

# Measuring 21st Century Competencies

## Guidance for Educators



Global Cities  
Education Network

A GLOBAL CITIES EDUCATION NETWORK REPORT



Global Cities  
Education Network



# MEASURING 21<sup>ST</sup> CENTURY COMPETENCIES

## GUIDANCE FOR EDUCATORS

**RAND Corporation, November 2013**

Jim Soland, Laura S. Hamilton, and Brian M. Stecher

# TABLE OF CONTENTS

Tables .....	III
Figures .....	III
Acknowledgments .....	IV
<b>1. Introduction .....</b>	<b>1</b>
The Global Cities Educational Network (GCEN) .....	1
<b>2. Examples of 21<sup>st</sup> Century Competencies .....</b>	<b>3</b>
A Closer Look at Some 21 <sup>st</sup> Century Competencies .....	4
Cognitive Competencies .....	4
Interpersonal Competencies .....	5
Intrapersonal Competencies .....	7
Conclusion .....	8
<b>3. A Framework for Selecting Measures of 21<sup>st</sup> Century Competencies .....</b>	<b>9</b>
Purpose .....	9
Instructional Considerations .....	9
Formative versus Summative .....	10
Usefulness of the Data for Students and Teachers .....	10
Influence of Testing on Teaching and Learning .....	11
Practical Considerations .....	12
Cost .....	12
Logistical Requirements .....	12
Technical Considerations .....	12
Reliability .....	13
Validity .....	13
Fairness .....	14
Conclusion .....	15
<b>4. Examples of 21<sup>st</sup> Century Assessments, by Format .....</b>	<b>16</b>
Multiple Choice .....	18
Common Core State Standards Consortia Assessments .....	20
Self-Report (Likert) .....	22
Closed-Ended Computer-Based Items .....	23
Open Response .....	24
Portfolios .....	27
Performance Assessments and Simulations .....	28
Measuring Leadership: A Lesson in the Difficulties of Assessing 21 <sup>st</sup> Century Competencies .....	31
Additional Sources of Data .....	32
Cross-Cutting Measures: A Deeper Look at the Mission Skills Assessment (MSA) .....	32
The Intent of the Mission Skills Assessment .....	32
Practical Considerations .....	33
Technical Considerations .....	34
Instructional Benefits and Challenges Associated with the MSA .....	34
Instructional Challenges .....	36
Conclusion .....	37

5. Guidelines for Measuring 21 <sup>st</sup> Century Competencies .....	38
Considerations When Adopting or Adapting Assessments of 21 <sup>st</sup> Century Competencies .....	38
Appendix. Case Studies of Cutting-Edge Measures .....	43
Alelo Language and Culture Simulations .....	43
Common Core State Standards Consortia Assessments .....	44
EcoMUVE .....	45
The Graduation Performance System .....	47
Mission Skills Assessment .....	48
PISA Collaborative Problem Solving .....	49
Queensland Performance Assessments .....	51
SimScientists .....	52
Singapore Project Work .....	53
World Savvy Challenge .....	54
References .....	56

## Tables

Table 1. Examples of 21 <sup>st</sup> Century Competencies by Category .....	4
Table 2. Considerations When Selecting Measures of 21 <sup>st</sup> Century Competencies .....	9
Table 3. Examples of Measures of 21 <sup>st</sup> Century Competencies .....	17
Table 4. PISA Draft Collaborative Problem-Solving Framework .....	30
Table 5. Key Takeaways from an Investigation of Available Measures of 21 <sup>st</sup> Century Competencies .....	38

## Figures

Figure 1. Raven's Progressive Matrices .....	19
Figure 2. PARCC Prototype Item: Julia's Garden .....	20
Figure 3. Sample PARCC Spreadsheet Item .....	24
Figure 4. Sample ETS Formulating Hypotheses Item .....	26
Figure 5. Graduation Performance System (GPS) Framework .....	27
Figure 6. Alelo Oral Language Simulation .....	29
Figure 7. Alelo Oral Language Simulation Learning Objectives .....	30
Figure 8. MSA Situational Judgment Test .....	33
Figure 9. EcoMUVE Ecosystem Science Learning .....	46
Figure 10. SimScientists Science Simulation .....	52

## ACKNOWLEDGMENTS

The impetus for this paper came from Asia Society, whose Global Cities Education Network (GCEN) is helping international educators promote their students' global competence. Tony Jackson and other Asia Society staff helped us frame the paper, and representatives from the GCEN cities provided useful information about their assessment efforts. The William and Flora Hewlett Foundation provided resources to support the writing of the paper and its dissemination by Asia Society to members of the GCEN and other stakeholders. The foundation also provided access to Merrilea Mayo, an assessment expert who helped with background research on assessment quality and validity for predicting college and career outcomes. We want to thank the educators we interviewed about their experiences with innovative assessments, particularly Lisa Pullman at the Independent School Data Exchange (INDEX), who helped us arrange interviews with a number of teachers. We also benefited from the insights of many test developers, including those responsible for designing many of the assessments highlighted in this report. Two anonymous reviewers offered valuable suggestions on an earlier draft. Finally, we are grateful to James Torr for assistance with editing and Donna White for help with document preparation.

# INTRODUCTION

Public school systems are expected to promote a wide variety of skills and accomplishments in their students, including both academic achievement and the development of broader competencies, such as creativity, adaptability, and global awareness. The latter outcomes, which are often referred to as “21<sup>st</sup> century skills” or “21<sup>st</sup> century competencies,” have recently taken a more central role in policy discussions, because they are seen as critical components of college and career readiness. For example, in the United States, more than forty states have adopted the Common Core State Standards (CCSS), which are designed to embody a broader view of the knowledge and skills needed for success in college and careers. This growing emphasis on outcomes beyond simple academic content knowledge is the result of a confluence of factors, including perceptions among some business and government leaders that globalization, technology, migration, international competition, and changing markets require a greater emphasis on these outcomes than was required in the past. As a result, school systems are facing increasing pressure to produce graduates with this range of competencies (i.e., knowledge, skills, attitudes, and dispositions), a demand that generates challenges in terms of pedagogy and assessment.

In a previous Asia Society report, Saavedra and Opfer (2012) summarized lessons from research on learning to identify promising strategies for teaching 21<sup>st</sup> century competencies. For example, they stressed the importance of making curriculum relevant, helping students learn how to teach themselves, and fostering creativity. This report builds on that foundation by examining how to assess 21<sup>st</sup> century competencies.<sup>1</sup> Fortunately, data systems and measurement techniques that provide opportunities to assess students’ attainment of these outcomes are increasingly available. This report is intended to acquaint teachers, school leaders, and district administrators with the current state of 21<sup>st</sup> century competencies assessment, provide examples of relevant measures that educators in the field may wish to consider using, and offer some guidance to help educators compare measures and implement an assessment system.

Given these objectives, the report proceeds as follows. In Chapter 2, we describe a number of important 21<sup>st</sup> century competencies that will serve as our focus for the remainder of the report. Chapter 3 presents a set of criteria that educators should consider when deciding whether, when, and how to measure these competencies. These criteria guide the selection of assessment examples of 21<sup>st</sup> century competencies presented in Chapter 4. The inclusion of an assessment in this chapter is neither an endorsement of that assessment nor a confirmation that it achieves particular measurement goals; rather, we try to provide examples that are representative of the tests in the field, so that practitioners can make more informed assessment decisions. Finally, Chapter 5 draws on the previous information to offer a set of guidelines to help potential users of assessments make more informed choices.

## The Global Cities Educational Network (GCEN)

This report was prepared as a resource for schools and school systems, and the GCEN in particular. Globalization of the economy, increasingly diverse and interconnected populations, and rapid technological change are posing new and demanding challenges to individuals and societies alike. School systems are rethinking what knowledge and skills students will need for success and the educational strategies and systems required for all children to achieve them. In both Asia and North America, urban school systems are at the locus of change in policy and practice—at once the sites of the most critical challenges in education and the engines of innovation needed to address them.

---

<sup>1</sup>Throughout the report, we use the terms “assessment,” “examination,” “test,” and “measure” fairly interchangeably. For example, when we discuss assessments, we are also including the examinations on which education systems in much of Asia and Europe rely.

Asia Society organized the GCEN, a network of urban school systems in North America and Asia, to focus on challenges and opportunities for improvement common to them and to virtually all city education systems. A critical element of high-performing school systems is that they not only benchmark their practices against those of other countries, but they also systematically adapt and implement best practices within their own cultural and political contexts. The GCEN is intended as a mechanism for educators and decision makers in Asia and North America to collaboratively dream, design, and deliver internationally informed solutions to common challenges with which education systems are currently grappling.

## 2. EXAMPLES OF 21<sup>ST</sup> CENTURY COMPETENCIES

The term “21<sup>st</sup> century competencies” means different things to different people, and descriptions of these competencies rarely match one another exactly. Our purpose is not to provide a comprehensive catalog of 21<sup>st</sup> century competencies but to explore the challenge of assessing such outcomes, which are not widely measured and are not always amenable to the traditional assessment formats used for academic achievement tests. For this purpose, it is important that we identify a diverse set of 21<sup>st</sup> century competencies, not that we identify all such outcomes. It is also important that we select competencies that illustrate a broad range of assessment approaches.

To ensure that our analyses reflect the full range of competencies, we begin by describing broad categories of outcomes deemed to be vital in the 21<sup>st</sup> century, and then we identify a diverse set of exemplars within each category. This approach helps ensure wide coverage and reasonable balance among the differing skillsets subsumed under the general heading of 21<sup>st</sup> century competencies. This two-level structure mirrors the approach taken by a number of organizations in the United States, Asia, and internationally. At the top level, we loosely follow the categories used by the National Research Council, which also reflect the priorities articulated by GCEN members and various organizations with expertise in 21<sup>st</sup> century competencies. These broad categories are the following:

- Cognitive competencies
- Interpersonal competencies
- Intrapersonal competencies

There are a number of specific competencies within each category that educators and policy makers deem essential to success in the global economy. We selected a small number of these competencies as our focus (see Table 1). Our selection was based primarily on two factors: (a) there is research suggesting that educators can influence student mastery of these competencies, and (b) the competencies are defined well enough to permit a range of assessment options.

The cognitive category includes mastery of core academic content, including but not limited to mathematics, science, language arts, foreign languages, history, and geography. Cognitive competencies also include critical thinking and creativity, both of which were identified by GCEN members and global organizations as fundamental in a transnational economy. While these various competencies all relate to cognition, the name of this category does not imply that other competencies do not. In fact, all of the competencies in this report involve higher-level thinking of some kind. The interpersonal category covers the competencies that students need in order to relate to other people. These competencies begin with the basic capacity to communicate. Communication, in turn, sets the foundation for more multifaceted outcomes, such as collaboration and leadership. Given the GCEN’s focus on global competence, we also include global awareness among the interpersonal competencies; global awareness refers to a demonstrated empathy for the diverse circumstances that people experience across countries and an understanding of the interrelatedness of people, institutions, and systems. Beyond its altruistic value, global awareness has been shown to predict a heightened ability to navigate the complexities of international markets.

The intrapersonal category includes the competencies “that reside within the individual and aid him or her in problem solving” (Koenig 2011, 21). This category can also be conceptualized as the attitudes and behaviors that influence how students apply themselves in school, work, and other settings. Such attitudes



**TABLE 1**

<b>Examples of 21st Century Competencies by Category</b>
Cognitive Competencies Academic mastery Critical thinking Creativity
<b>Interpersonal Competencies</b> Communication and collaboration Leadership Global awareness
<b>Intrapersonal Competencies</b> Growth mindset Learning how to learn Intrinsic motivation Grit

---

include having a growth mindset, learning how to learn, being motivated to succeed, and showing grit in pursuing goals. Recently, the intrapersonal category has received considerable attention from educators and employers. This attention results from emerging research showing that intrapersonal competencies can, in some contexts, predict long-term academic and economic outcomes (Duckworth, Peterson, and Matthews 2007; Dweck 2008; Walton and Cohen 2011). These studies also find that educators can generate specific, cost-effective interventions to improve these competencies for students, especially those at risk of low achievement and attainment (Dweck 2008; Walton and Cohen 2011; Yeager and Walton 2011; Yeager, Walton, and Cohen 2013), though uncertainty remains about whether these studies generalize to a broad range of student populations.

## **A CLOSER LOOK AT SOME 21<sup>ST</sup> CENTURY COMPETENCIES**

This section provides greater detail on the specific competencies we sampled in each of the three broad categories. It is important to have a precise definition for a given competency when trying to measure it, so we present commonly used definitions of the competencies we've chosen to examine. We also provide some evidence about the importance of these competencies for students in the 21<sup>st</sup> century. Readers might find that some competencies are more relevant to their contexts than others, and they can focus on those throughout the document.

### **Cognitive Competencies**

Among the many cognitive competencies, we highlight three in this paper: academic mastery, critical thinking, and creativity. These three are grouped together because they incorporate either knowledge in core academic subjects or the skills that relate to how one processes this knowledge. While intrapersonal competencies also influence how one thinks about academic content, they are more attitudinal in nature than the competencies that relate directly to mastery of core content and are therefore treated separately.

### **Academic Mastery**

Learning academic content is fundamental to education, and mastery of such content serves as the basis for higher-order thinking skills as well as the impetus for improved interpersonal and intrapersonal competencies. Academic content includes instruction in subjects such as mathematics, science, reading, global studies, and foreign languages. We include global studies on this list because of its importance

to GCEN members and because the World Bank (among others) has found it to be important for economic growth in general, and for building effective partnerships between Asia and the United States in particular (Almeida 2009). Understanding regional norms and being able to communicate across cultures has gained increased prominence in recent years as a result of the increasingly global economy. Academic mastery is also important, because efforts to improve interpersonal and intrapersonal competencies like communication, academic mindset, and learning to learn are best accomplished within the context of academic instruction in a specific subject (Boix-Mansilla and Jackson 2011; Gardner and Boix-Mansilla 1994; Grotzer and Basca 2003; Mansilla and Gardner 1998; Saavedra and Opfer 2012). Although 21<sup>st</sup> century competencies such as communication appear to be important to a variety of later educational and career outcomes in their own right, they do not operate independent of academic outcomes.

### **Critical Thinking**

Critical thinking is highlighted in almost every discussion of key competencies for the 21<sup>st</sup> century. According to Facione and colleagues (1995), critical thinking includes inductive and deductive reasoning, as well as making correct analyses, inferences, and evaluations. These competencies are important for deeply understanding academic content, and they also relate to later career performance. Research suggests that for a company to compete in the global economy, it needs workers who will think about how to continuously improve its products, processes, or services. According to many executives, the heart of this continuous improvement process is knowing the right questions to ask (Wagner 2010), a function of critical thinking. Studies also tie critical thinking to other important societal outcomes. For example, Facione (1998) argues that citizens who think critically are likelier to be self-sufficient and therefore less of a drain on state resources. Meanwhile others suggest that such citizens are better equipped to give back to society, including through social entrepreneurship aimed at benefiting others (Peredo and McLean 2006).

### **Creativity**

Many educators and employers see creativity as a vital 21<sup>st</sup> century competency. While researchers have not settled on a single definition of creativity, Jackson and Messick (1965, 319) suggest that “unusualness, appropriateness, and transformation in varying combinations characterize a large proportion of those things we are willing to call creative.” Though creativity has been defined variously since Jackson and Messick, many prominent definitions do not stray too far from these authors’ conceptualization (El-Murad and West 2004; Parkhurst 1999; Runco 1996). Given its broad applicability and value to entrepreneurship, creativity is included among the key 21<sup>st</sup> century competencies by a range of organizations and scholars, including the Organisation for Economic Co-operation and Development (2013), the National Research Council (Pellegrino and Hilton 2013), the Hewlett Foundation (Conley 2011), ETS (Kyllonen 2008), and the World Bank (Di Gropello 2011). Innovation in particular has consistently been identified as a driving force in 21<sup>st</sup> century economic development (Archibugi and Lundvall 2002; Sawyer 2006; Wagner 2010). Partially as a result, creativity has gained increasing focus in educational programs globally. For example, China and Singapore have devoted resources to fostering more creativity in their schools. It is important to note that definitions of creativity can be culturally specific (Trompenaars and Hampden-Turner 1998). Though it is beyond the scope of this report to address nuances in how different cultures conceptualize creativity, schools and school systems should be alert to cultural specificities when defining creativity (as well as many other constructs considered in this report).

### **Interpersonal Competencies**

We highlight three interpersonal competencies: communication and collaboration, leadership, and global awareness. We combine communication and collaboration because the latter relies heavily on the former, and it is difficult to measure collaboration independent of communication. In fact, there is overlap among many of these competencies, because they are founded on interpersonal interaction; e.g., leadership involves both communication and collaboration. Many formulations of 21<sup>st</sup> century competencies include

tolerance and sensitivity; we subsume these characteristics under the heading of global awareness, which involves empathy for people in diverse circumstances, among other competencies.

### **Communication and Collaboration**

We consider communication and collaboration jointly for practical reasons, although each is itself a broad concept. For example, communication is sometimes broken down into three qualities: clarity, information shared, and balance among participants (Mohr, Fisher, and Nevin 1996). Similarly, collaboration can be thought of as communication plus additional competencies related to conflict resolution, decision making, problem solving, and negotiation (Lai 2011).

Communication and collaboration are identified as vital 21<sup>st</sup> century competencies by almost all of the organizations we consulted, by the GCEN members with whom we spoke, and by most researchers. For example, Pellegrino and Hilton (2013) suggest that communication is vital to facilitate teamwork and lies at the core of empathy, trust, conflict resolution, and negotiation. For instance, effectiveness with clients often hinges on effective communication and the teamwork necessary to produce a superior product. The importance of communication and collaboration in the workforce has generated increased focus on these skills in schools. As an example, the Houston Independent School District (a GCEN member) requires students to give group presentations in front of local executives to improve the students' ability to communicate and collaborate.

### **Leadership**

Leadership can be difficult to define because it includes aspects of communication and collaboration, along with a sense of vision for the future and competencies involving working with people. More broadly, leadership is not just a competency but a set of competencies. For example, a study conducted across Asian countries suggested that leadership involves having initiative, building consensus, innovating new strategies, and implementing policies and programs in collaboration with or under the direction of others (Berman et al. 2013). Moreover, because leadership involves working with and managing other people, including their competing priorities, collaboration is an important competency for a leader to possess. Research also suggests that the nature of leadership may be changing. Statistics show that an increasing number of college graduates will find employment in an organization they started themselves (Ball, Pollard, and Stanley 2010; Di Addario and Vuri 2010; Nabi, Holden, and Walmsley 2010; Tanveer et al. 2013). As start-up businesses play a larger role in the economy, it will be critically important for leaders of these businesses to have the ability to not only act on a vision but also to nimbly organize others around that vision ( Ball, Pollard, and Stanley 2010).

### **Global Awareness**

Global awareness has grown in importance in the 21<sup>st</sup> century as economic, social, and cultural connections among countries have increased. Though global awareness involves a major academic, cognitive component, we include it in the interpersonal competencies section because it also requires a bevy of complex interpersonal skills. Perhaps the best-studied interpersonal competency that underlies global awareness is empathy. For instance, a student might demonstrate global awareness if he or she feels empathy for people in different cultural or geopolitical circumstances (Bachen, Hernández-Ramos, and Raphael 2012). However, this specific form of empathy is only one facet of the interpersonal nature of global awareness. To be globally aware, a person must also show an understanding of the interrelatedness of people, institutions, and systems. Being able to connect how actions in one area of the world affect other areas, or how local events influence and are influenced by global events, is a core part of global awareness. These interpersonal aspects of global awareness have been tied to success in the international economy. For example, Boix-Mansilla and Jackson (2011) suggest that students must know how to investigate the world, weigh perspectives, communicate ideas, take action, and apply expertise in order to prosper in a global, multicultural workforce.

