support of learning and instruction: Bridging the gap between large-scale and classroom assessment*. Washington, DC: National Academies Press.

Ng, P. T. (2008). Educational reform in Singapore: From quantity to quality. *Education Research on Policy and Practice, 7*, 5-15.

PARCC (Partnership for Assessment of Readiness for College and Careers) (2014). *The PARCC assessment: Item development*. Retrieved from http://www.parcconline.org/assessment-development

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (vol. 24, pp. 307-353). Washington, DC: American Educational Research Association.

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

Pellegrino, J. W., DiBello, L. V., & Brophy, S. (2014). The science and design of assessment in engineering education. In A. Johri & B. Olds (Eds.), *Cambridge handbook of engineering education research* (pp. 571-598). Cambridge, England: Cambridge University Press.

Pellegrino, J. W., & Hickey, D. (2006). Educational assessment: Towards better alignment between theory and practice. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), *Instructional psychology: Past, present and future trends. Sixteen essays in honour of Erik De Corte* (pp. 169-189). Oxford, England: Elsevier.

Pellegrino, J. W., & Hilton, M. L. (Eds.) (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

Pellegrino, J. W., Wilson, M., Koenig, J. & Beatty, A. (Eds.) (2014). *Developing assessments for the next generation science standards*. Washington, DC: National Academies Press.

Quellmalz, E., & Pellegrino, J. W. (2009). Technology and testing. *Science, 323*, 75-79.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*, 369-393.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

SBAC (Smarter Balanced Assessment Consortium) (2014). *The SBAC assessment: Item writing and review*. Retrieved from http://www.smarterbalanced.org/smarter-balanced-assessments/item-writing-and-review/

Schum, D. (1987*). Evidence and inference for the intelligence analyst*. Lantham, MD: University of America Press.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Snow, C. E., Burns, M., & Griffin, M. (Eds). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.

Stiggins, R. J. (1997). *Student-centered classroom assessment*. Upper Saddle River, NJ: Prentice-Hall.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (National Institute for Science Education and Council of Chief State School Officers Research Monograph No. 6). Washington, DC: Council of Chief State School Officers.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Wiliam, D. (2007). Keeping learning on track: formative assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053-1098). Greenwich, CT: Information Age Publishing.

Wilson, M., Draney, K., & Kennedy, C. (2001). *GradeMap* [computer program]. Berkeley, CA: BEAR Center, University of California, Berkeley.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181-208.

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

Zieky, M. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20*, 79-87.

Appendix

Project Work in Singapore

In Singapore, Project Work (PW) is an assessment that is compulsory for all pre-university students. There is dedicated curriculum time for students to carry out their collaborative interdisciplinary project tasks over an extended period. The assessment tasks, which are set by the Singapore Examinations and Assessment Board, are designed to be sufficiently broad to allow students to carry out a project that they are interested in while meeting the task requirements.

In groups formed by the teacher, students agree on the project that the group will undertake, brainstorm and evaluate each other's ideas, and decide on how the work should be allocated. Project Work tasks result in:

-- A written report which shows evidence of the group's ability to generate, analyze and evaluate ideas for the project.

-- An oral presentation in which each individual group member is assessed on his/her fluency and clarity of speech, awareness of audience as well as response to questions. The group as a whole is also assessed in terms of the effectiveness of the overall presentation.

-- A group project file in which each individual group member submits three documents related to 'snapshots' of the processes involved in carrying out the project. These documents show the individual student's ability to generate, analyze, and evaluate (i) preliminary ideas for a project, (ii) a piece of research material gathered for the chosen project, and (iii) insights and reflections on the project.

The SEAB specifies task setting, conditions, assessment criteria, achievement standards, and marking processes. Classroom teachers carry out the assessment of all three components of PW using the assessment criteria provided by the board. All schools are given exemplar material that illustrates the expected marking standards. The Board provides training for assessors and internal moderators. Like all other assessments, the grading is both internally and externally moderated to ensure consistency in scoring.

In carrying out the PW assessment task, students are intended to acquire self-directed inquiry skills as they propose their own topic, plan their timelines, allocate individual areas of work, interact with teammates of different abilities and personalities, gather and evaluate primary and secondary research material. These PW processes reflect life skills and competencies such as knowledge application, collaboration, communication and independent learning, which prepare students for the future workplace.

Extended Experimental Investigations in Queensland

In Queensland (Australia) science courses, like those in Singapore, Hong Kong, and other Australian states, students must complete an extended experimental investigation that they design, conduct, and evaluate. In Queensland, the task is defined as follows:

Within this category, instruments are developed to investigate a hypothesis or to answer a practical research question. The focus is on planning the extended experimental investigation, problem solving and analysis of primary data generated through experimentation by the student. Experiments may be laboratory or field based. An extended experimental investigation may last from four weeks to the entirety of the unit of work. The outcome of an extended experimental investigation is a written scientific report. For monitoring, the discussion/conclusions/evaluation/recommendations of the report should be between 1500 and 2000 words.

To complete such an investigation the student must:

- Develop a planned course of action
- Clearly articulate the hypothesis or research question, providing a statement of purpose for the investigation.
- Provide descriptions of the experiment
- Show evidence of modification or student design
- Provide evidence of primary and secondary data collection and selection
- Execute the experiment(s)
- Analyze data
- Discuss the outcomes of the experiment
- Evaluate and justify conclusion(s)
- Present relevant information in a scientific report

Graduate Certificate in Secondary Education (GCSE) Task in Interactive Computer Technology, England

In England, students choose a number of domains in which to be examined as part of the high school assessment system. Most of these examinations, which are linked to high school courses, include a project-based component that typically counts for 60% of the total examination score. The project below has been used as part of the Interactive Computer Technology examination.

Litchfield Promotions works with over 40 bands and artists to promote their music and put on performances in England. The number of bands they have on their books is gradually expanding. Litchfield Promotions needs to be sure that each performance will make enough money to cover all the staffing costs and overheads as well as make a profit. Many people need to be paid: the bands; sound engineers; and lighting technicians. There is also the cost of hiring the venue. Litchfield Promotions needs to create an ICT solution to ensure that they have all necessary information and that it is kept up to date. Their solution will show income, outgoings and profit.

Candidates need to: 1) work with others to plan and carry out research to investigate how similar companies have produced a solution. The company does not necessarily have to work with bands and artists or be a promotions company; 2) clearly record and display your findings; 3) recommend a solution that will address the requirements of the task; 4) produce a design brief, incorporating timescales, purpose and target audience.

Produce a solution, ensuring that the following are addressed: 1) it can be modified to be used in a variety of situations; 2) it has a friendly user interface; 3) it is suitable for the target audience; 4) it has been fully tested. You will need to: 1) incorporate a range of: software features, macros, modeling, and validation checks - used appropriately; 2) obtain user feedback; 3) identify areas that require improvement, recommending improvement, with justification; 4) present information as an integrated document; 5) evaluate your own and others' work.