1

The Learning Sciences in Educational Assessment: An Introduction

Victor Hugo is credited with stating that "There is nothing more powerful than an idea whose time has come." In educational achievement testing,[1] a multi-billion-dollar activity with profound implications for individuals, governments, and countries, the idea whose time has come, it seems, is that *large-scale achievement tests must be designed according to the science of human learning*. Why this idea, and why now? To begin to set a context for this idea and this question, a litany of research studies and public policy reports can be cited to make the simple point that students in the United States and abroad are performing relatively poorly in relation to expected standards and projected economic growth requirements (e.g., American Association for the Advancement of Science, 1993; Chen, Gorin, Thompson, & Tatsuoka, 2008; Grigg, Lauko, & Brockway, 2006; Hanushek, 2003, 2009; Kilpatrick, Swafford, & Findell, 2001; Kirsch, Braun, & Yamamoto, 2007; Manski & Wise, 1983; Murnane, Willet, Dulhaldeborde, & Tyler, 2000; National Commission on Excellence in Education, 1983; National Mathematics Advisory Panel, 2008; National Research Council, 2005, 2007, 2009; Newcombe et al., 2009; Phillips, 2007; Provasnik, Gonzales, & Miller, 2009). According to a 2007 article in the *New York Times*, Gary Phillips, chief scientist at the American Institutes for Research, was quoted as saying, "our Asian

[1] The terms "testing" and "assessment" are used interchangeably in the present volume to denote formal measurement techniques and evaluation procedures.

economic competitors are winning the race to prepare students in math and science." Phillips made this comment in relation to a report equating the standardized large-scale test scores of grade-eight students in each of the fifty U.S. states with those of their peers in forty-five countries. Underlying the sentiment in this quote is the supposition that test scores reveal valuable information about the quality of student learning and achievement[2] for feeding future innovation and economic growth. If test scores reveal that U.S. students are underperforming relative to their peers in other countries, then learning is likely being compromised, and innovation and economic growth may also falter.

To change this potentially grim outcome, there are at least three options: Change the educational system, change the large-scale tests,[3] or change both. In the balance of this book, we focus on the second option – changing the large-scale tests. This decision does not indicate that the first and third options lack merit. In fact, the third option is ideal. However, in this first chapter, we present a rationale for why it makes sense to focus on changing tests, that is, to design and develop large-scale educational assessments based on the learning sciences. To start, we discuss the relatively poor test performance of many U.S. students as an impetus for the growing motivation in North America to enhance the information large-scale educational

[2] Although we recognize that learning and achievement sometimes connote different ideas (e.g., learning might be used in relation to the processes involved in the acquisition of knowledge and skills, and achievement might be used in relation to demonstrations of those knowledge and skills), learning and achievement are used interchangeably in this volume. The goal of most educational initiatives and institutions is to have learning and achievement overlap significantly. In addition, developers of achievement tests strive to design measures that are sensitive to progressions in learning.

[3] The focus is on large-scale educational testing because testing companies and assessment branches of government agencies have the human and financial capital to consistently develop, refine, and administrate standardized, psychometrically sound assessments based on scientific cognitive learning models for large numbers of students. Although classroom tests could also be developed from findings in the learning sciences, these are less likely to be developed according to the same models due to a lack of resources.

assessments currently provide about student learning. Next, we present well-accepted knowledge from the learning sciences about the nature of thinking, learning, and performance to help determine the types of knowledge and skill components that may be required for measurement as large-scale educational assessments are designed and developed. After that, illustrative empirical studies in the field of educational measurement are reviewed to demonstrate the nature of the attempts to design and develop assessments based on the learning sciences (also commonly referred to as *cognitive diagnostic assessments* [CDA] or *cognitively based tests*; see Leighton & Gierl, 2007). Then, we offer a view on what is needed in the field of educational assessment to incorporate and systematically evaluate cognitive models in the design and development of large-scale assessments. Finally, we present a conclusion and roadmap for the present volume that outlines the rationale and content of the next six chapters, including what may be needed to change large-scale tests and ensure they provide the information about student learning and achievement many stakeholders seek.