## Intrapersonal Competencies

Intrapersonal competencies can be conceptualized as the attitudes and behaviors that influence how students apply themselves in school, work, and a range of other settings. For example, being motivated leads to better performance in a number of contexts and on a variety of tasks. Increasingly, educators believe that improving intrapersonal competencies is key to maximizing student potential. Research shows, for instance, that at-risk students achieve at higher levels when they possess a growth mindset (Dweck 2006; Walton and Cohen 2011; Yeager, Walton, and Cohen 2013). We sample four competencies from this category: growth mindset, learning how to learn, intrinsic motivation, and grit.

### **Growth Mindset**

Students with a growth mindset see intelligence as malleable and as a function of effort, whereas those with a fixed mindset treat intelligence as an innate ability, immune to the efforts of the individual to improve it (Dweck 2006; Dweck 2008; Dweck 2009; Dweck 2010). Over more than a decade, Carol Dweck has chronicled the advantages of having a growth mindset rather than a fixed mindset when it comes to learning core academic content in subjects such as mathematics and science. For example, she shows that students often perform better in mathematics when they possess a growth mindset, because they are more willing to engage with difficult material and overcome setbacks (Dweck 2008). More recently, Dweck and others have applied her mindset framework to 21<sup>st</sup> century competencies frameworks, suggesting that students are in a much better position to develop 21<sup>st</sup> century competencies if they exhibit the right mindset (Dweck 2009; Stevenson et al. 1990; Stevenson, Lee, and Stigler 1986; Stigler and Stevenson 1991).

### **Learning How to Learn**

Learning how to learn, or “metacognition,” refers to a student’s ability to determine how to approach a problem or task, monitor his or her own comprehension, and evaluate progress toward completion (Landine and Stewart 1998; Vrugt and Oort 2008). Much research documents the importance of metacognition to academic achievement (Landine and Stewart 1998; Paris and Winograd 1990; Schunk and Zimmerman 2012; Vrugt and Oort 2008; Zimmerman 1990; Zimmerman 2001). David Conley in particular ties metacognition directly to college readiness and uses the term to encompass many of the other competencies discussed in this report (Conley 2008). We interpret learning how to learn broadly to include related competencies, such as self-regulation, which has been shown to predict achievement, attainment, and workforce success. For example, Vrugt and Oort (2008) show that self-regulated learners have significantly better achievement test scores than their peers. Research shows that metacognitive competencies also influence how students respond to classroom instruction. A student who understands his or her own learning processes is better able to self-motivate, respond to teacher feedback, and develop stronger self-perceptions of academic accomplishment (Zimmerman 1990).

### **Intrinsic Motivation**

Motivation refers to the process that prompts people to take action to attain particular goals, and psychologists generally distinguish between two types of motivation. Extrinsic motivation refers to pressures to act that derive from sources outside the individual; these include incentives such as money or, in the case of students, praise or grades. Intrinsic motivation refers to forces within the individual that activate behavior. Students might demonstrate intrinsic motivation if they spent extra time learning about a topic in science because they are interested in it, whether or not it was a required assignment. Research shows that without motivation of some type, students are unlikely to master core academic content (Barca-Lozano et al. 2012; Deci et al. 1991; Goslin 2003; Guay et al. 2010). In fact, there is evidence (including research in schools in Asia) that both types of motivation matter to educational achievement. Since extrinsic motivation is largely a feature of setting, we focus our examination on assessing intrinsic motivation, which is an attribute of the learner. Interestingly, intrinsic motivation has been shown to be

an important element in complex problem solving and achievement (Deci and Ryan, 2012; Deci et al., 1991; Ryan and Deci, 2000a, 2000b; Stone, Deci, and Ryan, 2009).<sup>2</sup> Although in many cases extrinsic incentives can result in students working harder than they would in the absence of those incentives, there is some evidence that for certain (though not all) complex tasks, the use of extrinsic incentives can reduce performance and long-term interest in the task (Deci et al. 1991; Ryan and Deci 2000a). Intrinsic motivation has been found to be especially important in higher education, where enrollment is more at the discretion of the student (Lin, McKeachie, and Kim 2001; Ryan and Deci 2000b).

### **Grit**

Grit is an emerging construct that relates directly to intrinsic motivation and has been shown to predict important outcomes, including achievement (Duckworth et al. 2007; Duckworth and Quinn 2009). Whereas motivation refers to an immediate interest or desire to engage in an activity, grit refers to perseverance and passion for long-term goals (Duckworth et al. 2007). Thus a person could be motivated but still lack grit if he or she focused attentively on everyday work but lost interest in a project over time and failed to complete it. Research shows that students reach higher levels of educational attainment when they demonstrate grit (Duckworth et al. 2007). Moreover, teachers that evince more grit tend to stay in the profession longer and generate larger test-score gains for their students in mathematics and language arts (Duckworth, Quinn, and Seligman 2009). Although there is a positive correlation between measures of grit and later outcomes of interest, there are still unanswered questions about the role of grit in student success and how grit relates to other constructs. For example, there seems to be overlap among discussions of motivation, grit, and resilience. While it is important to clarify the distinctions among these concepts before measuring them, sorting them out is beyond the scope of this report.

### **Conclusion**

This report attempts to focus on a subset of 21<sup>st</sup> century competencies that are measurable, supported by quality research, and subject to improvement through teacher actions. However, 21<sup>st</sup> century competencies are an emerging area of research, and we do not always have a clear understanding of the processes through which these competencies develop. While there is extensive research on how students progress from one skill to the next in mathematics and writing, there is very little research on the stages of development for many of the competencies described in this chapter. For example, researchers cannot yet describe the “learning progressions” students follow to go from novice to accomplished in terms of collaboration or grit. The absence of learning progressions for many 21<sup>st</sup> century competencies also complicates efforts to measure performance and growth on these competencies, an issue discussed later in the report.

---

<sup>2</sup>Though people can be intrinsically motivated to perform all sorts of tasks, in this paper, we are generally referring to intrinsic motivation for academic achievement and attainment, as well as any related skills and attitudes that inform those academic outcomes.

### 3. A FRAMEWORK FOR SELECTING MEASURES OF 21<sup>ST</sup> CENTURY COMPETENCIES

Educators trying to decide whether, when, and how to measure 21<sup>st</sup> century competencies are confronted with a dizzying array of options. Few resources exist to help practitioners and policy makers balance the competing factors involved, such as cost and instructional value. In this chapter, we try to address this gap by presenting a framework educators can use to evaluate measures of 21<sup>st</sup> century competencies. As shown in Table 2, these considerations tend to fall into three categories: instructional, practical, and technical.

**TABLE 2**  
Considerations When Selecting Measures of 21<sup>st</sup> Century Competencies

<b>Instructional</b>
Formative or summative
Actionable information to teachers
Useful feedback to students
Grade/context appropriate
Engaging, meaningful, and authentic for students
Encourages effective teaching and learning
<b>Practical</b>
Cost
Ease of training
Ease of scoring
Ease of administration
Ease of technological implementation
<b>Technical</b>
Reliability
Validity
Fairness

#### Purpose

When determining what test to use, potential users of assessments must consider the purposes of the assessment (Haertel 1999; Messick 1994; Kane 2012). In general, there are four broad purposes for which assessments might be used: (1) monitoring system performance, (2) holding schools or individuals accountable for student learning, (3) setting priorities by signaling to teachers and parents which competencies are valued, and (4) supporting instructional improvement (Schwartz et al. 2011). Though we consider all four of these potential uses in the remainder of the report, we largely focus on the fourth one—providing information that can be used to improve instruction. Such information might relate to 21<sup>st</sup> century cognitive, interpersonal, or intrapersonal competencies.

#### Instructional Considerations

In this section, we describe the instruction-related criteria that schools should weigh when choosing an

assessment. In essence, schools must determine what instructional value an assessment will provide as a basis for deciding whether the value justifies its cost.

### Formative versus Summative

One important consideration is whether the measure is to be used for formative or summative purposes; i.e., to inform ongoing instructional decisions or to determine whether instruction has been effective after the fact. For example, the use of frequent formative assessment in elementary reading might help a teacher determine whether a student is struggling with vocabulary, phonemic awareness, phonics, or comprehension. In fact, many scholars argue that formative assessment is a process rather than a test (Heritage 2010) and that effective formative assessment practice involves teachers setting goals, engaging in frequent feedback cycles with students, adjusting instructional practices in response to assessment data, and engaging students in the assessment process by providing individualized instruction and opportunities for self-assessment (Black et al. 2003; Heritage 2010; Herman, Osmundson, and Silver 2010).

In contrast, other tests serve a very different purpose; namely, evaluating teaching and learning after it has occurred. For example, China's summative end-of-high-school test, the Gao Kao, is designed to provide a summary of a given student's learning up through secondary school. The test is not meant to adjust instruction so much as determine whether the combination of a student's initial skills with that instruction has generated the type of outcomes valued by the country's higher education system. Similarly, Advanced Placement examinations at the end of an accelerated course provide information about whether the student has mastered the content of the course.

The distinction between formative and summative purposes is not always clear-cut. Some schools administer interim (or benchmark) assessments that are designed to predict whether students are likely to perform well on a test used for accountability. Although these interim assessments might be used to adjust instruction, they typically mirror the form of the accountability tests and do not necessarily provide information that supports day-to-day instructional decision making (Perie, Marion, and Gong 2009).

### Usefulness of the Data for Students and Teachers

It is important to consider whether the information produced by an assessment will be useful for instructional improvement. Will the reports provide teachers with any new information that help identify student learning challenges or change instructional activities? As an example, potential users might question the value of a measure to diagnose areas of academic difficulty if the results do not tell the teacher anything new, or if they prove discouraging for the student without simultaneously offering useful information to help students improve. While these considerations largely apply to formative measures, they can also relate to summative assessments. A district would not, for instance, want to use a measure that is hard to interpret and provides little information on how educators can improve performance. In many ways, this criterion relates to validity (which is discussed below)—test users need to think carefully about an assessment's intended uses, then determine whether the information provided is actionable and relevant, given that purpose.

Assessment results can also produce benefits related to communication and organization. For example, even if teachers have strong intuition or anecdotal evidence about a student's 21<sup>st</sup> century competencies, having concrete data can increase how intentional teachers are about fostering these competencies, in part by developing a common vocabulary among teachers, parents, and students. Regular data might also provide organizational benefits, including providing information at consistent intervals that help ensure teachers meet regularly to discuss student needs.

## Influence of Testing on Teaching and Learning

Educators should also attend to the way that an assessment may influence classroom practice; researchers have found that assessment can have both positive and negative consequences for schools and classrooms. On the positive side, research shows that the implementation of new assessment systems in K–12 schools can lead to beneficial changes (Hamilton, Stecher, and Yuan 2012). Similarly there is some evidence that the implementation of standards-aligned accountability tests has been accompanied by efforts to improve the quality of curriculum and instruction, particularly when the assessment includes open-ended items that measure complex problem solving (Center on Education Policy 2006; Hamilton et al. 2007; Lane, Parke, and Stone 2002; Stecher 2002). Teachers and administrators have reported responding to assessments by taking steps to improve school performance, such as adopting new programs to address the needs of low-performing students, increasing the use of data to improve decision making, providing professional development and other supports to promote improved teaching, and increasing time spent on instruction (see Hamilton 2003, Stecher 2002, and Faxon-Mills et al. 2013 for more comprehensive reviews of this research).

Moreover, reports from Queensland suggest that their portfolio-based system (described in the appendix), which is scored and largely run by teachers, can provide useful information about student learning and can develop a sense of professionalism among educators. Other studies support this theory as well. For example, Stecher (1998) shows that assessments requiring more teacher involvement can positively influence attitudes and beliefs about teaching, inform curriculum and instruction in meaningful ways, and change how teachers go about measuring skills in their own classrooms (though results are mixed, in particular on this latter point). Positive consequences were also reported for the implementation of the Mission Skills Assessment of 21<sup>st</sup> century competencies (see Chapter 4).

However, there is substantial evidence of responses to assessments that would generally be considered undesirable, and these detrimental effects are often unanticipated (as well as unintended). We mention two well-documented examples of this phenomenon. First, researchers and practitioners have raised concerns that the use of high-stakes, multiple-choice tests has led to a narrowing of the curriculum and reduced emphasis on nontested skills, including many of the skills highlighted in this report (Longo 2010; McNeil 2000; Hamilton et al. 2007; Hamilton 2003; Koretz et al. 1991; Koretz 2008; Smith 1991). Second, performance levels that often accompany high-stakes tests can negatively label students and influence attitudes of students and teachers alike (Jussim, Eccles, and Madon 1996; Papay, Murnane, and Willett 2010).

We would argue that effect on instruction is one of the most important criteria that educators can consider, and it might be worthwhile to sacrifice some other features of assessment (e.g., cost and practicality) for increased positive instructional effects. For instance, a district might be willing to pay more for an assessment and even accept reduced reliability (discussed below) if the activity generates student engagement and a deeper mastery of the content. In particular, performance tests, which tend to measure tasks that more closely resemble the real-world actions they are meant to assess, often prove less reliable and more expensive than multiple-choice and other types of tests with structured responses (Gao, Shavelson, and Baxter 1994; Shavelson, Baxter, and Gao 1993). If the primary purpose of the assessment is to model good instruction, the use of performance assessments might be appropriate even if they have lower technical quality, but their use might be problematic if the scores are used to make high-stakes decisions about students or educators. We discuss more of these complex tradeoffs in Chapters 4 and 5.



## Practical Considerations

The documentation that accompanies a test may not fully discuss the practical issues that are important to teachers and schools that will be administering the assessments. We discuss two practical considerations: cost and logistical requirements.

### Cost

For school systems with limited budgets, cost is an important factor in deciding whether and when to use an assessment. Assessment costs are often driven by the complexity of the test format, which means that assessments of some 21<sup>st</sup> century competencies may be more expensive than more traditional assessments. Although some measures require only a brief paper-and-pencil survey, others involve complex computer simulations and rating schemes, with multiple observations of a student. As a result, the cost of purchasing and using different measures can vary substantially. To complicate matters further, the complexity of the test format often mirrors the complexity of the competency being measured, which means some of the highly valued competencies, such as creativity, are frequently the most expensive to assess. At the same time, technology has made it possible to reduce some costs associated with complex assessment. For instance, electronic scoring algorithms can replace human raters in some situations, and many of the computer-based simulations are much less costly than a similar, hands-on activity. We discuss these tradeoffs more in Chapter 4.

### Logistical Requirements

Clearly, cost does not come only in the form of a price tag on an assessment. Staff time in particular represents a cost, in terms of both dollars and time taken from direct instruction or other activities. More complex measures often require time to teach educators how to administer, score, interpret, and use them. For example, for test responses that are scored by teachers, test developers frequently try to promote high levels of rater agreement by providing detailed instructions and rubrics that help teachers score the test in a consistent manner. While this approach tends to help increase reliability, it typically requires teachers to reallocate time from instruction and other activities in order to participate in the necessary training. At the same time, reports from countries using this approach suggest that teacher involvement can be quite valuable as a professional development tool and can inform instruction. Given these tradeoffs, educators wishing to use complex assessments need to think carefully about whether the investment of time and resources will be worth the potential benefits.

Technological requirements are also an important consideration. Schools must be sure they have the technological infrastructure to administer and score the tests and to make sense of the data they produce. In particular, schools must ensure that the computers they have are powerful enough to run simulations, and that there are enough consoles to allow a reasonable number of students to complete the assessment. Schools also have to provide support to teachers who may be less familiar with the technology, as well as when glitches inevitably arise. Finally, practitioners will usually need to be cognizant of how facile students are with technology and, if relevant, whether students can access the necessary computing resources at home. These demands will only increase as the sophistication of the technology increases.

## Technical Considerations

In addition to instructional and practical considerations, educators must pay attention to the overall technical quality of the measure. Technical quality refers to factors such as whether the assessment measures what its developers claim it measures, and whether it provides consistent and meaningful results across students, tested tasks, and versions. While technical criteria can be hard to explain because of their statistical nature, it is important to consider these issues when examining assessment options.

If a test's technical quality is low, then it will not provide meaningful information for any potential use. In this section, we attempt to identify the most important considerations and explain them in relatively nontechnical terms; this should help educators navigate some assessment-related measurement jargon. By and large, any assessment used, even if for relatively low-stakes purposes, must at least meet some minimum standard for each of the below criteria.

### **Reliability**

Reliability has both technical and conversational meanings, though the two are not unrelated. Put simply, scores on a test are considered reliable if a student taking the test would get essentially the same score if he or she took it again under similar circumstances (and assuming no learning occurred as a result of the first administration). At heart, reliability is about consistency. Inconsistency results from the effects of measurement error on scores, and different sources of error can contribute to this lack of consistency.

A variety of methods can be used to estimate score reliability, and each of them addresses different sources of error. For example, test-retest reliability coefficients are used to estimate the consistency of scores across multiple occasions of testing. Sometimes it is impossible to administer the test more than once—for logistical reasons, or because it would result in an artificial increase in scores as a result of student exposure to the test content. An alternative approach to estimating this source of error is to divide the test into two parallel halves and estimate the correlation between scores on these half-tests. Other methods build on this approach, such as by estimating the average correlation across all possible pairs of half-tests. This method is referred to as Cronbach's Alpha (which we reference in Chapter 4 and the appendix 1), and it measures the internal consistency of the test—that is, the extent to which scores on the items cluster together. For tests that rely on human raters, the effects of raters are another potential source of error. In these cases, measures of rater agreement are often used to estimate reliability. Similarly, for tests that ask students to perform a range of tasks, such as scientific experiments, test developers also consider how much error is introduced by the tasks themselves. For example, if two tasks that are intended to measure the same construct produce different results, then tasks could be a major source of error (Shavelson, Baxter, and Gao 1993).

The primary consideration for test users who are interested in understanding score reliability on a particular test is the extent to which there is evidence of measurement consistency or precision that takes into account all of the relevant sources of error. Though these scenarios deal with different types of consistency, they are often reported on a similar scale that ranges from 0 to 1, with 1 being perfect reliability (i.e., no measurement error). While standards differ, a reliability of 0.7 or higher is often considered acceptable for standardized tests, though the adequacy of score-reliability evidence needs to be evaluated based on the specific uses proposed for the test. Understanding this reliability scale—and the technical details of reliability more generally—is not important for most practitioners. What is important, however, is the understanding that a test with low levels of reliability will not provide useful information about students. If a score is determined more by chance than by the student's skills in the tested area, the score will not be useful for decision making.

### **Validity**

Validity is the most important consideration when evaluating the technical quality of a test. The term refers to the extent to which there is evidence to support specific interpretations of test scores for specific uses or purposes. For example, a test claiming to measure student ability to conduct arithmetic operations with fractions may produce consistent scores but would not be considered valid if it tested only addition and subtraction of fractions but not multiplication and division. While this example is clear-cut, others are not. For instance, how would one show definitively that all topics relevant to understanding algebra have been covered on a given test? Or that performance on the test is not unduly

influenced by mastery of content other than algebra (e.g., reading ability)? For some purposes, one might consider reading ability to be a legitimate focus of measurement for an algebra assessment, whereas for other purposes, reading ability might be considered irrelevant. Because such issues are often murky, validity is never proven: the best we can hope for is a convincing validity argument accompanied by evidence that supports that argument (American Educational Research Association et al. 1999; Kane 2001; Kane 2006; Kane 2013).

A convincing validity argument generally involves synthesizing evidence from a variety of sources. Examples of the types of evidence that can support a validity argument include evidence based on test content (e.g., expert evaluations of the extent to which test items are representative of the domain that the test is designed to measure), evidence based on response processes (e.g., interviews with examinees as they “think aloud” while taking the test in order to determine whether the test elicits the intended responses), and evidence based on relationships with other measures or other information about examinees collected either at the same time or in the future (e.g., the extent to which scores on a reading test correlate with scores on a different reading test, or the extent to which they predict later performance in postsecondary education) (American Educational Research Association et al. 1999). Examining multiple sources of evidence can help test users understand the extent to which the test measures what they think it measures and whether it is an appropriate tool for the particular decision they are interested in making. This type of validity investigation can also help users identify sources of “construct-irrelevant variance”—that is, instances in which scores are influenced by a skill or attribute other than the one(s) the test is intended to measure. If, for instance, scores on a mathematics test correlate more highly with scores on a reading test than with scores on other mathematics tests, this finding would raise concerns that students’ scores are unduly influenced by their reading ability. As another example relevant to the assessments discussed in this report, there may be construct-irrelevant variance if a student’s score on a computer-based assessment is determined partially by his or her facility with the technology (unless understanding of that technology is meant to be part of the construct). Evidence regarding construct-irrelevant variance relates directly to issues of fairness, which we discuss next.

### **Fairness**

Fairness is perhaps the easiest concept to understand because it extends well beyond testing. It also relates directly to validity: a test should measure the same construct for all examinees, regardless of whether they are members of particular groups (e.g., racial/ethnic or gender groups), and should support valid interpretations of examinee performance for the intended purposes of the test. Issues of fairness arise when a test wrongly characterizes the performance of a given student subgroup in some systematic way. For example, much research shows that standardized tests of academic content can be biased against students who do not speak the native language, because getting the right answer is determined more by language status than understanding of the tested subject (Abedi 2002; Abedi 2006a; Abedi 2006b; Haladyna and Downing 2004). Implicit in this example is an important distinction: just because a test is harder for one group than another does not make it unfair. Rather, bias (unfairness) arises when students with the same ability in the subject from two different groups perform differently. As a clarifying example, bias would not be present if poor students receive lower scores than their peers due to lack of sufficient instruction or low levels of family resources to support education (these are certainly major problems, just not ones of test bias), but it would be a fairness issue if poor students receive lower scores because they are less familiar with the language, scenarios, or logic of the test than their peers, despite having equal knowledge of the tested subject.

Industry standards suggest that tests and test administration conditions should be designed with fairness in mind by minimizing the number of group-specific modifications or accommodations required (a concept that is often referred to as “universal design”). For example, if students with certain disabilities often need more time to read a question and communicate an answer, and if the ability to answer the

question quickly is not central to the construct the test is designed to measure, then it might make more sense to allow all students to have more time to respond rather than setting a shorter time limit for most students and creating an exception for those with those specific disabilities. One benefit of technology is that it often makes reducing the number of accommodations or modifications easier by allowing more fluidity in test administration and format. For example, the font size on a computer-based test can be easily changed to accommodate the needs of visually impaired students, a process that is likely to be easier and less expensive than creating large-print paper copies of a test.

## Conclusion

In this chapter, we reviewed three broad criteria that practitioners may wish to consider when selecting an assessment: instructional, practical, and technical. This information is meant to help educators boil down the numerous considerations that go into selecting a measure into a few key themes. Laying out these criteria also sets the stage for Chapter 4, which provides examples of assessments that readers may wish to consider using. We have intentionally selected a broad array of 21<sup>st</sup> century competencies and assessment formats to help educators see how these criteria play out in practice.

## 4. EXAMPLES OF 21<sup>ST</sup> CENTURY ASSESSMENTS, BY FORMAT

In this chapter, we use examples to bring to life the range of considerations that arise when selecting an assessment. We help capture this range by presenting examples that span from established to cutting-edge measures, distinctions that are often a function of the assessment's format. The format types we review include multiple choice, self-report, open response, portfolio, performance, and cross-cutting (definitions for each will be provided as we go). Both across and within assessment formats, we begin with established assessments that measure a single construct, have a consistent track record of use, can be administered fairly easily, and oftentimes have a strong research base. From there, we progress to cutting-edge assessments that have not been administered on a large scale and therefore typically lack solid evidence of technical quality. These assessments often measure multiple constructs at once using multiple methods and involve the use of sophisticated technologies for administering, scoring, or both. We conclude with a rich description of a measure—the Mission Skills Assessment—that illustrates features of the most cutting-edge assessments. Altogether, deciding which measure to use—especially which format to use—involves difficult tradeoffs that practitioners should keep in mind as they read this chapter.

Table 3 provides a sample of the sorts of assessments we will discuss in this chapter. In addition to presenting the organization of the chapter visually, the table also helps make the characteristics of established versus cutting-edge assessments concrete. For example, the table shows that even a single cutting-edge measure might assess multiple competencies in different domains. All in all, this chapter is intended to be useful for schools just starting in assessing 21<sup>st</sup> century competencies through those ready-to-employ assessments relying on emergent technologies and formats. The format-based structure and content of this chapter builds on work done by Patrick Kyllonen (2012) at Educational Testing Service (ETS), documenting assessment formats that can be used to measure 21<sup>st</sup> century competencies in ways that are aligned with the Common Core State Standards.

As an additional caveat, inclusion of an assessment in this section is neither an endorsement nor confirmation that it is reliable and valid. We make this caveat for two main reasons. First, many cutting-edge assessments are still prototypes, so little research exists to establish their validity for certain purposes. While this does not mean a test provides little value or should be avoided, consumers should be aware that cutting-edge measures often have less evidence showing that they (1) cover the appropriate content, (2) produce scores that correlate with outcomes of interest, and (3) assess the construct of interest. Second, though technology is helping make test scores more reliable, there are some problems that innovation cannot fix, especially those related to defining the construct. That is, a test may be quite reliable in measuring some aspect of a given 21<sup>st</sup> century competency, but that is no guarantee that all aspects of the relevant competency have been measured (for example, a test may measure some facet of leadership consistently but fail to capture other essential aspects of that competency). Though we pay close attention to this aspect of validity, passing judgment on whether a test adequately captures the notion of, say, communication, is beyond the scope of this report.

**TABLE 3**  
Examples of Measures of 21<sup>st</sup> Century Competencies

Measure	Format	Competency			Purpose	
		Cognitive	Interpersonal	Intrapersonal	Formative	Summative
<b>Established Measures</b>						
Advanced Placement	multiple choice	language				x
Formulating Hypotheses	multiple choice, open response	creativity			x	
Watson-Glaser	multiple choice	critical thinking			x	
Global Empathy Scale	self-report		global awareness		x	
Theory of Mind	self-report			mindset	x	
College and Career Ready School Diagnostic	self-report			mindset, learning how to learn, intrinsic motivation, grit	x	
Work Extrinsic Intrinsic Motivation Scale	self-report			intrinsic motivation	x	
Grit Scale	self-report			grit	x	
<b>Cutting-Edge Measures</b>						
PARCC and Smarter Balanced*	multiple choice, open response, performance	math, reading, critical thinking	communication		x	x
Singapore Elementary Portfolios	portfolio	math, science, reading, critical thinking	communication			x
World Savvy Challenge	performance	global awareness, critical thinking	global awareness		x	x
PISA*	performance	math, science, reading, critical thinking	communication, collaboration			x
Graduation Performance System	portfolio	math, reading, critical thinking	communication, global awareness	learning how to learn	x	x
Alejo language and culture simulator*	performance	language, critical thinking	communication, global awareness	learning how to learn	x	
SimScientists*	performance	science, critical thinking	collaboration	learning how to learn	x	
EcoMUVE*	performance	science, critical thinking	collaboration, communication	learning how to learn, intrinsic motivation	x	x
Mission Skills Assessment	cross-cutting	creativity	collaboration	resilience, intrinsic motivation, learning how to learn	x	x
Queensland Performance Assessment	cross-cutting	math, science, reading, critical thinking	communication	learning how to learn	x	x

## Multiple Choice

Most educators today are familiar with standardized multiple-choice tests. Though popular for decades, the stakes attached to multiple-choice tests have grown in the wake of accountability movements across the globe (Carnoy, Elmore, and Siskin 2013; Kamens and McNeely 2010; Kell 2010; Koretz et al. 1991; Nichols and Berliner 2007; Wilson 2007). In the United States, the No Child Left Behind (NCLB) Act of 2001 required states to provide reliable and valid measures of reading and mathematics in grades 3 through 8. Moreover, the law required assessments specific to certain student subgroups, such as English learners. Examples of standardized multiple-choice tests can be found throughout the world. China has been administering high-stakes exams for centuries, including the Gao Kao, a test designed to determine college eligibility. The Australian Council for Education Research has developed a wide range of tests that involve multiple-choice items, including benchmark tests for schools and admissions assessments for colleges. Internationally, though not used explicitly to hold educators accountable, the Programme for International Student Assessment (PISA) is a test administered by the Organisation for Economic Co-operation and Development (OECD) that relies in part on multiple-choice items to compare achievement in reading, mathematics, and science literacy across countries.

Beyond large-scale accountability measures, multiple-choice tests are currently being used for a variety of purposes, including assessing some 21<sup>st</sup> century competencies. Measuring the acquisition of a foreign language is a prime example. Both Advanced Placement (AP) and Test of English as a Foreign Language (TOEFL) assessments rely in part on multiple-choice responses to measure proficiency in a foreign language, including English for non-native speakers. Like many tests of language proficiency, the TOEFL includes sections on reading, writing, speaking, and listening. Almost all of these sections rely on multiple-choice items, though in different ways. The listening portion, for instance, involves hearing two native English speakers engage in conversation, then responding to multiple-choice questions about the content of that discussion.

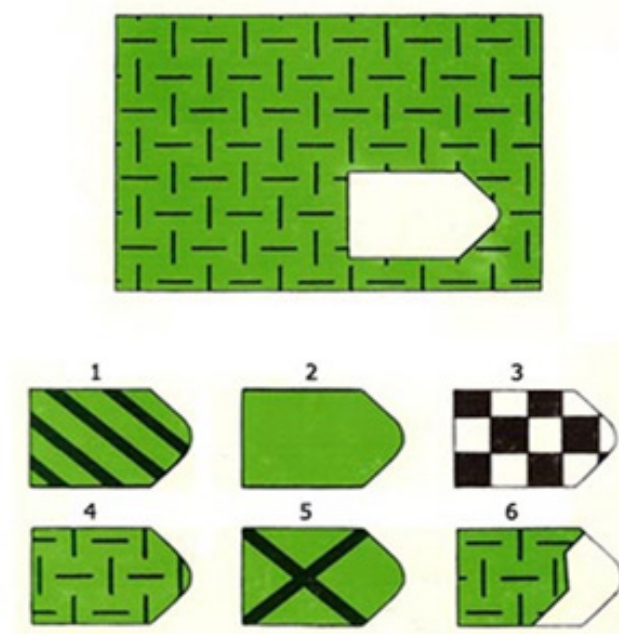
Scores from tests such as the TOEFL and AP are quite reliable, and both are supported by validity evidence for the purposes for which they are typically used. At the same time, these tests are quite different from the cutting-edge assessments of foreign language acquisition we present in this chapter's section on performance assessments, which have less evidence regarding technical quality but attempt to better replicate the actual experience of conversing in a new tongue.

Some assessments also use multiple-choice items to measure critical thinking. For instance, Raven's Progressive Matrices shows pictures of patterns with a five-sided piece missing (see Figure 1). Test takers are then given options for the piece that best completes the picture. Unlike most traditional IQ tests, Raven's Progressive Matrices is nonverbal, which means it can be used for students of different ages and from varying language backgrounds. Perhaps unsurprisingly, newer multiple-choice tests of critical thinking tend to be computer-based. For example, the California Critical Thinking Skills Test (CCTST) has been translated into a variety of languages and uses computer-based administration<sup>3</sup>. Although most of the questions are multiple choice, they ask students to examine reading passages, charts, pictures, paintings, and the like and then draw inferences from them. For each student, the assessment returns scale scores for analysis, evaluation, inference, deduction, induction, and overall reasoning.

---

<sup>3</sup><http://www.insightassessment.com/About-Us/California-Critical-Thinking-Skills-Test-Family>

**FIGURE 1.**  
Raven's Progressive Matrices



Retrieved from <http://www.raventest.net/raven-test.html>.

As noted by Kyllonen (2012), multiple-choice measures called “situational judgment tests” (SJTs) have been used increasingly, especially to measure intrapersonal competencies. SJTs present students with a scenario meant to test their mindset, motivation, or the like, then ask them to respond to that situation. For example, the Mission Skills Assessment (MSA), which we discuss in detail later in the chapter, tests students’ collaboration skills by asking them what they would do when confronted with an important deadline for a group project and a group member who is not willing to contribute. Such a scenario uses the multiple-choice format while still attempting to place students in a more real-world environment, especially one in which the right answer is not always obvious.

Looking ahead, several assessment organizations are attempting to make multiple-choice assessments more dynamic and less rote (see Figure 2). Paramount among these groups are the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (Smarter Balanced), both of which are developing large-scale tests aligned with the Common Core State Standards. In particular, these assessments will rely to a large extent on multiple-choice items, but the questions are meant to better emphasize critical thinking. For example, students may be asked to support an argument about a reading passage by selecting two quotes that best match their contention from a list. Much greater detail on PARCC and Smarter Balanced efforts—including an update on their progress and timeline for implementation—is included in the box below.




**FIGURE 2.**  
PARCC Prototype Item: Julia's Garden


Julia is planting flowers. She wants to cover  $\frac{3}{4}$  of the garden with flowers.

Drag a tile onto Julia's garden that will finish covering  $\frac{3}{4}$  of her garden with flowers.

**Possible tiles:**



**Julia's garden**



Submit Answer

Retrieved from [http://www.ccsstoolbox.com/parcc/PARCCPrototype\\_main.html](http://www.ccsstoolbox.com/parcc/PARCCPrototype_main.html)

### Common Core State Standards Consortia Assessments

As many states begin to implement the Common Core State Standards (CCSS), the US Department of Education has given grants to two state consortia to develop assessments aligned to the new standards. These tests are intended to replace the ones currently being used to meet federal accountability requirements for participating states. Though the two funded consortia—Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (Smarter Balanced)—differ in important ways, they share a number of fundamental priorities. Perhaps the most important of these priorities is an emphasis on increasing the representation of 21<sup>st</sup> century competencies in statewide assessments. Both assessments, which are meant to provide comparability across participating states, are scheduled to be in full use by 2015.

#### New Tests, New Goals

Beyond assessing mathematics and language arts, the new tests are being designed to measure 21<sup>st</sup> century competencies, with particular emphasis on critical thinking and communication. To that end, both consortia aim to do the following:

- Balance multiple-choice items with essay questions and performance assessments (for PARCC, this will include an end-of-the-year test of speaking and listening)
- Diversify the types of assessments by including formative and summative measures
- Use formative tasks to assess skills that are harder to measure, including planning, management of information, and critical thinking

### **Balancing Multiple Choice with Other Item Formats**

Both the summative and formative assessments will use innovative formats, item types, and scoring routines to help achieve the consortia's objectives. Though the summative assessments will include many standard multiple-choice questions, they will also include items that give students more flexibility to develop their own responses. For example, mathematics questions might allow students to graph functions, write equations, identify pieces of evidence to support a conclusion, or compose short answers.

### **From Paper to Computer**

Beyond specific items, both consortia will administer summative assessments on the computer, one of which (Smarter Balanced) will adapt questions to the student's level of understanding as the test progresses. One benefit to the adaptive computer-based approach is that the tests will likely do a better job of measuring performance for exceptionally high- or low-achieving students by ensuring that the questions a student sees are commensurate with his or her understanding of the material. More generally, computer-based testing can lead to faster scoring; in some cases, test results will be available a week or two after administration. The formative (interim) assessments, meanwhile, will often use even more innovative approaches. To measure communication and critical thinking, these measures will include tasks in the form of an oral presentation, essay, product development, or the like.

The consortia are also paying close attention to matters of validity. For example, Smarter Balanced is conducting validity studies of how well certain items and scores predict success in college or the workplace.

### **New Tools under Development**

Both PARCC and Smarter Balanced emphasize the importance of supporting teachers, and instruction more generally, by designing pedagogically useful assessments and providing a variety of classroom resources. These objectives are in line with the broader goals of the CCSS, which emphasize deeper learning of core subject matter through better-articulated learning progressions. Beyond developing a range of formative assessments, the tools available through PARCC and Smarter Balanced may include the following:

- An online interface for developing custom reports on schools or students
- Measures of growth to track student progress toward college readiness
- Banks of test questions for classroom use
- Model lessons
- Curricular frameworks
- Established cadres of educational leaders tasked with supporting districts as they implement the tests

All of these tools will be organized into a warehouse of research-based supports and interventions to support students falling behind academically, including subgroups like English learners. Both consortia are also piloting evaluation tools that educators can use to provide feedback as implementation gets under way. To support these activities, the US Department of Education has awarded both groups add-on grants for transition supports.

### **Additional Resources**

Given that enhancing teaching is a primary goal of both consortia, detailing all of the supports available to educators and students is well beyond the scope of this report. For more information, please refer to the following resources.

#### **Smarter Balanced site for teachers**

<http://www.smarterbalanced.org/>

#### **PARCC site on educational resources**

<http://www.parcconline.org/>

#### **ETS update on CCSS assessments**

[http://www.k12center.org/rsc/pdf/Assessments\\_for\\_the\\_Common\\_Core\\_Standards\\_July\\_2011\\_Update.pdf](http://www.k12center.org/rsc/pdf/Assessments_for_the_Common_Core_Standards_July_2011_Update.pdf)

#### **ETS (Kyllonen) on 21<sup>st</sup> century skills measures**

<http://www.k12center.org/rsc/pdf/session5-kyllonen-paper-tea2012.pdf>  
[http://www.usc.edu/programs/cerpp/docs/Kyllonen\\_21<sup>st</sup>\\_Cent\\_Skills\\_and\\_CCSS.pdf](http://www.usc.edu/programs/cerpp/docs/Kyllonen_21<sup>st</sup>_Cent_Skills_and_CCSS.pdf)

As these examples illustrate, multiple-choice items often involve tradeoffs among technical, practical, and instructional considerations. Practically and technically, the format is popular because it is inexpensive to administer, easy to score, allows for many questions in a short amount of time, and, due in part to this large sample of items, tends to produce more-reliable scores than other formats. Instructionally, multiple-choice items can be used to measure much more than merely factual recall or declarative knowledge. Nonetheless, there are constructs that cannot be assessed simply using multiple choice, such as the ability to generate a large number of possible solutions to a problem or the ability to compose an essay. Other types of assessments can help address such limitations, but often with accompanying technical challenges related to reliability and validity. As we continue to move through these format types, it will be clear that as the responses required of students become more complex, the technical hurdles do as well.

### **Self-Report (Likert)**

Likert-style self-report items are a subset of multiple-choice items that warrant additional consideration, given their prevalence in surveys, which are often used to measure 21<sup>st</sup> century competencies. These items ask students to rate themselves on a variety of factors, and responses are usually along a scale, such as “strongly agree” to “strongly disagree.” For example, the College and Career Ready School Diagnostic (developed by David Conley)—a measure that assesses several 21<sup>st</sup> century competencies—asks students to respond to statements like “I take complete, organized, and accurate notes during class” and “I talk to adults when I have concerns about school.” In general, questions in self-report format are commonly used in part because they are a less expensive means of assessing 21<sup>st</sup> century competencies than some of the more open-ended formats. Nonetheless, there are several technical problems that can arise with self-report items, including trouble assessing oneself accurately, which results in biases.

Self-report items are being used to measure some emerging psychological constructs. Several of these measures are shown in Table 3. For example, Carol Dweck, an expert in motivation and personality, assesses growth mindset using a very brief self-response questionnaire that can be found on her website: <http://mindsetonline.com/testyourmindset/step1.php>. Test takers respond to several items using a six-

point scale ranging from “strongly agree” to “strongly disagree.” Specifically, students or teachers react to prompts that include “Your intelligence is something about you that can’t change very much” and “You can learn new things, but you can’t really change your basic level of talent.” Responses to these questions are compiled in order to place test takers on a scale from having a fixed mindset to having a growth mindset. Other measures of important intrapersonal constructs use similar methods. For instance, Angela Duckworth and colleagues have developed a grit scale, for which they have measured reliability and begun to make a validity argument (Duckworth and Quinn 2009). Though more established, the Work Extrinsic/Intrinsic Motivation Scale (WEIMS) uses a comparable self-rating approach to measure a student’s level of intrinsic motivation (Ryan and Deci 2000a).