THE IMPETUS FOR CHANGE: LOW ACHIEVEMENT TEST RESULTS

The U.S. Department of Education (2008) posted the following summary of the results achieved by fifteen-year-old American students in reading, science, and mathematics on the Programme for International Student Assessment (PISA™) administered by the Organization for Economic Cooperation and Development (OECD, 2007, 2009; see also U.S. results on Third International Mathematics and Science Study [TIMSS], Hanushek, 2009):

1. In the 2003 PISA administration, which focused on reading literacy, U.S. students received an average score just higher than the OECD average of approximately 500 (i.e., 495 versus 494, respectively; see OECD, 2003) but lower than seventeen

OECD jurisdictions. Reading literacy scores were not compiled for U.S. students in 2006 due to an administrative problem with the test booklets.

2. In the 2006 PISA administration, which focused on science literacy, U.S. students received lower scores relative to their peers in sixteen of the other twenty-nine OECD jurisdictions and in six of the twenty-seven non-OECD jurisdictions; in two specific types of scientific literacy (i.e., explaining phenomena scientifically and using scientific evidence), U.S. students received lower scores than the OECD average (i.e., 486 versus 500, and 489 versus 499, respectively).
3. In the 2006 PISA administration, which also measured mathematical literacy, U.S. students received an average score (i.e., 474) that was lower than the OECD average of 498, and lower than twenty-three OECD jurisdictions and eight non-OECD jurisdictions.

As an introduction to this section on the impetus for change, we focus on the PISA results posted by the U.S. Department of Education for three reasons: First, as a psychometrically sound assessment, PISA results are noteworthy; second, PISA provides a broad view of how well students can apply what they are learning to novel tasks, because the assessment measures *literacy* in reading, science, and mathematics as opposed to measuring knowledge of a specific curriculum (e.g., see de Lange, 2007, for a discussion of the Third International Mathematics and Science Study [TIMSS] in relation to specific curricula); and third, forty-one countries participated in the 2003 administration of PISA (fifty-seven countries participated in 2006), and the sheer size of this endeavor renders the assessment results relevant to many stakeholders who can initiate substantial change in education policies and practices, especially in the United States.

Although the United States may be the most notable country struggling with the literacy performance of its adolescents, it is not the only

one. Students in countries such as Australia, Canada, Japan, and the United Kingdom may be performing better than American students on PISA, but there is still substantial room for improvement (OECD, 2007). For example, in the domain of scientific literacy, an average of only 1.3 percent of students across OECD countries were classified into the top category of science proficiency (i.e., level 6 on the PISA 2006 proficiency scale; see OECD, 2007, p. 14). Finland and New Zealand had 3.9 percent of their students classified into this top category, whereas countries such as Australia, Canada, Japan, and the United Kingdom only had between 2 and 3 percent of their students meet this high level of performance (OECD, 2007). By including level 5, the next highest category of science proficiency, the percentage of students considered to be high performers across OECD countries rose to an average of 9 percent. Again Finland and New Zealand had the highest percentage of students classified into categories 5 or 6 (21 and 18 percent, respectively). Countries such as Australia, Canada, and Japan had between 14 and 16 percent of students classified into one of these top two categories. Twenty-five countries had less than 5 percent of their students reaching the highest categories (levels 5 and 6) of science proficiency, and fifteen had less than 1 percent.

There are many incentives for wanting students to be classified into the highest level of proficiency in core academic domains (see Hanushek, 2005). For example, in science, the classification of students into the highest category provided by PISA is assumed to mean that students are able to engage the types of higher-order thinking skills that will be necessary for many twenty-first-century jobs (Hanushek, 2009). These higher-order thinking skills include (a) identifying, explaining, and applying scientific knowledge in a variety of multifaceted life situations; (b) connecting distinct sources of information and explanations, and making use of evidence from those sources to defend decisions; (c) demonstrating advanced thinking and reasoning, and using scientific understanding to justify solutions to novel scientific and technological situations; and (d) using scientific knowledge and