Likert-style self-report items have strengths and weaknesses when it comes to their technical, practical, and instructional value. On the plus side, they can be used to measure psychological constructs such as grit at a low cost. Technically and instructionally, like multiple choice, these items provide the ability to ask many questions quickly in a standard format, which can generate a broad range of information for teachers and increase reliability. Yet other technical problems can undermine these benefits. For example, students taking a survey in class may be concerned about how the teacher will perceive their answers, incentivizing responses that are more favorable to them but do not reflect reality (Lombardi, Seburn, and Conley, 2011)<sup>4</sup>. In addition, responses to hypothetical situations on paper may be less nuanced or altogether different than if students actually engaged in the activity being described. These threats to validity can undermine the utility of these measures for understanding related student needs.

### Closed-Ended Computer-Based Items

One way that computers are allowing for more complicated item formats than multiple choice is by allowing test takers to construct their own answers to questions, but in highly constrained ways that involve selecting or moving items on the screen rather than entering their own responses. This approach is being used frequently by both PARCC and Smarter Balanced, especially for math questions. The item provided in Figure 3 is an example. In this question, high school students are asked to examine a spreadsheet being used to develop a plan to pay off a credit card. Unlike multiple-choice items, where several possible answers are provided and the test taker must choose one, this question requires construction of different formulae using the available numbers, functions, and cells from the spreadsheet.

This format has potential to provide advantages over standard multiple-choice items. Instructionally, while these items do not give examinees complete freedom to generate their responses, they do significantly reduce the chance that a student will obtain the right answer by guessing, and they provide opportunities for students to construct responses (albeit in a constrained way) rather than select from a small number of preexisting response options. In general, compared with multiple-choice items, these closed-ended computer-based items show potential for better balancing technical and instructional issues, eliciting more-complex responses without eschewing the standardization that facilitates reliable scores.

---

<sup>4</sup>Such issues of self-perception can bias other item formats as well, but the issue is especially germane to items dealing directly with student beliefs about their own knowledge, skills, and attitudes.

**FIGURE 3.**  
Sample PARCC Spreadsheet Item

Isabella owes a balance of \$300 on her credit card. She has stopped making purchases with the card, and she plans to make a \$40 payment each month until her debt is paid and her credit card balance is \$0. The monthly rate is 1.5%, and interest is added each month to the balance that remains.

Consider the spreadsheet. In a spreadsheet, each entry (cell) is referred to by its column letter and row number. For example, 260.00 is the entry in cell D2 of this spreadsheet.

	A	B	C	D	E
1	Month	Amount owed (\$)	Monthly payment (\$)	Remaining amount owed after payment (\$)	Amount owed after 1.5% interest charge(\$)
2	1	300.00	40.00	260.00	263.90
3	2	263.90	40.00		

A3	B3	C3	D3	E3	0.015	1.015	×	÷	+	-
----	----	----	----	----	-------	-------	---	---	---	---

Drag the tiles to write a formula to find the value of cell D3.

D3 =

Drag the tiles to write a formula to find the value of cell E3.

E3 =

[Submit Answer](#)

Retrieved from [http://www.ccsstoolbox.com/parcc/PARCCPrototype\\_main.html](http://www.ccsstoolbox.com/parcc/PARCCPrototype_main.html)

### Open Response

Open-response questions elicit written responses on paper or on a computer or other electronic device. For instance, responses to these items often come in the form of a few written sentences, a paragraph, or even a full essay. This format facilitates the measurement of constructs that can be difficult or impossible to measure using the more constrained multiple-choice format, and it may be particularly well suited to measures of nonroutine problem solving or creativity. At the same time, these items pose technical challenges by making high levels of reliability more difficult to attain. Though response formats are usually constrained to simplify technical challenges (students are required to write an essay of a certain length, for example), student answers vary more than for the formats discussed above, and scores often demonstrate lower levels of reliability. These items typically need to be scored by human raters, though advances in automated scoring of written responses have reduced this need in many cases. In addition, because these items take longer to complete, fewer of them can be administered in a given testing time than is possible with shorter, multiple-choice items, and a reduction in the number of items typically reduces reliability. Although comparing reliability across multiple-choice and open-response items is complicated because the latter format has a source of error that is not present in the former (i.e., error due to raters), research does show that the difficulties in producing assessments that generate reliable scores are typically greater for open-response than multiple-choice and other closed-ended items (Ackerman and Smith 1988; Koretz 1998).

A variety of common standardized tests have addressed some of these technical challenges by providing explicit scoring criteria, training raters on those criteria, and blending open-response and multiple-choice items. One example is the essay portion of the Graduate Record Examination (GRE). Students are given a prompt and then write a structured essay supporting an argument related to the prompt. The quality of the essay is then rated on a scale from one to five, with five being a perfect score. A sample writing prompt on the ETS website asks students to agree or disagree with the following statement: “As people rely more and more on technology to solve problems, the ability of humans to think for themselves will surely deteriorate.” Another example of a test with open-response items is the Gao Kao, which includes free-response mathematics items and an essay question. The latter is eight hundred characters long and attempts to measure a student’s writing and critical-thinking skills. In the past, students have been asked what Thomas Edison would think of mobile phones if he were alive. Other topics have included living a balanced life and the future of youth (Carlson and Chen 2013).

Some test developers have begun to adapt open-response items to measure 21<sup>st</sup> century competencies. One example is the Formulating Hypotheses test, a measure of creativity developed by Educational Testing Services (ETS; see Figure 4). The structure of the test is relatively simple: Students are given a prompt and then asked to provide as many responses as possible to address the prompt. For instance, students are given a graph showing the declining rate of death from infectious disease in a fictitious developing country; then they are asked to list as many potential causes of the decline as possible in a finite amount of time. In addition to being fairly simple to administer, Formulating Hypotheses also presents a glimpse of the future. Currently, ETS is working on a more comprehensive version of the measure for potential inclusion in some of its major standardized tests. Many critical-thinking tests also give students some freedom to respond, albeit constrained. For instance, the Assessment of Reasoning and Communication, which was developed by ACT, requires students to produce three short written essays and speeches on a variety of topics. Students then receive scale scores on social, scientific, and artistic reasoning.

**FIGURE 4.**  
Sample ETS Formulating Hypotheses Item

00:29
GRE - Section 1: Formulating Hypotheses
1 of 4

**Disease in Alcadia**  
Declining Rate of Death From  
Infectious Diseases in Alcadia

Year	Deaths per 100,000 Persons
1900	750
1920	600
1940	500
1960	350
1980	250

According to the graph above, the rate of death (per 100,000 people) from infectious diseases in Alcadia declined steadily from 1900 to 1980.

Think of hypotheses (possible explanations) to account for this decline.

Write each hypothesis as a separate answer of no more than 15 words.

1. More people are vaccinated
2. Better sanitation
3. More doctors in the area
4. People have healthier life-styles
- 5.
- 6.
- 7.
- 8.
- 9.
- 10.
- 11.
- 12.
- 13.
- 14.
- 15.

Edit
Save

Improved health education in schools

Section Time

? Answer →  
Help Confirm Next

Source: Bennett, R. E., and D. A. Rock. (1995). "Generalizability, Validity, and Examinee Perceptions of a Computer-Delivered Formulating Hypotheses Test." *Journal of Educational Measurement* 32(1): 19-36.

This item format has potential benefits to instruction. In many situations, open-response items offer a viable alternative to the highly prescriptive answers used in multiple-choice and Likert-scale items, potentially improving the instructional value of an item. Compared to closed-ended items, the open-ended nature of responses arguably gives teachers more information about how students think, thereby adding instructional value to the measurement of constructs such as critical thinking. This format can also be used to measure academic mastery in certain domains, such as students' ability to analyze historical documents or to generate a well-written essay.

However, use of these items involves somewhat murkier tradeoffs among technical, practical, and instructional considerations, as addressing the technical issues associated with this approach is difficult and time consuming, especially if school systems wish to use them to compare across a broad range of students. Even when a clear set of criteria are used to assign scores to an open-ended response, scores assigned by different raters can vary considerably. Ensuring consistency across raters typically requires raters to be trained on what it means to reach a given performance level on those criteria. Needless to say, developing criteria and training teachers on them requires time and money. By the same token,

addressing these technical issues can provide additional instructional benefits by encouraging teachers to discuss and decide upon clear-cut standards for both content knowledge and analytical writing. Therefore, school systems can overcome technical challenges and can potentially gain information valuable to instruction, but only if they are willing to devote the resources required to engage in the process of developing the exam.

## Portfolios

Portfolios are collections of student work that are scored against some predetermined criteria. Though these assessments, like open-response measures, can be difficult to score consistently, they are gaining in popularity as a way to assess some 21<sup>st</sup> century competencies. This rise in usage stems in part from increased work in the assessment community to ensure that portfolios can be scored in a way that ensures high levels of reliability and validity. Whereas teachers and schools have used portfolios to assess student progress for decades, these newer prototypes treat collections of student work more like a series of test items with measurable properties. That is, by using a highly structured scoring process not unlike those developed to score open-response items, consistency of the criteria for scoring—and consequently, the agreement among raters—increases. Some of the most cutting-edge assessment work related to portfolios has been conducted by the Stanford Center for Assessment of Learning and Equity (SCALE), which has worked with a number of school systems to make this testing format useful for teachers and generalizable across students.

Recently, Asia Society partnered with SCALE to develop the Graduation Performance System (GPS). An important part of the GPS is a portfolio of student work that is used to measure student progress in a number of areas, with particular emphasis on global competence. In the GPS framework, global competence is broken down into constituent skills, including investigating the world, weighing perspectives, communicating ideas, taking action, and applying expertise within and across disciplines. More broadly, the GPS is intended to assess critical thinking and communication, among other competencies. The GPS gives local practitioners a great deal of flexibility in terms of what the portfolios include, though Asia Society fosters consistency by providing standards for the portfolio content (referred to as a “graduate profile”), a series of discipline-based performance targets and rubrics, sample curricula, and examples of student work (such examples can be seen on Asia Society website: <http://asiasociety.org>). All in all, the GPS goes beyond a portfolio system by providing guidance on rubrics, model design and implementation, assessment of student work, combining that work into a portfolio, and determining whether that final product meets standards to deem the student globally competent (see Figure 5).

**FIGURE 5.**  
Graduation Performance System (GPS) Framework

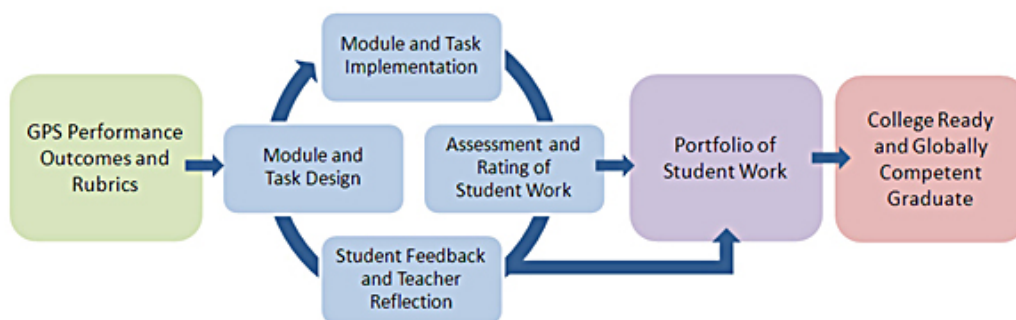


Image courtesy of Asia Society.



Singapore has also instituted a portfolio system for measuring student development in the elementary grades. These portfolios include drafts of individual work, reflections, journal writing, paper-and-pencil tests, self-assessments, and feedback from teachers, parents, and peers. By incorporating a broad range of student work, the Singapore Ministry of Education is attempting to foster competencies such as knowledge application, communication, collaboration, and learning to learn. In particular, the inclusion of much work that is still in progress is meant to place emphasis on the process of learning, including developing student self-awareness of that process.

Use of portfolios involves a delicate balance between technical and instructional priorities not unlike those associated with the open-ended format. Research shows that even some portfolio systems with significant financial and research support may fall short of technical soundness. For example, a portfolio system developed in Tennessee to evaluate teachers was dropped after only a year due to teacher resistance and technical setbacks (Sanders and Horn 1998). Similar findings have been documented elsewhere (Koretz et al. 1994; Koretz 1998). Another major challenge is that teachers have more power to give feedback to students, which means they can influence results and therefore bias scores as a measure of student performance (Stecher and Herman 1997). Yet this involvement of teachers is also, in some cases, a pedagogical strength. Research suggests that the process used to improve reliability on these measures—namely, training teachers on academic standards and involving them in testing decisions—can positively influence instruction and curriculum design (Stecher 1998). Certainly, anecdotal evidence from Australia and Singapore suggests that portfolios can make teachers more vested in improving performance on the standards being measured. Ultimately, developing portfolios involves a tradeoff between technical soundness and instructional richness that make them an appealing alternative to multiple-choice tests.

## Performance Assessments and Simulations

One of the drawbacks of multiple choice and other forms of assessment with constrained responses is they do not appear to be “true to life” or “authentic.” Performance assessments and simulations are forms of assessment that are more authentic by having students perform tasks that look very much like real-world activities. For instance, a performance assessment of communication and collaboration would not ask students to respond to scenarios or rank personality traits on paper; rather, students might be asked to give a group presentation to peers and teachers, on which they would be scored. These tests have the potential to make assessment more meaningful for students while encouraging teachers to engage in the type of instruction that 21<sup>st</sup> century employers value—or so the argument goes. However, issues related to technical shortcomings and cost significantly complicate the use of performance assessments. Research shows, for instance, that the cost of administering performance assessments is three times greater than for open-response measures when reliability of scores is comparable (Stecher and Klein 1997).

Performance assessments have been around for quite some time yet have not been incorporated into large-scale assessment due to their cost and complexity. Precursors to the more sophisticated, technology-based versions now on the market typically involve having a student perform an activity, then asking multiple raters to score the performance. Academic mastery in science, in particular, has been the focus of several different performance assessments over the past few decades. For example, one task that was the subject of research involved giving students three paper towels with different characteristics (e.g., thickness), then asking them to use laboratory equipment to determine which of the towels absorbed the most water. Though seemingly straightforward, scores on tasks like these often had low reliability, because a student’s score depended greatly on both the particular task and the rater assigned (Gao et al. 1994; Klein et al. 1997; Ruiz-Primo and Shavelson 1996). They were also difficult to set up, requiring equipment, props, lab materials, and the like.

Nonetheless, in recent times, a wide range of performance assessments that appear to meet minimum technical standards have been developed. In addition, many of these assessments measure multiple, overlapping competencies, such as critical thinking, academic mastery, and communication. Though we cannot discuss them all here, many are detailed in the case studies in the appendix. Given these resources, we will focus here on two examples. The first is a computer-based simulation designed to measure not only proficiency in a foreign language, but also communication and problem solving (see Figures 6 and 7). Developed by Alelo Inc., the computer program allows a student to interact directly with an avatar—a realistic, computer-generated human being—in a variety of languages. Moreover, each language is culturally specific. For instance, if the simulation is focused on Argentine Spanish but the student uses, say, a Mexican colloquialism, the avatar will respond accordingly. Each time students enter the simulation, they must negotiate a specific outcome. For beginners, this might include scheduling a time to study. For advanced students, this might involve negotiating between conflicting parties in an argument. At both levels, critical thinking and communication are involved. Here, the very definition of assessment is relaxed. Students do not take a formal test of any kind; rather, they receive constant feedback on their performance, which they can then use to improve at their own rate.

**FIGURE 6.**  
Alelo Oral Language Simulation



Source: Johnson, W. L., and S. B. Zaker. (2012). “The Power of Social Simulation for Chinese Language Teaching.” Proceedings of the 7th International Conference & Workshops on Technology and Chinese Language, University of Hawai‘i at Manoa.

The second example of a performance assessment or simulation is a new portion of the PISA (developed in partnership with ETS) that will measure collaborative problem solving (CPS; see Table 4). Unlike tests that focus only on problem solving or collaboration, this one intentionally blends the two. For example, students are judged on their ability to resolve a situation through effective exchanges with their partners. Rather than using a live collaborator, PISA uses a system like Alelo’s (and many other promising performance assessments): a computer simulation with avatars. Researchers at PISA and ETS decided against pairing actual students together for the assessment because of threats to reliability and validity stemming from the potential for a student’s score to be influenced by the skills and behaviors of his or her partner. By using an avatar, test developers can directly control the skillset that a given student’s partner will possess. As a result, the difficulty of the simulation can be varied through the questions asked and the avatars assigned. We describe both the Alelo and PISA simulations in greater detail in the appendix.

**FIGURE 7.**

**Alelo Oral Language Simulation Learning Objectives**



Johnson, W. L., and S. B. Zaker. (2012). “The Power of Social Simulation for Chinese Language Teaching.” Proceedings of the 7th International Conference & Workshops on Technology and Chinese Language, University of Hawai‘i at Manoa.

**TABLE 4**

**PISA Draft Collaborative Problem-Solving Framework**

Probe	Skill Assessed
What does A know about what is on your screen?	(A1) Discovering perspectives/abilities of team members
What information do you need from B?	(C1) Communicating with team members about the actions being performed
Why is A not providing information to B?	(D1) Monitoring and repairing the shared understanding
What task will B do next?	(B2) Identifying and describing tasks to be completed
Who controls the factory inputs?	(B3) Describe roles and team organization
Write an email to your supervisor explaining whether there is consensus of your group on what to do next	(B1) Building a shared representation and negotiating the meaning of the problem (B2) Describing tasks to be completed
Write an email to your group explaining what actions the group will need to do to solve the problem	(B2) Identifying and describing tasks to be completed (C2) Enacting plans

Office of Economic Cooperation and Development. (March 2013). PISA 2015 Draft Collaborative Problem Solving Framework. Retrieved from [www.oecd.org](http://www.oecd.org).

In sum, performance assessments involve some of the greatest risks and rewards when it comes to balancing technical, practical, and instructional issues. While instructionally they may be the most true to life in terms of the demands placed on students, they also introduce many possible sources of error that complicate technical matters. For example, reliability could be undermined by inconsistent scores across raters, tasks, and even a student's own performance on the same task repeated at different times. Therefore, these measures may be useful for formative purposes but are difficult to use for accountability purposes. (There's a reason performance assessments have been around for decades but have largely not been incorporated into accountability systems across the globe.) The development of technology-based performance assessments has addressed some of the limitations of hands-on assessments. Although they require computing infrastructure, which costs money, once this infrastructure is in place, simulated environments can save teachers time and decrease costs associated with nonsimulated performance tasks, such as physical materials. Further, automated essay-scoring technology can eliminate costs associated with rater training and scoring.

Beyond the performance tasks we describe in this chapter, the appendix also explores two science simulations, one aimed at blending instruction with assessment (EcoMUVE), the other developed with an intentional option to use it for accountability purposes (SimScientists). Though most of these assessments are in prototype stages, they nonetheless represent the direction that many measures of 21<sup>st</sup> century competencies may be heading.

### Measuring Leadership: A Lesson in the Difficulties of Assessing 21<sup>st</sup> Century Competencies

One competency included in this report, leadership, is not like the others in one important way: There are still very few technically sound and practically feasible measures of leadership. Even with the potential of technology to improve score reliability by making test items and environments more stable, there are complexities inherent in measuring leadership that cannot be solved by technology alone. Although we could have omitted leadership from the report because of our focus on competencies that are actionable and measurable, we included leadership in the report because (a) it is seen as valuable by nearly all the organizations interested in 21<sup>st</sup> century competencies, and (b) the complications inherent in measuring leadership are relevant to other competencies.

The greatest hindrance to the measurement of leadership is that the construct is not well defined. Even experts disagree about what makes an effective leader (Berman et al. 2013; Childers 1986; Walumbwa et al. 2008). Part of the disagreement arises because leadership requires so many constituent skills. As previously discussed, leadership involves effective communication, collaboration, and creativity. These skills are difficult to measure on their own, let alone in concert. Subsequently, measures of leadership that are reliable and assess more than some small aspect of effective leadership are sparse.

This finding does not, however, mean that schools and school systems cannot foster leadership in students and attempt to assess it. For example, the Houston Independent School District asks students to give group presentations of research findings to executives from local businesses (especially from the oil industry) that hire their students. As part of the process, the executives score student performance on their presentations. While these scores may not be very reliable, students learn how leaders in local industry think about a given topic, and they receive constructive feedback that can be used to improve communication, collaboration, and leadership. This Houston example shows that even when there are no off-the-shelf assessments available, and even when developing an assessment locally may not generate a measure that meets basic technical standards, there can still be educational value inherent in the process of assessment that outweighs such technical limitations.

## Additional Sources of Data

Although this report does not discuss sources of data on student acquisition of 21<sup>st</sup> century competencies beyond assessments, educators should nonetheless be aware that information other than test scores is important to building a thorough, balanced assessment system. In particular, most educators have access to administrative data that can be used to measure aspects of relevant competencies. For example, research shows that data on student attendance, punctuality, and behavioral infractions (e.g., suspensions or expulsions in the US school system) are relevant measures of intrapersonal competencies such as motivation and grit (Conley 2005; Conley 2008). When practitioners consider whether and how to assess a 21<sup>st</sup> century competency, they can start by determining whether a source of data other than an assessment already captures that information, as well as how test scores can be combined with these additional data to create a more nuanced picture of students' competency levels.

## CROSS-CUTTING MEASURES: A DEEPER LOOK AT THE MISSION SKILLS ASSESSMENT (MSA)

The previous examples of assessments using different formats illustrate some of the tradeoffs inherent in selecting a measure; however, they do not offer much insight into how assessments can provide information that changes what happens in the classroom. This section provides an example of an assessment of 21<sup>st</sup> century competencies, the Mission Skills Assessment (MSA), which is currently being used to improve student outcomes in dozens of independent schools. To understand how the MSA is being used to shape instruction, we interviewed teachers, principals, and other administrators at three participating schools.

We chose the MSA for a variety of reasons but primarily because it is innovative yet not overly costly or overly reliant on technology—most of the components are completed using pencil and paper. In addition, the MSA combines several different measures, thereby safeguarding against some threats to reliability and validity that might be associated with a particular format by using different methods to triangulate a student's score. This section of the report allows us to consider how educators are using an innovative, relatively inexpensive, high-quality measure to change practice.

### The Intent of the Mission Skills Assessment

The Mission Skills Assessment (MSA) is a collection of instruments being developed by ETS in conjunction with the Independent School Data Exchange (INDEX). Broadly, the purpose of the MSA is to measure the interpersonal and intrapersonal competencies that many independent schools value, both in classrooms and during the admissions process. In particular, the MSA allows schools to measure several competencies, some of which we emphasize explicitly in this report: collaboration, creativity, ethics, resilience (similar in many respect to grit), intrinsic motivation, and learning to learn (especially time management). With the exception of growth mindset, the MSA assesses virtually the entire set of competencies included in the intrapersonal category.

At this stage in its development, the MSA consists of student surveys, multiple-choice questions (especially related to situational judgment), and teacher observations of student behavior. Though not all competencies are measured using each of these approaches, virtually all of the competencies are measured using multiple assessment formats, a strategy we discuss in greater detail when considering the technical quality of the MSA. Students are scored on each competency, and then those results are aggregated up to the school level. All in all, the MSA is meant to provide schools guidance on whether they are succeeding in their mission to promote not only academic but also inter- and intrapersonal development.

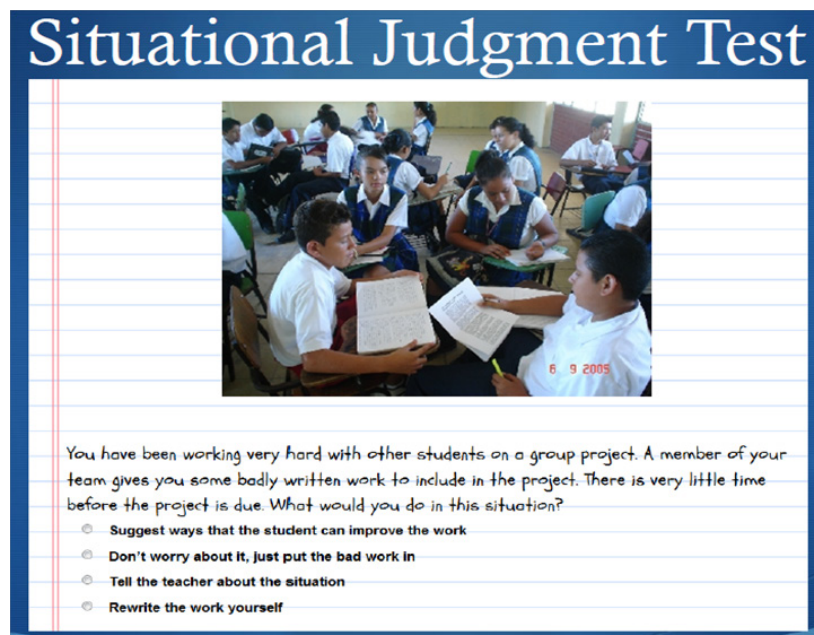
The MSA is used for institutional improvement purposes, and the results are reported only at the school level. This approach is intended to provide schools with useful aggregate information and to allow

them to compare results with other INDEX schools while protecting students from potentially unfair labeling at a very young age. Nonetheless, the MSA can play a formative role. Our interviews with teachers and school leaders suggest that the MSA's ability to chart a school's progress in promoting these competencies can help generate conversation around them and ultimately establish a culture focused on these priorities. This focus has, in the schools we consulted, translated into more regular, informal measurements of these competencies by teachers.

## Practical Considerations

Designers of the MSA attempt to overcome technical difficulties in measuring these competencies by triangulating across different measures. That is, the MSA includes student self-reports, teacher observations, and situational judgment tests. By using multiple instruments to assess the same construct, MSA developers can better disentangle sources of error and thereby increase the precision of the measurement. Though the measures for each competency differ, they generally include the following components each year: sixty minutes of student self-assessment, situational judgment tests (SJTs; see Figure 8) and other performance measures, teacher ratings of students, and outcome data such as grades. For example, resilience is measured by asking students and teachers to rate the student's ability to overcome setbacks, and by recording student multiple-choice responses to a hypothetical situation in which he or she has too much homework or is faced with another stressful situation. One benefit of this approach is that it does not rely on technology, which can improve the feasibility of administration in schools that lack the required technology infrastructure. The time constraints also do not appear to be overly burdensome. Educators with whom we discussed the MSA said that the test requires some additional teacher time to fill out ratings of students, but that the requirement is minor, taking roughly a day. The same teachers also reported that the ratings process itself proved valuable to their teaching, a suggestion we discuss in the section on instructional considerations.

**FIGURE 8**  
MSA Situational Judgment Test



Copyright Independent School Data Exchange, Educational Testing Services (ETS).

## Technical Considerations

Initial evidence suggests the MSA at least meets minimum technical standards. The reliability of scores on the combined assessments for each competency, as measured by both internal consistency and test-retest reliability, is high (in fact, the reliability of scores for measures of some competencies is on par with similar estimates for the SAT). Further, researchers at ETS have shown that MSA measures predict not only academic outcomes but also the student's overall well-being, as assessed by self-reports of life satisfaction<sup>5</sup>. Many MSA measures also do a better job of predicting both academic and nonacademic outcomes than do scores on standardized academic achievement tests. For example, both time management and teamwork have higher correlations with well-being than mathematics and reading scores. Similarly, intrinsic motivation has a higher correlation than achievement scores with teacher reports of student quality.

## Instructional Benefits and Challenges Associated with the MSA

Here we discuss some of the instructional benefits and challenges associated with measuring 21<sup>st</sup> century competencies using the MSA. None of the benefits occurred automatically for the schools in question; they arose from purposeful changes in practice facilitated by access to assessment scores. One administrator commented that the teachers themselves must have a growth mindset about their practice when trying to make data on 21<sup>st</sup> century competencies actionable and integrated into the curriculum, because the entire endeavor occurs in uncharted territory. While these perspectives are from only a handful of schools and are therefore not necessarily generalizable, they nonetheless provide a glimpse of how the measures in this report are being used in practice.

### *Being More Intentional*

Perhaps the most common benefit cited by practitioners was that having data on these competencies allowed them to be more intentional about fostering them. That is, having MSA scores allowed educators to be more strategic about incorporating 21<sup>st</sup> century competencies into the school's environment. Several examples help illustrate how this intentionality works. For one, by seeing how the MSA measures creativity, teachers across schools were better able to incorporate it into their own teaching and assessment. For instance, several teachers have built the MSA measures into the rubrics they use to grade projects in core academic content. Similarly, other teachers have created displays on classroom bulletin boards showing examples of students' work that demonstrates the desired competencies. Administrators often play a major role in incenting these practices. One head of school requires teachers to provide evidence on a regular basis of how they are measuring and teaching these competencies in the classroom. Professional development is then provided based in part on that evidence: Teachers visit one another's classrooms with a curriculum specialist to view different approaches to teaching and measuring the competencies, then follow up with a discussion about aspects that did or did not work well.

### *Building a Common Vocabulary*

Teachers also reported a heightened ability to collaborate on fostering 21<sup>st</sup> century competencies by sharing a common vocabulary. According to practitioners, conversations about improving these skills are likelier to occur and unfold more efficiently when all the teachers are intimately familiar with the skills highlighted by the MSA. As an example, teachers at several schools are now posting videos to a website of successful lessons, categorized by skill on the MSA. Similarly, participating schools often use these competencies on report cards, replacing more nebulous yet oft used concepts like "effort."

---

<sup>5</sup>Life satisfaction was measured with the Students' Life Satisfaction Scale (Huebner 1991), which includes responses to statements about how well the respondent's life is going and whether the respondent has what he or she wants in life.

This common vocabulary improves communication not only among educators but also with students and parents. According to interviewees, students are now aware that they are being assessed on these competencies and, as a result, frequently work hard to demonstrate possession of them. Parents in turn learn from their children and their teachers that such competencies are valued and can think about them when helping with homework.

#### *Tying Individual Skills to Outcomes of Interest*

Several educators with whom we spoke also highlighted that measuring competencies individually and with specificity better allowed them to connect these competencies to outcomes of interest. For example, results from MSA-based research suggests that time management is correlated with outcomes of interest, including grade point average. Though teachers suspected that such a correlation existed, the confirmed evidence proved helpful, in part by disentangling the effect of time management from other factors, such as motivation. Being more detailed in turn facilitated the generation of targeted supports. In the case of time management, research from certain schools showed that students in the bottom two quintiles did much worse on the outcomes of interest, but there were no substantive differences among students in the top three quintiles. Therefore, rather than attempt to move students not at the top quintile into the upper 20 percent, teachers are instead focused on improving time management among the bottom two quintiles, such that those students end up with a score that matches those of students in the middle quintile. More broadly, this practice suggests that teachers can potentially set more accurate and nuanced performance benchmarks for students using measurable data.

#### *Making Curriculum More Engaging*

Teachers across sites suggested that, somewhat surprisingly to them, measuring these 21<sup>st</sup> century competencies actually helped make curricula more engaging. As an offshoot of incorporating creativity, communication, and collaboration into projects and their grading schema, the assignments became more multifaceted and therefore more engaging. For example, one school requires students to complete a culminating academic project at the end of the 8th grade. This project has always involved research, writing, critical thinking, and a verbal presentation of findings on a self-chosen topic that relates to content in core academic subjects. Whereas students used to be graded on these facets, the rubric now also includes creativity, ethics, resilience, and intrinsic motivation. According to teachers, the emphasis on creativity in the final grade has led to more inventive projects. Additionally, students are assessed on ethics based on how well the project articulates benefits for others, especially people from diverse backgrounds and circumstances.

#### *Making Students Feel More Valued*

Teachers report that with the right approach, measuring 21<sup>st</sup> century competencies can make students feel more valued in the schooling community, especially students who are not as academically gifted. This finding is especially true for students with disabilities. According to one teacher who works specifically with dyslexic students, “these kids are all about resiliency.” In order to overcome the challenges that can accompany a disability, students must remain resilient in the face of setbacks. According to several practitioners, measuring nonacademic competencies means these students can receive well-deserved recognition, as they often score quite highly on MSA measures of these competencies. Conversely, interviewees reported that students who are gifted academically can also be shown that there are additional competencies they need and must work to master.

The same practitioners, however, caution that producing these results involves framing them thoughtfully and productively. Administrators across schools try to consistently frame talk around skills using the vocabulary of growth mindset, suggesting repeatedly to teachers that students can improve skills. This emphasis occurs despite inconclusive research on the mechanisms that underlie some of these skills and, as a result, uncertainty over how much power, exactly, teachers have to shape them. As



one educator pointed out, there is no downside to assuming that these skills are entirely teachable and proceeding accordingly. Taking this approach helps ensure that students do not receive stigmatizing labels that could occur if they have low scores on a particular skill and feel they cannot do anything to improve their performance.

## Instructional Challenges

Despite the benefits articulated by educators with whom we spoke, there are accompanying challenges, including fostering buy-in, managing ambiguity, and avoiding labeling of students. Interviewees also stressed, however, that the challenges mainly arise because much of the work around fostering 21<sup>st</sup> century competencies is new and therefore requires a degree of trial and error. Though adjusting practices can at times be challenging, practitioners report that the process itself can be valuable in improving instruction and, ultimately, student outcomes.

### *Fostering Buy-In*

Even in schools with a long-standing dedication to fostering nonacademic competencies, the process of measuring them yielded initial skepticism on the part of many teachers, which our interviewees attributed to two main factors. First, teachers increasingly tend to associate measurement with accountability and are oftentimes wary that an assessment will be used primarily as an evaluation tool, either formally or otherwise. Administrators at the schools devoted significant time to crafting and emphasizing a message around the intended use of the MSA; namely, that it would be used to monitor the effectiveness of the school as a whole but is primarily meant to help teachers generate effective practices around measuring 21<sup>st</sup> century competencies. Second, some teachers expressed concerns that the MSA and other measures being implemented might take too much time away from instruction. However, according to practitioners, this fear was assuaged once teachers administered the MSA and discovered that it was not overly burdensome. In fact, several teachers mentioned that filling out their assessment of students—the largest draw on their time—provided a valuable opportunity to reflect on student needs and think about being more intentional in emphasizing these competencies in the classroom. As with most new initiatives adopted by schools, measuring 21<sup>st</sup> century competencies was not accepted without skepticism, but it diminished as teachers became comfortable with the assessment and came to understand its intended uses.

### *Managing Ambiguity*

A main facet of generating buy-in is what one educator called “managing ambiguity.” Given best practices are still being developed around teaching 21<sup>st</sup> century competencies, teachers need to be flexible in their approaches and willing to try new techniques. For instance, educators mentioned being uncertain at times about what a construct was capturing, whether their classroom measures developed using the MSA as a guide actually assessed what they wanted, how student understanding of technologies being used related to the construct, and how best to teach the competencies amid such uncertainty. In most cases, practitioners attempted to manage this ambiguity by using it to spark conversation and drive improvement. For example, professional development often involved discussing instructional approaches, including what went well or did not, and having them critiqued by fellow teachers. Only by making intentional and steady adjustments to practice have teachers slowly begun to develop instructional techniques for teaching competencies of interest.

To become effective at fostering 21<sup>st</sup> century competencies, teachers may need to be open to trying new, untested approaches and seeking feedback from colleagues. The experienced teachers we interviewed had developed many lessons in core content over time and had opportunities to refine them. Lessons based around 21<sup>st</sup> century competencies, meanwhile, were necessarily a work in progress. For example, in the schools we contacted, teachers often worked on improving the delivery of instruction around these competencies by allowing other teachers to observe lessons. As a result, any flaws in a lesson were

suddenly on display. In some cases, teachers felt they were incorporating 21<sup>st</sup> century competencies into a lesson, when in fact the format used related vocabulary but still relied on rote learning (for instance, some teachers, especially in the beginning, would lecture on 21<sup>st</sup> century competencies). In other cases, a lesson did emphasize these competencies in meaningful ways but did not succeed for any number of reasons, including misunderstanding of the objectives among students or inadequate structure to the activity. Whatever the reason, administrators reported that making teachers more comfortable with imperfection is key to using measures of 21<sup>st</sup> century competencies effectively in the classroom. Instilling this comfort level meant changing the goal of classroom observation, which administrators often use to evaluate the quality of instruction rather than to learn collectively about improving practice.

### *Avoid Labeling of Students*

Finally, as discussed in the benefits section, practitioners reported the need to be very careful about how 21<sup>st</sup> century competencies were discussed, especially with students and parents. This delicacy was meant to avoid assigning students labels that might be stigmatizing, which research shows can have negative consequences for student achievement (Rosenthal and Jacobson 1968). According to teachers with whom we spoke, even when educators have the best interests of students at heart, being careless about communicating assessment results to students, especially when measuring intrapersonal or interpersonal competencies, can damage their self-perception. As one teacher suggested, at a time when test scores are used to make major determinations, including college admissions and ranking test takers, students are quick to associate measurement results with ability and potential. As a result, students can become discouraged if they feel that basic aspects of their personality, such as interpersonal effectiveness, are a reflection of their potential.

Teachers used several techniques to avoid having students interpret results in detrimental ways (even though the MSA is aggregated up to the school level for exactly this reason, teachers nonetheless embed measures based on the MSA into their practice). For example, the skills challenges previously described always begin with a discussion about how other students have improved those skills in the past, explicitly employing some of Dweck's growth mindset strategies (Dweck 2007). Another approach is to ensure that classroom assessments mirror the MSA's approach of measuring growth over time. In so doing, teachers can show that a student's overall performance level on a particular skill is less important than improvement. For instance, several teachers recognize students in their classroom who greatly improve on a skill, such as communication, between one assignment and the next rather than only recognizing overall performance on a single task.

## Conclusion

In this chapter, we provided examples of 21<sup>st</sup> century competencies assessments by format type in order to demonstrate the variety of approaches that developers have taken to measuring these competencies. The descriptions of each format show a pattern: as the tasks included on a measure look more like they do in the real world, the practical and technical challenges typically increase, sometimes dramatically. However, test developers are beginning to get around some of these issues by relying more heavily on technology, computer simulations in particular. Assessment developers are also beginning to clear technical hurdles by measuring a single skill with multiple formats, a triangulation strategy meant to increase reliability. Schools using one such set of measures, the MSA, have shown that assessing skills of interest does not need to be time consuming and can generate much richer dialogue among teachers than when evidence is entirely anecdotal.

## 5. GUIDELINES FOR MEASURING 21<sup>ST</sup> CENTURY COMPETENCIES

This paper is neither a comprehensive guide to 21<sup>st</sup> century competencies nor a comprehensive guide to educational assessment; however, by providing an overview of both topics combined with examples of 21<sup>st</sup> century competencies assessments, we hope to have provided educators with the background they need to make more informed choices about what to assess and how to assess it. This chapter offers school and school-system leaders some guidelines for promoting more effective and thoughtful assessment programs.

### Considerations When Adopting or Adapting Assessments of 21<sup>st</sup> Century Competencies

Our review of 21<sup>st</sup> century competencies assessments suggested a number of guidelines that could help improve the implementation of these assessments (see Table 5). These guidelines are not meant to be rules to follow as much as principles to keep in mind. Specific assessment needs will vary from site to site, and local priorities should dictate how these criteria are weighed and how decisions are made.