developing arguments in support of recommendations and decisions that focus on personal, social, or global situations. A student classified into one of the top categories of science performance is arguably better prepared than a student classified into the lowest level for tackling science in the classroom and ultimately contributing and pursuing scientific innovation in the labor market. At the lowest level of proficiency, students know and can do few things. According to the OECD (2007, p. 14), "At Level 1, students have such a limited scientific knowledge that it can only be applied to a few, familiar situations. They can present scientific explanations that are obvious and that follow explicitly from given evidence." When countries are ranked by the percentages of fifteen-year-olds classified above the lowest level of 1 (i.e., levels 2, 3, 4, 5, and 6) on the PISA proficiency scale, Finland is ranked first, Canada is ranked fourth, and the United States is ranked thirty-sixth of fifty-seven countries (OECD, 2007, p. 20, table 1).

Policy makers recognize that large-scale educational testing for students in kindergarten through grade twelve can be a powerful measure of accountability and a driver of educational reform and improvement. There is also mounting hope in the United States that federal legislation, in the form of the No Child Left Behind Act (NCLB, 2002; see also Race to the Top Fund), can improve educational outcomes by mandating states that accept Title 1 funds to administer annual large-scale educational testing in at least seven grade levels (Koretz & Hamilton, 2006, p. 531). On the one hand, the mandate of NCLB makes sense given that results from national (and international) large-scale educational assessments provide a snapshot of student achievement and the success of educational systems for attaining specific outcomes. At a minimum, the results from these assessments allow us to generate conclusions about whether students are performing as expected (i.e., being classified into projected categories of proficiency) or whether performance could be improved. On the other hand, it seems reasonable to ask whether these test results will provide additional information about student learning aside from their performance at a single point in time, including

partial knowledge and skills, misconceptions, and areas of genuine cognitive strength. For example, could these test results shed light on why U.S. students are struggling with explaining phenomena scientifically and using scientific evidence (see OECD, 2009)? Given the cost and time spent on designing, developing, and administering these large-scale educational assessments, it seems wasteful not to have them provide at least some information about the possible sources of students' learning impasses and incorrect responses. In fact, these tests do not even provide unequivocal information about the quality of higher-order thinking students possess, because the items are not typically evaluated for whether they elicit in students the appropriate thinking skills of interest (Schmeiser & Welch, 2006). In short, many of these large-scale educational tests have not been designed to provide information on the quality of students' thinking and learning. Consequently, a poor or good test result conveys little information about how students think, reason, or problem-solve with the knowledge and skills presented in test items. Hence, little information can be gleaned from many large-scale test results about possible student misconceptions or other pedagogically based reasons students may struggle with core academic concepts and skills.

Billions of dollars are spent every year on education, but these expenditures do not translate to superior test results for American students (Hanushek, 2005). In fact, there appears to be little association between educational expenditures and large-scale achievement test scores (see Hanushek, 2009, p. 49). The cost for public elementary and secondary education in the United States was estimated at approximately $543 billion for the 2009–2010 school year (Hussar & Bailey, 2008), and the national average current expenditure per student was estimated at around $10,844 for 2009–2010, which rose from $9,683 in 2006–2007 (Zhou, 2009). These figures are startling not because the United States spends too little or too much, but rather because one would expect a significant, positive correlation to exist between amounts spent on education and educational outcomes. Moreover, the United States is a leader

in research and innovation, and one would expect that children educated in one of the richest and most resourceful countries in the world would perform better than children in countries who do not have the same financial or human intellectual capital.