**TABLE 5**  
Key Takeaways from an Investigation of Available Measures of 21<sup>st</sup> Century Competencies

1. The process of selecting an assessment should begin with a determination of what purpose the assessment is intended to serve.
2. Tests that will be used to make consequential decisions need to meet higher technical standards than tests that are used for lower-stakes decisions.
3. The cost of assessment (both expenditures and time) should be weighed against the value of the uses it will serve.
4. More-complex assessments may be needed to measure more-complex competencies.
5. Innovative assessments (involving simulations, remote collaboration, etc.) can require substantial time and resources (e.g., training, computing power, telecommunications infrastructures).
6. 21 <sup>st</sup> century competencies cannot be measured equally well, and competencies that are not well defined are particularly difficult to measure.
7. If the desired assessments do not exist, districts can work with partners to develop them (partners can include other districts, researchers, and assessment organizations).
8. Context and culture matter, and assessments that work in one setting might not work as well in another. It is often necessary to conduct additional research to validate measures locally.
9. Acquiring information about students' understanding of 21 <sup>st</sup> century competencies can make educators and students more intentional about improving the competencies.
10. Educators (and learning scientists) do not know as much about teaching and learning 21 <sup>st</sup> century competencies as they do about teaching traditional academic content, so expectations for improvement need to be realistic.
11. Assessments can have unintended consequences, which should be monitored in each local context.
12. Measures of 21 <sup>st</sup> century competencies should be part of a balanced assessment strategy.

**The process of selecting an assessment should begin with a determination of what purpose the assessment is intended to serve.** The first thing educational leaders should consider when thinking about adopting an assessment is its purpose: Why measure critical thinking, leadership, or learning how to learn? How will the information be used? As discussed in Chapter 3, there are a number of possible purposes that might be served by an assessment, and each might lead to different assessment choices. For example, administrators might want to use the information as a monitoring tool to track the performance of the system and to make high-level decisions about allocating resources, assigning staff, etc. Alternatively, the purpose might be accountability—to identify teachers or schools that are performing poorly so as to intervene, or to find schools or teachers that are performing well to use as models. The purpose might be to set priorities; i.e., to send a clear signal to teachers and school leaders about what student outcomes are most important. Or the purpose might be to improve instruction by providing information to students or teachers that they can use to diagnose student weaknesses and prescribe new learning activities. The first step should be answering the question “for what purpose?”

In this paper, we focused on assessment for the purpose of improving teaching and learning; i.e., our approach was that the reason for adopting a new assessment is instructional. Assuming this is the case, educators may still be tempted to start the process of developing a measurement infrastructure by surveying the available assessments, then choosing the ones that seem most appropriate. While there is nothing wrong with this approach per se, it frequently means that tests drive instructional strategy rather than the other way around. When measurement products drive related practices, schools and school systems are at greater risk of adopting assessments that do not measure exactly what they care about, that fail to provide teachers information they do not already have, that prove duplicative of other efforts, or that do not provide information that is actionable for teachers. In short, it can be easy to be attracted to the innovativeness of different measures, but educators forget to ask a basic question: will adopting a measure result in better teaching and learning?

**Tests that will be used to make consequential decisions need to meet higher technical standards than tests that are used for lower-stakes decisions.** Rigorous technical standards are vital for assessments used to make decisions about student placement, attainment, and accountability. For instance, if students’ scores fluctuate significantly from one administration to the next purely by chance (low reliability), then real concerns arise that students might not be placed appropriately or, even worse, might be prevented from progressing to the next level in their schooling for arbitrary reasons unrelated to their academic ability. Concerns like these are the main reason that few large-scale accountability systems use performance assessments. When tests carry important consequences, it is critical that a student’s score not be influenced by an individual rater or task.

While the need for technical rigor for high-stakes tests may seem obvious, the situation is murkier for tests that are not used for such decisions. In some cases, these technical standards can be relaxed (though not ignored) when the assessment is formative, especially if educators wish to place a premium on the types of teaching and learning the test inspires. For example, if a performance assessment results in students performing tasks that are much like what will be asked of them in the workplace, and these tasks, in turn, encourage teachers to emphasize 21<sup>st</sup> century competencies more and rote learning less, then one might be willing to sacrifice a bit of reliability. This is especially true if the assessment is being used not to determine a grade or other benchmark but to provide teachers and students with useful information and practice in carrying out 21<sup>st</sup> century competencies. At the same time, technical standards should not be ignored completely for tests that are used only for formative purposes. If the reliability of scores on a particular measure is low, those scores will not provide useful information and should probably not be used even for low-stakes decisions.

**The cost of assessment (both expenditures and time) should be weighed against the value of the uses it will serve.** The assessments reviewed in this report vary both in terms of their financial cost and the time required for administration and scoring. Because educational resources are typically limited, the decision to adopt a costly or time-consuming test is likely to require a reduction in other activities, such as instructional time or professional development. Unfortunately, it is rarely possible to do a clear cost–benefit analysis of an assessment, because it is difficult to measure accurately either costs or benefits. Nevertheless, one should consider these issues when thinking about new assessments, particularly assessments of 21<sup>st</sup> century competencies.

**More-complex assessments may be needed to measure more-complex competencies.** Some 21<sup>st</sup> century competencies can be measured individually with established assessments that use paper and pencil and are not costly. This may be an encouraging fact for schools just beginning to develop measurement infrastructures or with limited budgets (or both). For instance, a district that is interested in measuring motivation but has limited technology can likely rely heavily on the WEIMS, a paper-and-pencil test. However, assessments measuring multiple or especially complex competencies and using tasks matching real-world activities often require innovative formats. That is, as the complexity of the competency increases, so too does its measurement. This continuum results in important cost tradeoffs. If educators wish to use measurement to generate a holistic picture of a student’s cognitive, interpersonal, and intrapersonal competencies, then significant cost and resources may be required. Much of this cost is related to expanding computing power.

**Innovative assessments (involving simulations, remote collaboration, etc.) can require substantial time and resources (e.g., training, computing power, telecommunications infrastructures).** Many assessments of 21<sup>st</sup> century competencies involve computer simulations of real-world scenarios. Beyond the increased availability of computing power, these simulations are popular because they help address technical considerations. As we pointed out in our discussion of the PISA CPS subtest, test makers at ETS prefer partnering students with avatars rather than with live peers, because the ability of the former can be controlled, which increases reliability and validity. These formats are also popular because manipulating an avatar’s responses to a situation means that the simulation can elicit specific skills not necessarily tied to academic mastery, such as critical thinking and resilience. While the benefits of computer- and simulation-based assessment are myriad, they can also tax a school’s resources, especially in a tough budget environment. When determining whether to invest in a software or online assessment package, districts and schools should consider their technological capacity, then determine what level of investment is justified in order to have a more realistic and multifaceted assessment.

**21<sup>st</sup> century competencies cannot be measured equally well, and competencies that are not well defined are particularly difficult to measure.** Despite all the advances in measurement and technology, some competencies still cannot be measured well. Oftentimes this phenomenon occurs because a competency involves several different overlapping component skills and therefore lacks clear definition. For example, as discussed in the previous chapter, there are assessments that measure leadership styles, but measuring a student’s overall leadership ability (regardless of style) is not especially feasible, in part because leadership involves so many facets, including communication, collaboration, and creativity. Though there are options for measuring leadership, there are none that have established high levels of reliability and validity. Although advances in technology will probably improve the quality and feasibility of measures of some skills, technology is unlikely to be helpful for measuring competencies that are not clearly defined.

**If the desired assessments do not exist, districts can work with partners to develop them (partners can include other districts, researchers, and assessment organizations).** Many of the assessments cataloged in this report were developed jointly by educators, government agencies, and

research organizations with psychometric expertise. For example, the MSA and PISA CPS were both developed by education policy makers and educators in collaboration with ETS. Similarly, Asia Society developed the GPS in part through a contract with SCALE at Stanford. While these partnerships relied on support from organizations that are staffed with professional psychometricians, some governments have relied primarily on internal expertise. The Queensland Performance Assessment (see the appendix) involved consultation with external psychometric experts at universities but was largely developed through collaboration between the Queensland government and local schools. These examples illustrate that new measures can be developed through partnerships if the assessments already in existence do not meet local needs.

**Context and culture matter, and assessments that work in one setting might not work as well in another. It is often necessary to conduct additional research to validate measures locally.** A particular measure might be supported by evidence that it predicts valued academic and social outcomes in a particular setting. However, the fact that a measure is predictive in one or a few settings does not mean it will be predictive in all settings or under all circumstances. Extra caution is warranted when considering measures of 21<sup>st</sup> century competencies, particularly interpersonal and intrapersonal competencies, because these may be more culturally and contextually dependent than traditional academic skills. To the extent possible, the validity of scores on a given measure should always be confirmed locally.

**Acquiring information about students' understanding of 21<sup>st</sup> century competencies can make educators and students more intentional about improving the competencies.** Even educators who already place great emphasis on fostering the competencies discussed in this report can still potentially benefit from measuring them. According to several practitioners, the measurement process often makes attempts to generate these competencies more intentionally. For example, an awareness of how a construct is measured can result in increased use of such measures in assignments and scoring rubrics, including measures shown to be reliable and valid. As a result, students also become more attuned to the importance of these competencies and think about them more concretely when doing their work. Moreover, the assessment process can make fomenting valued competencies more intentional across classrooms by providing a common vocabulary. For instance, teachers might attend a professional development meeting and not only be able to discuss specific competencies, but also have concrete student data to accompany those conversations. These concrete discussions can also occur with students and parents, especially if incorporated explicitly into report cards and parent–teacher meetings.

**Educators (and learning scientists) do not know as much about teaching and learning 21<sup>st</sup> century competencies as they do about teaching traditional academic content, so expectations for improvement need to be realistic.** While new research on teaching 21<sup>st</sup> century competencies is emerging all the time (Saavedra and Opfer 2012), much uncertainty remains around best practices for instruction, especially since measures that could be used to document effective practices are only just being developed. Such uncertainty leaves questions about how to respond to results from competency assessments. Lack of clarity on instructional approaches results for a variety of reasons, not least of which is that the mechanisms underlying many of these competencies remain murky. In particular, studies have yet to show whether certain skills are influenced more by factors inside or outside of school. Without that information, additional evidence is needed on how teachers can expect to influence those competencies. Creativity is an example. Some schools in the United States and Asia are actively measuring and teaching creativity, and research suggests that creativity can be taught (Ball, Pollard, and Stanley 2010; Shallcross 1981; Sternberg 2010). Nonetheless, uncertainty remains as to whether creativity is driven more by factors inside or outside the classroom, and what this lack of clarity means for classroom practice (Craft et al. 1997; Dudek 1974). Therefore, in many school settings, uncertainty remains about how best to use the data from measures of 21<sup>st</sup> century competencies to have the greatest possible influence on them.

**Assessments can have unintended consequences, which should be monitored in each local context.** The decision to adopt measures of 21<sup>st</sup> century competencies typically reflects a desire to promote attention to those competencies in schools and to improve students' mastery of them. Despite the potential benefits of adopting such measures, any test carries risks of unintended and undesirable consequences. For example, measuring a student's motivation or grit could help students understand their own strengths and weaknesses in these areas, thereby fostering improvement on these competencies. But if used in certain ways, these measures could lead to students' receiving stigmatizing labels that hinder their development. As we discussed in Chapter 4, the MSA safeguards against such potential unintended consequences by reporting only results aggregated at the school level. In taking this approach, students will not be labeled as unmotivated, low on time-management skills, or the like. A large body of research shows that these labels can have serious consequences for teachers and students, in part because they can generate self-fulfilling prophecies. Another unintended consequence might come in the form of an unanticipated or unwanted change to teaching practices as a response to the assessments. For example, as previously discussed, research shows that high-stakes tests can narrow the curriculum covered when teachers teach to the test. Many of these unintended consequences can be avoided through the careful design of testing and accountability policies (e.g., refraining from attaching stakes to measures that are too easily "taught to") and through monitoring of the administration process and the subsequent uses of test scores to identify instances in which results are misused.

**Measures of 21<sup>st</sup> century competencies should be part of a balanced assessment strategy.** While this report provides examples of innovative assessments measuring competencies that have traditionally not been the focus of measurement, adoption of such assessments should not come at the expense of other, more common assessments. For instance, we would not recommend abandoning achievement testing in favor of focusing purely on critical thinking. Using existing measures of achievement in mathematics, reading, science, and the like to ensure students are mastering core academic content remains important. As discussed in the previous chapter, these measures might also include administrative data on factors such as attendance or behavior, which should be integrated with more formal assessments of 21<sup>st</sup> century competencies. In short, the assessments discussed in this report are not meant to replace existing measures so much as supplement them with the goal of producing a more balanced, holistic system of assessment in schools and school systems.

## APPENDIX. CASE STUDIES OF CUTTING-EDGE MEASURES

### Alelo Language and Culture Simulations

Alelo, a company that spun off of a project at the University of Southern California, uses social simulation to teach foreign languages. Using an online format, students interact with an avatar—a digital, fully interactive replica of a human being—in whatever language is being learned. This approach allows students to learn through simulated interactions with native speakers at a skill-appropriate level. Even more importantly, the simulator is designed to be culturally specific. For example, a student interested in learning Spanish will not interact with a speaker from an arbitrarily chosen Spanish-speaking region. Instead, if the pupil will be going to, say, Argentina, then the avatar will use not only local pronunciations but also particular idioms. If a student speaks in a way that is not natural for a given country, the avatar will react accordingly. As a result, simulations give students opportunities to retry scenarios and thereby generate improved outcomes from their interactions. By focusing on how language and culture interact, Alelo’s simulation software attempts to enhance fluency, cultural awareness, communicational skill, and critical thinking. The measures built into the program are largely formative rather than summative.

#### **Practical Considerations**

Of all the assessments included in this report, Alelo’s simulation arguably goes farthest in blending curriculum and measurement. This approach is designed to mirror the real process by which individuals develop language fluency; this process typically involves interacting with people in the language and thereby receiving a constant assessment of their skill from the other person in the conversation. When a mistake is made, it is recognized immediately by the native speaker and often communicated, either intentionally or in the form of body language. Like these real-world scenarios, the simulation also assesses students in real time, albeit more overtly than might occur conversationally. For example, if the avatar suggests meeting over the weekend, but the student tries to schedule the appointment for a weekday, the avatar will react and, more concretely, the program will alert the student to the oversight with hints for correction.

#### **Technical Considerations**

Because the measures built into Alelo’s simulation are real-time and informal, traditional reliability and validity estimates are not especially feasible. Instead, much attention is focused on determining how learning in the simulation compares with learning in a classroom, especially for long-term fluency outcomes. The developers have conducted research on the product’s effectiveness by examining pre- and post-tests of language skill that rely on more standard measures of proficiency. For example, Alelo participated in a study designed to compare outcomes for students learning Danish in traditional classrooms with those primarily using the simulation. They found that these students could use the simulations for a comparable number of hours without any drop in the passage rates on a Danish language exam (Hansen 2013).

#### **Instructional Considerations**

Currently, Alelo is only just expanding into the K–12 and higher-education sectors. Therefore, very little information is available on its classroom uses. Nonetheless, these simulations have been used extensively by the military and have won a number of awards, including being a finalist in the Ed Tech Innovation Incubator Program. Studies on these military and government agency uses suggest that the interface is fairly straightforward (at least for students who are young adults or older) and academically beneficial (Johnson and Zaker 2012). For example, students using the Chinese language program gave it an average score of 4.24 out of a maximum possible score of 5 on a Likert scale measuring the speech-recognition



capability of the software, and only 2 percent of students gave this facet of the simulation a negative rating (Johnson and Zaker 2012). An additional study showed that the software can improve language learning and reduce costs by supplementing in-class instruction with additional hours of practice in the simulator (Hansen 2013). Alelo's research for Denmark's primary educational agency shows that, though expensive, the software can still generate overall savings by allowing more students to participate at flexible times, which reduces the need for in-class conversational time.

## Common Core State Standards Consortia Assessments

As many states begin to implement the Common Core State Standards (CCSS), the US Department of Education has given grants to two state consortia to develop assessments aligned to the new standards. These tests will presumably replace the ones currently being used to meet federal accountability requirements for participating states. Though the two funded consortia—Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (Smarter Balanced)—differ in important ways, they share a number of fundamental priorities, including an effort to link assessment to instruction, support for formative assessment, and test questions that tap 21<sup>st</sup> century competencies. Both sets of assessment, which are meant to provide comparability across states, are scheduled to be in use by 2015.

These tests are being designed to measure critical-thinking and communication skills in both mathematics and languages arts. To that end, both consortia hope to balance multiple-choice items with essay questions and performance assessments, including an end-of-the-year test of speaking and listening being developed by PARCC. In addition to changing the types of questions asked, both consortia aim to diversify the types of assessments used by including formative, diagnostic, and summative measures. For example, both PARCC and Smarter Balanced will provide what they are calling “interim” assessments designed to give teachers actionable information about whether a given student will be prepared for the summative exam used for accountability purposes. Some of these interim measures will ask students to demonstrate skills that are hard to measure, including planning, management of information, and critical thinking.

### Format and Scoring

Both the summative and formative assessments will use innovative formats, item types, and scoring routines to help achieve the consortia's objectives. Though the summative assessments will include many standard multiple-choice questions, they will also include items that give students more flexibility to develop their own responses. For example, math questions might allow students to graph functions, write equations, identify pieces of evidence to support a conclusion, or compose short answers. Beyond specific items, both consortia will administer summative assessments on the computer, one of which (Smarter Balanced) will adapt questions to the student's level of understanding as the test progresses. One benefit of the computer-based approach is that the scoring of many items can occur much faster; in some cases, test results will be available a week or two after administration. The interim assessments, meanwhile, will often use even more innovative approaches. To measure communication and critical thinking, these measures will include tasks in the form of an oral presentation, essay, product development, or the like.

### Overall Quality

To date, very little information exists on the quality of the CCSS assessments, in part because both are still in development. However, in order to be used for accountability purposes as intended, these tests will need to meet the highest of technical standards. To ensure this level of quality, both consortia have partnered with organizations expert in test development and will convene technical groups made up of measurement professionals. One benefit to the computer-adaptive approach being employed by Smarter Balanced is that the tests will likely do a better job of measuring performance for exceptionally high- or low-achieving students by ensuring that the questions a student sees are commensurate with his or her

mastery of the material. The consortia are also paying close attention to matters of validity. For example, Smarter Balanced plans to conduct validity studies of how well certain items and scores predict success in college or the workplace.

### **Classroom, School, and District Use**

Both PARCC and Smarter Balanced emphasize the importance of supporting teachers, and instruction more generally, through the design of the assessments and by providing a variety of pedagogic resources. These objectives are in line with the broader goals of the CCSS, which emphasize deeper learning of core subject matter through better-articulated learning progressions. Beyond developing a range of formative assessments, the tools available through PARCC and Smarter Balanced may include (1) an online interface for developing custom reports on schools or students, (2) measures of growth to track student progress toward college readiness, (3) banks of test questions for classroom use, (4) model lessons, (5) curricular frameworks, and (6) established cadres of educational leaders tasked with supporting districts as they implement the tests. All of these tools will also be organized into a warehouse of research-based supports and interventions to support students who are falling behind academically, including subgroups, such as English learners. Both consortia are also piloting evaluation tools that educators can use to provide feedback as implementation gets under way. To support these activities, the US Department of Education has awarded both groups add-on grants for transition supports.

Given that enhancing teaching is a primary goal of both consortia, detailing all of the supports available to educators and students is well beyond the scope of this report. For more information, please refer to the following resources:

- Smarter Balanced site for teachers: <http://www.smarterbalanced.org/>
- PARCC site on educational resources: <http://www.parcconline.org>
- ETS update on CCSS assessments: [http://www.k12center.org/rsc/pdf/Assessments\\_for\\_the\\_Common\\_Core\\_Standards\\_July\\_2011\\_Update.pdf](http://www.k12center.org/rsc/pdf/Assessments_for_the_Common_Core_Standards_July_2011_Update.pdf)
- ETS (Kyllonen) on 21<sup>st</sup> century skills measures:  
<http://www.k12center.org/rsc/pdf/session5-kyllonen-paper-tea2012.pdf>  
[http://www.usc.edu/programs/cerpp/docs/Kyllonen\\_21<sup>st</sup>\\_Cent\\_Skills\\_and\\_CCSS.pdf](http://www.usc.edu/programs/cerpp/docs/Kyllonen_21st_Cent_Skills_and_CCSS.pdf)

### **EcoMUVE**

Developed at Harvard, EcoMUVE stands for Ecological Multi-User Virtual Environment. Like some of the other innovative assessments discussed, EcoMUVE is a computer simulation designed to foster and assess multiple skills in tandem, including science knowledge, problem solving, collaboration, communication, and learning how to learn (see Figure 9). Like SimScientists, an assessment we discuss later, EcoMUVE is both formative and summative. The simulation is structured as a two-week course on ecology with ways to glean information on students during and at the end of the unit. Teachers are meant to use the assessment to respond to students' needs relative to all of these skill sets, as well as ensure that students have mastered basic knowledge of ecological science by the end of the unit. Though the test overtly focuses on content knowledge, its broader intent relates directly to fostering motivation. Research shows that students using EcoMUVE were more motivated by the scientific content than under normal instructional conditions and that as a direct result, their self-belief and motivation around scientific inquiry increased (Metcalf et al. 2011).

**FIGURE 9.**  
EcoMUVE Ecosystem Science Learning



Left: Screenshot of EcoMUVE pond ecosystem. Right: Example of information provided to students exploring the EcoMUVE environment.

Source: Metcalf, S., A. Kamarainen, M. S. Tutwiler, T. Grotzer, and C. Dede. (2011). "Ecosystem Science Learning via Multi-User Virtual Environments." *International Journal of Gaming and Computer-Mediated Simulations* 3(1): 86.

In terms of formatting, EcoMUVE creates a three-dimensional world that students can explore with unfettered access. Specifically, students explore a pond and the surrounding landscape—both natural and manmade—to investigate a given problem with the ecosystem, such as increased deaths among a fish species. The unrestricted access to the landscape means that students can decide what to measure (pH levels, for instance), talk to whomever they want, and visually inspect whatever might be contributing to the problem. Students are supposed to be broken into project teams then assigned a specific role within the team, such as chemist or biologist. These teams are meant to encourage group problem solving, including communication of fairly complex issues, with particular emphasis on cause and effect. The entire unit culminates in a letter written by each student to the local mayor explaining what likely causes the problem being examined. Though evidence within the scenario points to several probable causes, correctly diagnosing the issue is not the primary focus. Rather, students are rated on the quality of the argument made about their hypothesis. Test developers argue that this approach increases engagement and self-efficacy, especially among students who struggle in their science courses. The format is also meant to better resemble scientific inquiry in the real world, which typically affords no absolute solutions.

### Technical Considerations

Given the blending of assessment into curriculum that occurs under the EcoMUVE platform, researchers developing the test focus less on standard measures of reliability and validity. Instead they use external measures to show that using the software changes attitudes and understanding over time. For example, a 2011 study of EcoMUVE shows that students not only increase their understanding of ecosystem concepts such as abiotic factors and photosynthesis, but also develop a deeper understanding of these concepts (Grotzer et al. 2011). Similarly, another preliminary study shows that students are engaged by both the novel format and the scientific content, an engagement that translates into an enhanced desire among participants for self-driven learning (Metcalf et al. 2013). In short, researchers on the project focus less on making the scores on letters written by students at the end of the project replicable and

more on making sure that the process of exploration culminating in the letter-writing process consistently generates the shifts in knowledge and attitudes with which the project is concerned.

### **Instructional Considerations**

Research and anecdotal feedback from the EcoMUVE project suggest that the blending of curriculum and assessment can help support teachers in a variety of ways. According to EcoMUVE researchers, teachers often struggle to convey concepts in hands-on, engaging ways, given how time delays, spatial distance, nonobvious causes, and population-level effects influence problems of cause and effect. By contrast, the test designers argue that using a computer simulation makes time, space, and population influences easier to manipulate and therefore more overt. Though some teachers involved in piloting the simulation expressed concerns that students were responding to the novelty of the technology rather than the pedagogic benefits of the simulation, research at Harvard suggests that these fears are largely unwarranted (Metcalf et al. 2013). Both teachers and students alike reported decreasing fascination with the format of the unit and increasing engagement with the scenario over time.

### **The Graduation Performance System**

The GPS was developed collaboratively by Asia Society and SCALE as a portfolio used to measure student progress in a number of areas, with particular emphasis on global competence. In the GPS framework, global competence is broken down into constituent skills, including investigating the world, weighing perspectives, communicating ideas, taking action, and applying expertise within and across disciplines. More broadly, the GPS is intended to assess critical thinking and communication, among other skills. The GPS gives local practitioners a great deal of flexibility in terms of what the portfolios include, though Asia Society fosters consistency by providing standards for the portfolio content (referred to as a “graduate profile”), a series of discipline-based performance targets and rubrics, sample curricula, and examples of student work (such examples can be seen on Asia Society’s website: <http://asiasociety.org/pos>). All in all, the GPS goes beyond a typical portfolio system by providing guidance on rubrics, model design and implementation, assessment of student work, combining that work into a portfolio, and determining whether the final product meets standards to deem the student globally competent.

### **Practical Considerations**

The GPS is formatted as a portfolio system with embedded curriculum modules that suggest how core academic content can be reframed to better emphasize global competence. As such, a school or school system that adopts the GPS system would not necessarily be adopting new academic content standards (the GPS uses the CCSS as the basis for core academic material) so much as supplementing existing standards with new ones that help sharpen the focus of education on issues of global relevance. Learning to score the portfolios, however, involves a time commitment from teachers in order to overcome some of the reliability issues that can influence this mode of assessment—though, as discussed below, this training process is part of what might make the GPS valuable to teachers. The GPS also uses technology to make administration and scoring of portfolios as efficient and technically sound as possible. In particular, the GPS uses a digital platform that provides a repository for GPS materials, a “bank” of GPS modules and module development frameworks, a process for teachers to score student work electronically, and professional development modules to support teachers’ onsite and online professional learning—as well as an online learning community for teachers to share the learning modules and performance assessment tasks they have designed, and to receive feedback.

### **Technical Considerations**

Given that the GPS is fairly new, evidence of technical quality is still limited. However, a few important factors should be mentioned. First, the technical complexities associated with the GPS are much like those described in Chapter 4 for portfolios in general and include issues of inter-rater reliability.

Like other portfolio systems, the GPS attempts to address this issue by providing clear standards for performance and structured training for the educators doing the scoring. Second, much of Asia Society's ongoing work with SCALE around the GPS involves studying the effectiveness of this approach to generating reliable scores. In particular, SCALE's work focuses on establishing a validity argument for the assessment. Research at SCALE will seek evidence that the measure covers appropriate and relevant content related to global competency, and that the measure generates reliable scores. In terms of the latter, SCALE will examine inter-rater agreement and will conduct a study to determine what other sources of error may be influencing scores.

### **Instructional Considerations**

The GPS is built around instructional considerations, especially the value it will add to teaching and thereby the student's experience of the embedded curriculum modules. Schools implement the GPS through a collaborative model of professional development. Specifically, teachers are engaged in professional learning that is itself largely experiential and performance-based through a sequence of professional development modules. Each module engages teachers in learning activities that iterate between learning in small groups; trying out what they have learned in their classrooms; collaboratively reviewing the classroom experience with peers, including examination of student work; and subsequently applying what they have learned in their teaching. Teacher learning is facilitated through onsite and online coaching by a trained and certified staff member or consultant from the GPS project team. The GPS curriculum also uses the CCSS as the basis for core academic material, which means teachers in the United States can build GPS work into their ongoing adoption efforts, and schools in other countries can see how the United States is attempting to revamp its core standards.

### **Mission Skills Assessment**

The Mission Skills Assessment (MSA) is a packet of instruments being developed by ETS in conjunction with the Independent School Data Exchange (INDEX). Broadly, the purpose of these instruments is to measure the noncognitive skills that many independent schools value, both in classrooms and during the admissions process. In particular, the MSA allows schools to measure several skills, some of which we emphasize explicitly in this report: collaboration, creativity, ethics, resilience, intrinsic motivation, and learning to learn (especially time management). With the exception of growth mindset, the MSA assesses virtually the entire set of skills included in the intrapersonal category presented earlier in this paper. As discussed previously, the MSA is used primarily for summative purposes.

### **Practical Considerations**

Designers of the MSA attempt to overcome technical difficulties in measuring these skillsets by triangulating with different measures. That is, they measure the same skills using student self-reports, teacher observations, situational judgment tests, and the like. By using multiple instruments to assess the same construct, they can better disentangle sources of error and, thereby, increase the precision of the measurement. Though the measures for each skill differ, they generally include the following components each year: 60 minutes of student self-assessment followed by situational judgment tests and other performance measures, teacher ratings of students, and outcome data, such as grades. For example, resilience is measured by asking students and teachers to rate the child's ability to overcome setbacks, and recording student multiple-choice responses to a hypothetical situation in which he or she has too much homework or is faced with another stressful situation.

### **Technical Considerations**

Initial evidence suggests the MSA at least meets minimum technical standards (INDEX 2013). The reliability of scores on the combined assessments for each skill, as measured by both internal consistency and test-retest reliability, is high (in fact, the reliability of scores for measures of some skills is on par

with similar estimates for the SAT). Further, researchers at ETS have shown that MSA measures predict not only academic outcomes, but also the student's overall well-being as assessed by self-reports of life satisfaction (INDEX 2013). Life satisfaction was measured with the Students' Life Satisfaction Scale (Huebner 1991), which includes responses to statements about how well the respondent's life is going and whether the respondent has what he or she wants in life. Many MSA measures also do a better job of predicting both academic and nonacademic outcomes than do scores on standardized academic achievement tests (INDEX 2013). Resilience, for instance, has a much higher correlation with well-being than mathematics and reading scores, as does intrinsic motivation with teacher reports of student quality.

### **Instructional Considerations**

The teachers we consulted for this report stated that schools using the MSA can be quite valuable strategically and pedagogically, especially because of the assessment's ability to focus conversations around desired outcomes. According to educators involved in the project, MSA results provide several advantages compared with the anecdotal evidence that teachers and administrators used previously to assess these skills. First, making data more systematic has generated not only more conversation around these skills, but also increasingly fine-grained discussion. For example, teachers can now disaggregate the relationship between specific skills and outcomes of interest, which allows for more targeted support when students appear to be falling behind. Second, data establish more transparency and a self-imposed accountability. Now schools know that they are successful in promoting the development of particular skills. Conversely, the data also force educators to confront shortcomings more immediately and directly. For instance, a school cannot easily ignore concrete data suggesting that students are behind peers in other schools on time management, curiosity, or the like. Finally, the first two advantages—better discussion and self-imposed accountability—help administrators and teachers support each other in achieving the school's mission. For example, New Canaan Country School has developed a teacher network among study schools to share curriculum designed to foster the skills measured by the MSA (Secondary School Admission Test Board 2013). Administrators, meanwhile, use feedback from these groups to provide related professional development (Secondary School Admission Test Board 2013).

### **PISA Collaborative Problem Solving**

In 2015, the Programme for International Student Assessment (PISA) will launch a new test devoted to measuring collaborative problem solving (CPS). Educational experts working at and supporting PISA define CPS as “the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution” (OECD 2013, 6). Toward that end, the test will include measures of collaboration, problem solving, and how the two interact with each other to generate a desired outcome. More specifically, the construct being measured incorporates three specific competencies: establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining team organization. Each student tested will receive multiple questions designed to measure all three competencies. Unlike many other innovative tests we discuss, the assessment is intended to be summative. While educators certainly are not restricted from using results to support students, several barriers to such a use exist, including the facts that only a small sample of students are tested and that the technological requirements of administering this computer-based simulation are not insignificant. As a summative assessment, the test makers are also concerned with protecting against cheating, which means that questions will not likely be made widely available.

### **Practical Considerations**

The PISA CPS test is entirely computer-based. In fact, the handful of countries that will not have switched to the computer version of the general PISA will be ineligible to administer the test. The main reason for using the computer is to overcome a technical problem with giving a test involving communication

between two people; namely, that the abilities and skills of one will influence the outcome for the other, a major threat to reliability and validity. To avoid this problem, test makers at PISA—in collaboration with ETS—have designed an assessment in which the person being tested interacts with a simulated collaborator. In so doing, the test knows the quality of a given student’s partner and can vary that quality from one set of questions to the next. At least in theory, this approach makes the test fairer, because assignment of a student’s teammates is not left to chance.

What does this simulated interaction look like, exactly? According to a PISA draft CPS framework (OECD 2013), each student will be assigned a two-hour test form, an hour of which could be devoted to CPS (not all students taking the PISA will receive a CPS assessment—only a subset is randomly assigned to do so). Within CPS, units will range from five- to twenty-minute collaborative interactions around a particular problem scenario. For each unit, multiple measurements of communications, actions, products, and responses to probes will be recorded. Each of these individual questions (five to thirty per unit) will provide a score for one of the three CPS competencies. Some of these questions will be multiple choice, while others will be open response. During a scenario involving factory efficiency, for instance, the student both responds to multiple-choice questions about what tasks his or her partner should perform and writes an email to the person in charge of the factory about what must happen next. The computer-based test is adaptive in several ways. For one, it does not provide more questions per unit than is needed to produce a satisfactory score. Second, the computer will assess how the student responds to his or her virtual partner, then tailor what the virtual partner does next given that response. The simulation is designed to help ensure that the problem-solving process does not stall, at least to the point where an accurate measurement of the student’s abilities cannot be taken. The prompts given by the computer are designed to maintain a balance between success and challenge. Specifically, prompts may include questions like the following:

- What information does your partner have or need?
- Why is your partner not providing information to other group members?
- What tasks will or should your partner do next?

By combining these sorts of multiple-choice questions with open-ended responses to the situation, the assessment is meant to provide a reliable estimate of the three competencies that make up collaborative problem solving. Clearly, the tests are quite complicated and cannot be fully described here. For more details, including example scenarios, please see the draft CPS framework at: <http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf>.

### **Technical Considerations**

To date, the CPS assessment is very much in a design phase. As a result, little information exists publicly on the reliability and the validity of the test scores. According to researchers at ETS and OECD representatives, the assessment will be field-tested in the coming year.

### **Instructional Considerations**

The assessment’s pilot status means few educators have had a chance to interact with the test. Nonetheless, there are several facts about the administration process that could be valuable as educators prepare for the 2015 PISA administration:

- For the majority of countries that have committed to doing the PISA by computer in 2015, there will be no additional technological requirements for the CPS.
- Technological support, such as laptops for administration, is often provided to countries, though this varies by region.

- PISA employs a somewhat complicated sampling strategy: not all schools take the PISA, nor do all students within a school take the PISA, nor will those selected students all be administered the CPS portion of the exam.

## Queensland Performance Assessments

The Queensland Performance Assessments (QPAs) measure academic knowledge, as well as problem solving, communication, and learning how to learn, among others. In response to Australia's high-stakes university entrance exams of the 1970s, which were often deemed unrealistically difficult in the sciences, Queensland developed its own externally moderated school-based assessment system. Rather than rely on information about a student from a single point in time, the new assessment system is built on a purposeful, systematic, and ongoing collection of data on student learning. At heart, the system is designed to create a tighter link between the goals of instruction and testing. To achieve this goal, teachers develop the tests—even those used for high-stakes decisions—based on national standards and with support from psychometric experts at the Queensland Studies Authority (QSA). Teachers also meet across schools in an attempt to ensure that standards are consistent through a process of negotiation, especially when it comes to making test-based decisions about student proficiency. According to the QSA, in addition to benefits for students, the approach promotes teacher professionalism (Queensland Studies Authority 2010).

### Practical Considerations

The primary feature of the QPAs is that they are extremely loose on format and tight on scoring. Teachers can develop a test in any format they want, so long as the standards used to determine proficiency are not only clear but also comparable across schools. For instance, one school could use a multiple-choice history test, while a neighbor school uses an essay and oral presentation to measure the same standard, so long as the teachers (and the QSA) agreed on how proficiency was determined in each case. While this example presents quite different formats, such discrepancies are not necessarily the norm, a fact driven in part by item banks that schools can draw from in developing their tests. Though describing the process by which Queensland attempts to generate comparability is too complicated to describe in this paper, it generally begins with QSA-developed standards and curricula, then involves consistent back and forth between schools and review boards made up of teachers from across the province to negotiate a consistent set of student classifications for a given test.

### Technical Considerations

Though one might expect reliability to be a major problem under this testing framework, evidence suggests the contrary. In a study conducted over more than a decade by the QSA, consistency across raters of student work has been shown to be quite high, in some cases exceeding industry standards set for more highly standardized tests (Queensland Studies Authority 2010). A study conducted independently of the QSA found similar results (Masters and McBryde 1994). (These results do not, however, rule out sources of measurement error due to factors other than raters, on which there appears to be little information available.) Despite these promising findings, tying the assessments to outcomes of interest becomes more difficult. Though the QSA can say that being rated at a particular proficiency level predicts long-term academic and professional outcomes, it is less clear that a given local assessment predicts these outcomes.

### Instructional Considerations

The QPAs appear to involve a series of tradeoffs when it comes to implications for teachers. On one hand, according to the QSA, there is evidence of a tighter link between instruction and assessment, and educators are much more empowered in the testing system (Queensland Studies Authority 2010). In particular, the QSA reports that advantages of the system include improved teacher professionalism,

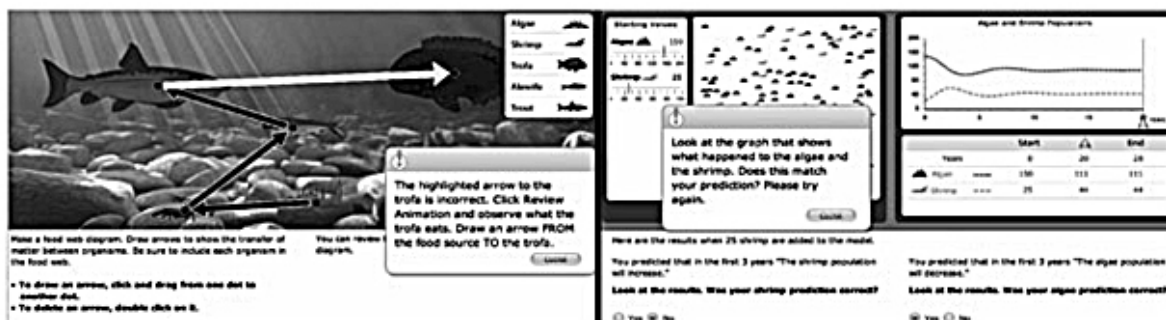


responsiveness to different learning styles, and connection between assessment and both academic and nonacademic skills, including higher-order critical-thinking skills. On the other hand, the system requires significant resources. Although the cost of test development is lower than when a major testing company is hired, Queensland nonetheless must invest in the complex process of “moderation” by which consistency is studied (Queensland Studies Authority 2010). Further, the system relies on teachers devoting a significant portion of their time not only to test development, but also to serving on panels designed to ensure cross-school comparability.

## SimScientists

In many ways, SimScientists assessments, which are developed by WestEd in San Francisco, overlap in content and purpose with EcoMUVE. For example, both assessments include formative and summative components, emphasize ecosystem components of standard science curriculum, and use computer-based simulations to elicit more complex responses from students. Like EcoMUVE, SimScientists also emphasizes a range of cognitive, interpersonal, and intrapersonal skills. However, the tests also differ in fundamental ways. Most importantly, the assessments built into SimScientists are formal and designed to be replicable; in fact, researchers at WestEd intend for the tests to be used not only at the end of the unit, but also as a component of city or state accountability systems. Therefore, as GCEN members begin to think about how innovative assessments fit into the broader accountability structure, SimScientists presents a glimpse of what the future may hold (see Figure 10).