The absence of a relationship between educational expenditures and large-scale educational test scores might lead one to conclude that infusing money into teacher professional development or other initiatives to boost instruction is wasteful because it does not translate into higher test scores. Alternatively, one could conclude that money spent on professional development or other initiatives is working, but we do not have the appropriate tools to measure their benefits (see Polikoff, 2010). For example, suppose we tried to use large-scale achievement tests to measure the learning outcomes derived from newly funded and innovative instructional programs. Suppose further that the learning goals driving these innovative instructional programs were focused on the quality of students' thinking processes, such as the nature of their representations and search strategies for problem solving, the depth and breadth of the inter-relations among their networks of knowledge and skills, novel frames of reference, and their combinatorial mechanisms to use analogy and metaphor. If this were the case, then learning strides could be occurring but would be missed with large-scale testing (see Hanushek, 2005), because current large-scale tests have not been designed to measure these thinking processes. Students taught to engage specific thinking processes might find few outlets in traditional[4] constructed-response and multiple-choice items to show off their newfound competencies. If the tests students took failed to measure what teachers were trying to teach, then it would not be surprising to find relatively poor test results. Pumping money into schools to

[4] The term "traditional tests" is used in the present chapter to denote test-item design and development that is not formally based on learning scientific research and commonly based on historical practice, such as the use of Bloom's taxonomy to develop test items of varying difficulty levels. Most operational large-scale educational tests are traditional (Ferrara & DeMauro, 2006).

enhance instruction without a concomitant effort to change or revise large-scale educational achievement testing might even be viewed as a setup to fail – as missing the boat in detecting any gains or improvements achieved in student learning and thinking (see Hanushek, 2009; Mislevy, 1993).

In sum, the large-scale educational assessments that proved effective decades ago may no longer be sufficient to measure twenty-first-century knowledge and skills. We now live in a time when most students are digitally prodigious, engaging multiple modes of electronic communication, and cultivating informal networks of expression, discussion, and collaboration outside of the classroom more often than inside the classroom (Collins, Halverson, & Brown, 2009; see also Russell, 2005; Shute et al., 2010; Thomas, Li, Knott, & Zhongxiao, 2008). Undoubtedly students are using sophisticated thinking processes to learn and navigate through these complex systems of communication, most of which are underwritten by technological gadgets. The rate at which technology changes and the scale of students' ability to learn and pick up new tools suggest that complex problem solving is occurring. Our task is to figure out how to assess it (NRC, 2005; Pashler, Rohrer, Cepeda, & Carpenter, 2007). As titanic as the challenge of educational reform appears to be, however, there is optimism that inroads may take place with the design and development of large-scale educational tests based on advances in the learning sciences. These new types of tests might even inform us about the knowledge and skills that characterize adaptability, innovation, and higher-order thinking for job and economic growth in the twenty-first century. This is the idea whose time has come.

THE LEARNING SCIENCES

The learning sciences are an inter-disciplinary domain of study. Although its foundations can be traced back to educational technology, socio-cultural studies, computing science, anthropology,

and cognitive science, the main focus is consistently on what is needed to make human learning more successful. To maximize learning, the mechanisms that enhance or hinder learning are identified and investigated. In this respect, cognitive science has played a particularly pivotal role. Its influence can be traced back to Piaget's *constructivism* (Piaget & Inhelder, 1967), which emphasized the qualitatively different structure of children's knowledge and thinking in relation to adult knowledge and thinking, as well as the instructional importance of recognizing these differences as new knowledge is introduced to learners (e.g., Siegler, 2005). Sawyer (2006, p. 2) emphasized this point:

> [B]eginning in the 1970s, a new science of learning was born – based in research emerging from psychology, computer science, philosophy, sociology, and other scientific disciplines. As they closely studied children's learning, scientists discovered that instruction was deeply flawed. By the 1990s, after about twenty years of research, learning scientists had reached a consensus on the following basic facts about learning – a consensus that was published by the United States National Research Council. (see Bransford, Brown, & Cocking, 2000)

According to Sawyer, the five basic facts about learning are the following:

1. *Deep conceptual understanding is needed to apply knowledge*. Knowledge about facts and procedures will not transfer to novel settings and will therefore be relatively useless unless students also know the situations for when these facts and procedures can be applied (see Kilpatrick et al., 2001; Kuhn, 2001). Helping students gain conceptual understanding involves helping them to recognize the structural features of problem-solving situations and not just their surface-level characteristics (Chi, Feltovich, & Glaser, 1981; Slotta & Chi, 2006).
2. *Learning, not just teaching, must be a focus*. The science of human learning emphasizes that individuals develop deep conceptual