**FIGURE 10.**  
SimScientists Science Simulation



Quellmalz, E. S., M. J. Timms, M. D. Stilbergitt, and B. C. Buckley. (2012). “Science Assessments for All Integrating Science Simulations into Balanced State Science Assessment Systems.” *Journal of Research in Science Teaching* 49(3): 363–93

### Practical Considerations

Though SimScientists is designed to provide formative information useful to teaching and learning, the assessment suite does not blend in its own curriculum, as EcoMUVE does. Rather, each of the two units—ecosystems and force and motion—build in formative assessments that teachers can administer when they feel students have mastered the necessary prerequisites. In these formative assessments, students are given a scenario to examine. Though they have a great deal of latitude to examine the environment, they are not provided free range as in EcoMUVE, a decision made to help ensure the reliability of assessment results. During the simulation, students complete tasks such as making observations, running trials in an experiment, interpreting data, making predictions, and explaining results. Responses to these tasks are recorded in a variety of formats, including multiple choice, changing the values in a simulation, drawing arrows to represent interactions in the system, and typing explanations to open-ended questions. For all questions except those involving open response, students are then provided graduated levels of coaching in the form of prompts based on their needs. As an example, students are given opportunities to fix incorrect answers based on the feedback provided.

Finally, toward the end of the unit, teachers decide when to administer the summative assessment. The final test consists of items much like those in the formative assessments but are translated into a different setting. For instance, if a student answered questions about a lake ecosystem in a formative assessment, the summative test might use a forest ecosystem. Teachers receive training, either online or in person, on how to score the open-ended responses for the end-of-unit exam. The report generated online for a given student following the summative assessment classifies an individual's proficiency as below basic, basic, proficient, or advanced. Further, these classifications are given for both content and inquiry proficiencies, which might help teachers understand whether students struggle with the core knowledge, critical-thinking skills, or both. Beyond the individual student, the report generates classroom-level data on content and inquiry mastery.

### **Technical Considerations**

Emergent research suggests that SimScientists meets technical benchmarks (Quellmalz et al. 2012). Estimates of reliability meet industry standards. In terms of validity, researchers at WestEd studied students and teachers using the program to help establish criterion evidence of validity. They found that results on the SimScientists summative test correlate highly with those on independent tests measuring the same content standards (Quellmalz et al. 2012). The study also found that English learners and special-education students perform much better on the SimScientists tests than the external tests, and that this increased subgroup performance appears to be attributable to the SimScientists' presentation of the testing material in multiple formats, including written, oral, and visual. In total, the technical quality of SimScientists is still being confirmed but appears high.

### **Instructional Considerations**

As part of its validation work, WestEd conducted feasibility studies to determine whether use of these tests in classrooms and schools is practicable (Quellmalz et al. 2012). WestEd found that after a short adjustment period, teachers were able to implement the assessments without much difficulty. More importantly, results suggest the assessments yielded educational benefits. For instance, teachers reported that the simulation-based tests were an improvement on their homegrown predecessors because SimScientists provided instant feedback, interacted with students to increase learning during the formative tests, and presented helpful visuals (Quellmalz et al. 2012). Results also show that students were highly engaged by the assessments and able to complete them successfully (Quellmalz et al. 2012). Overall, the studies only presented one practical drawback: the need for computers to be easily accessible in order to administer the assessments multiple times in a given school year.

### **Singapore Project Work**

The Singapore A-levels—tests required of all pre-university students—now require completion of a group project. This project work is meant to complement other A-level requirements, which include a formal paper and “mother tongue” assessment. Specifically, the group project measures application of core academic content, communication, collaboration, and learning to learn. This last 21<sup>st</sup> century competency in particular involves learning independently, reflecting on learning, and taking appropriate action to improve. Results from the A-levels, including the project component, help determine university admission within the country.

### **Practical Considerations**

Students are placed into groups by their teacher then are free to select a topic of their choosing. Past topics have included natural forces, momentum, tradition, groundbreaking individuals, and entertainment. Once students select a project, they work for several weeks preparing for the three associated requirements: a written report, an oral presentation, and a group project file. While the written report and oral presentation are largely self-explanatory, the project file is Singapore's approach to assessing the student's

skills in learning how to learn. Essentially, the file represents the student's way to track progress over time and to reflect on challenges and successes. In particular, students analyze three specific artifacts of their choosing from the project, artifacts that elucidate the thinking behind its design. For example, one might choose an early document outlining the argument that will underlie the paper. From there, the student could discuss how the ideas were formulated, or how they evolved over the course of work. Scoring of these requirements is conducted entirely by local teachers. Each requirement receives a specific weight in the final grade, which is assigned to only the group for the paper, only the student for the file, and both for the presentation.

### **Technical Considerations**

Because these projects are still in a fairly early stage, little technical information is available. However, Singapore is approaching reliability much like test designers in Queensland. The Ministry of Education (MOE) controls the assessment requirements, conditions, standards, and grading process. Standards in particular are key to the strategy of generating consistency. By being explicit about how to tell whether a student has met standards in core academic content, communication, collaboration, and learning to learn, Singapore is attempting to ensure agreement among raters. To enhance this consistency, teachers who will be scoring the projects receive training from the MOE on the standards and effective grading practices. As a final check on reliability, the MOE trains internal moderators and provides external moderation where necessary. Thus far, little available research shows whether the group projects predict outcomes of interest, including use of tested skills in a postsecondary setting.

### **Instructional Considerations**

The project portion of the A-levels is designed specifically to be useful and seamless for teachers. In addition to having teachers play a direct role in grading the assessment, the projects are intentionally integrated into regular instructional time. This approach is meant to ensure that the project is incorporated into the curriculum, with the intent of reducing time taken from core instruction. Moreover, inserting the assessment into class time is intended to build in a formative component, because teachers can spot areas of difficulty and provide guidance when students struggle with subject-related material (that is, teachers cannot directly influence the product but can help students when they have troubles with core content). Another potential benefit is the high degree of autonomy granted to students. Initial teacher reports suggest that, as a result of this independence, students are given an opportunity to focus on self-directed inquiry (Tan 2013). Given the ability to select the topic, divvy up work among team members, set a timeline, and formulate the final presentation, teachers report that they see the assessment as a tool that is useful for emphasizing skills associated with learning to learn.

## **World Savvy Challenge**

World Savvy Challenge is an international student competition designed to encourage and assess global awareness, critical thinking, and communication. The assessment takes the form of a project-based competition in which students develop a plan for addressing a policy area of international importance. For example, the 2013 theme is "sustainable communities," which means that students will research issues that could pose a threat to sustainability, such as global warming, poor governance, or threats to potable water. The competition centers on this theme and involves a number of stages. First, a classroom or school enrolls in the early fall then begins to develop understanding of the topic through initial project work, recommended field trips, and participation in the World Savvy Scavenger Hunt. As part of this initial phase, students develop a Knowledge-to-Action plan, which details how the students themselves can play a role in finding a solution to the identified problem. These activities take place either as part of the regular school day or after school, depending on the teacher's preference. Second, in the spring, students may compete in regional competitions (either in person or online for international groups) to determine which students have come up with the most promising solution. Finally, winners of the regional competition are invited to participate in the national competition.

### **Practical Considerations**

The competition is the main form of assessment in the World Savvy Challenge. Students attending the regional competition in person present their policy solutions as either a performance or a showcase. The former might include a skit, simulation, or some other creative format. By contrast, a showcase is a visual presentation of the same findings, such as a PowerPoint presentation, poster, or sculpture. Students competing online, meanwhile, showcase their work as either a documentary or website. Regardless of whether they are participating live or via the Internet, students must submit their Knowledge-to-Action plan in the form of a written strategic plan and take part in an action roundtable with peers. This roundtable involves a collaborative discussion with students from a variety of different educational institutions about available policy options and how effective they are likely to be. The action roundtable is meant to provide students with an opportunity to broaden their perspective on the issue and gain feedback on their own proposals. All three elements of the competition—presentation, action plan, and roundtable participation—are scored by judges, either from or trained by World Savvy.

### **Technical Considerations**

Currently, very little information exists on the reliability of the rating system used by World Savvy to judge student projects, though the organization appears to address consistency in part by having raters trained by the organization. However, the lack of information on reliability may result because of the formative nature of the measure. That is, participation is, according to the website, as important as winning. Regardless of whether students go on to the national competition, they will have done all the prework in their classrooms and competed in the regional competition, allowing them to interact with many peers interested in the same issues. Additionally, teachers can use the project work as an informal assessment of global awareness and critical thinking. These informal formative assessments are facilitated in part by the materials that teachers receive when they participate in the program.

### **Instructional Considerations**

Though teachers have significant latitude to shape the experience of their students in the World Savvy program, they also receive a number of supports in return for participation. These tools include a curriculum guide, professional development, and individual consulting with program staff, including ongoing curriculum support (the availability of each item is dependent on the package purchased). Professional development in particular is a major focus of the program. Developed jointly by World Savvy, Asia Society, and the Teachers College at Columbia University, professional development opportunities include workshops, institutes, and certification programs. Further, the organization is in the process of expanding these opportunities. For example, the Global Competency Certificate Program will allow teachers to earn either a certificate or master's degree in global-awareness instruction. Workshops and institutes, meanwhile, allow teachers to learn some of the content that will be presented to certificate program participants, but in smaller chunks. Teachers that attend these events gain insight into incorporating global awareness into regular classroom instruction, as well as a number of tools, such as a hundred-page curriculum guide, an online database of region-specific lessons and field trips, and an invitation to join the Global Educators Forum, an online community devoted to sharing best practices. Overall, participation in the program requires roughly twenty-five hours of class time, not including the competitions themselves. As previously mentioned, these hours can be incorporated into regular class time or offered as an afterschool program. Generally, the technological know-how required to participate is minimal. For instance, even students participating online who choose to develop a website will do so using a well-established template rather than coding the site themselves (unless they prefer to do so). Students can be registered by the teacher or school. Costs range from \$150 to \$1,000 per classroom entered, dependent on the level of professional development, staff support, and number of teams.

## REFERENCES

- Abedi, J. (2002). "Standardized Achievement Tests and English Language Learners: Psychometrics Issues." *Educational Assessment* 8(3): 231–57.
- . (2006a). "Language Issues in Item Development. In *Handbook of Test Development*, 377–98), edited by S. M. Downing and T. M. Haladyna. Mahwah, NJ: Erlbaum.
- . (2006b). Psychometric Issues in the ELL Assessment and Special Education Eligibility. *Teachers College Record* 108(11): 2282–303.
- Ackerman, T. A., and P. L. Smith. (1988). "A Comparison of the Information Provided by Essay, Multiple-Choice, and Free-Response Writing Tests." *Applied Psychological Measurement* 12(2): 117–28.
- Almeida, R. (2009). *Does the Workforce in East Asia Have the Right Skills? Evidence from Firm Level Surveys*. Mimeo. Washington, DC: World Bank.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Archibugi, D., and B. Lundvall. (2002). *The Globalizing Learning Economy*. Oxford University Press. Retrieved from <http://econpapers.repec.org/bookchap/oxpobooks/9780199258178.htm>.
- Bachen, C. M., P. F. Hernández-Ramos, and C. Raphael. (2012). "Simulating REAL LIVES Promoting Global Empathy and Interest in Learning through Simulation Games." *Simulation and Gaming* 43(4): 437–60.
- Ball, L., E. Pollard, and N. Stanley. (2010). "Creative Graduates, Creative Futures." CGCF Higher Education. Retrieved from [http://homepages.lboro.ac.uk/~adjjw/CreativeGraduatesCreativeFutures\\_ResearchReport\\_01.10.pdf](http://homepages.lboro.ac.uk/~adjjw/CreativeGraduatesCreativeFutures_ResearchReport_01.10.pdf)
- Barca-Lozano, A., L. S. Almeida, A. Ma Porto-Rioboo, M. Peralbo-Uzquiano, and J. C. Brenlla-Blanco. (2012). "School Motivation and Achievement: The Impact of Academic Goals, Learning Strategies and Self-Efficiency." *Anales de Psicología* 28(3): 848–59.
- Bennett, R. E., and D. A. Rock. (1995). "Generalizability, Validity, and Examinee Perceptions of a Computer-Delivered Formulating Hypotheses Test." *Journal of Educational Measurement* 32(1): 19–36.
- Berman, E., C.-Y. Wang, C.-A. Chen, X. Wang, N. Lovrich, C. Jan, . . . J. T. Sonco. (2013). "Public Executive Leadership in East and West: An Examination of HRM Factors in Eight Countries." *Review of Public Personnel Administration* 33(2): 164–84.
- Black, P., C. Harrison, C. Lee, B. Marshall, and D. William. (2003). *Assessment for Learning: Putting It into Practice*. Open University Press. Retrieved from <http://oro.open.ac.uk/24157/>
- Boix-Mansilla, V., and A. Jackson. (2011). *Educating for Global Competence: Preparing Our Students to Engage the World*. New York: Asia Society and the Council of Chief State School Officers. Retrieved from <http://dpi.state.wi.us/cal/pdf/book-globalcompetence.pdf>.

Carlson, B., and Z. Chen. (2013). "SATs Have Nothing on China's Dreaded Gaokao Exam." Salon.com, June 7, 2013. Retrieved from [http://www.salon.com/2013/06/07/sats\\_have\\_nothing\\_on\\_chinas\\_dreaded\\_gaokao\\_exam\\_partner/](http://www.salon.com/2013/06/07/sats_have_nothing_on_chinas_dreaded_gaokao_exam_partner/).

Carnoy, M., R. Elmore, and L. Siskin. (2013). *The New Accountability: High Schools and High-Stakes Testing*. Oxford, UK: Routledge.

Center on Education Policy. (2006). *From the Capital to the Classroom: Year 4 of the No Child Left Behind Act*. Washington, DC: Author.

Childers, T. L. (1986). "Assessment of the Psychometric Properties of an Opinion Leadership Scale." *Journal of Marketing Research* 23(2): 184–88.

Conley, D. T. (2005). "College Knowledge: Getting In is Only Half the Battle." *Principal Leadership* 6(1): 16–21.

———. (2008). "Rethinking College Readiness." *New Directions for Higher Education* 144: 3–13.

———. (2011). "Crosswalk Analysis of Deeper Learning Skills to Common Core State Standards." Educational Policy Improvement Center. Retrieved from <http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED537878>.

Craft, A., J. Dugal, G. Dyer, B. Jeffrey, and T. Lyons. (1997). *Can You Teach Creativity?* Nottingham, UK: Education Now. Retrieved from <http://oro.open.ac.uk/20308/>.

Deci, E. L., and R. M. Ryan. (2012). "Overview of Self-Determination Theory." In *The Oxford Handbook of Human Motivation*, edited by R. Ryan. Oxford, UK: Oxford University Press.

Deci, E. L., R. J. Vallerand, L. G. Pelletier, and R. M. Ryan. (1991). "Motivation and Education: The Self-Determination Perspective." *Educational Psychologist* 26(3–4): 325–46.

Di Addario, S., and D. Vuri. (2010). "Entrepreneurship and Market Size. The Case of Young College Graduates in Italy." *Labour Economics* 17(5): 848–58.

Di Gropello, E. (2011). *Putting higher education to work: skills and research for growth in East Asia*. World Bank, free PDF retrieved from <http://bit.ly/1ja5YKA>.

Duckworth, A. L., and P. D. Quinn. (2009). "Development and Validation of the Short Grit Scale (GRIT-S)." *Journal of Personality Assessment* 91(2): 166–74.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly. (2007). "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92(6): 1087.

Duckworth, A. L., P. D. Quinn, and M. E. Seligman. (2009). Positive Predictors of Teacher Effectiveness. *Journal of Positive Psychology* 4(6): 540–7.

Dudek, S. Z. (1974). "Creativity in Young Children—Attitude or Ability?" *Journal of Creative Behavior* 8(4): 282–92.

Dweck, C. (2006). *Mindset: The New Psychology of Success*. New York: Random House.

- . (2007). “The Perils and Promises of Praise.” *Educational Leadership* 65(2): 34–39.
- . (2008). *Mindsets and Math/Science Achievement*. New York: Carnegie Corporation of New York, Institute for Advanced Study, Commission on Mathematics and Science Education.
- . (2009). “Who Will the 21st Century Learners Be?” *Knowledge Quest* 38(2): 8–9.
- . (2010). “Even Geniuses Work Hard.” *Educational Leadership* 68(1): 16–20.
- El-Murad, J., and D. C. West. (2004). “The Definition and Measurement of Creativity: What Do We Know?” *Journal of Advertising Research* 44(2): 188–201.
- Facione, P. A. (1998). *Critical Thinking: What It Is and Why It Counts*. Millbrae, CA: California Academic Press.
- Facione, P. A., C. A. Sánchez, N. C. Facione, and J. Gainen. (1995). “The Disposition toward Critical Thinking.” *Journal of General Education* 44(1): 1–25.
- Faxon-Mills, S., L. S. Hamilton, M. Rudnick, B. M. and Stecher. (2013). *New Assessments, Better Instruction? Designing Assessment Systems to Promote Instructional Improvement*. Santa Monica, CA: RAND Corporation.
- Gao, X., R. J. Shavelson, and G. P. Baxter. (1994). “Generalizability of Large-Scale Performance Assessments in Science: Promises and Problems.” *Applied Measurement in Education* 7(4): 323–42.
- Gardner, H., and V. Boix-Mansilla. (1994). “Teaching for Understanding—within and across Disciplines.” *Educational Leadership* 51(5): 14–18.
- Goslin, D. A. (2003). *Engaging Minds: Motivation and Learning in America’s Schools*. Lanham, MD: Scarecrow Press.
- Grotzer, T. A., and B. B. Basca. (2003). “How Does Grasping the Underlying Causal Structures of Ecosystems Impact Students’ Understanding?” *Journal of Biological Education* 38(1): 16–29.
- Grotzer, T. A., M. S. Tutwiler, C. Dede, A. Kamarainen, and S. Metcalf. (April 2011). “Helping Students Learn More Expert Framing of Complex Causal Dynamics in Ecosystems Using EcoMUVE.” In *National Association of Research in Science Teaching Conference*, vol. 4. <http://ecomuve.gse.harvard.edu/publications/EcoMUVENARSTFull%20Paper7.2011.pdf>
- Guay, F., C. F. Ratelle, A. Roy, and D. Litalien. (2010). “Academic Self-Concept, Autonomous Academic Motivation, and Academic Achievement: Mediating and Additive Effects.” *Learning and Individual Differences* 20(6): 644–53.
- Haertel, E. H. (1999). “Performance Assessment and Educational Reform.” *Phi Delta Kappan* 80(9): 662–6.
- Haladyna, T. M., and S. M. Downing. (2004). “Construct-Irrelevant Variance in High-Stakes Testing.” *Educational Measurement: Issues and Practice* 23(1): 17–27.

- Hamilton, L. S. (2003). "Assessment as a Policy Tool." *Review of Research in Education* 27: 25–68.
- Hamilton, L. S., B. M. Stecher, and K. Yuan. (2012). "Standards-Based Accountability in the United States: Lessons Learned and Future Directions." *Education Inquiry* 3(2): 149–70.
- Hamilton, L. S., B. M. Stecher, J. A. Marsh, J. S. McCombs, and A. Robyn. (2007). *Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States*. Santa Monica, CA: RAND Corporation.
- Hansen, T. K. (2013). "The Danish Simulator: Exploring the Cost-Cutting Potential of Computer Games in Language Learning." [http://conference.pixel-online.net/ICT4LL2013/common/download/Paper\\_pdf/056-ITL11-FP-Hansen-ICT2013.pdf](http://conference.pixel-online.net/ICT4LL2013/common/download/Paper_pdf/056-ITL11-FP-Hansen-ICT2013.pdf).
- Heritage, M. (2010). "Formative Assessment and Next-Generation Assessment Systems: Are We Losing an Opportunity." National Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the Council of Chief State School Officers (CCSSO). Washington, DC: CCSSO. Retrieved from [http://www.sde.idaho.gov/site/formativeInterim/docs/Formative\\_Assessment\\_Next\\_Generation\\_2010.pdf](http://www.sde.idaho.gov/site/formativeInterim/docs/Formative_Assessment_Next_Generation_2010.pdf).
- Herman, J. L., E. Osmundson, and D. Silver. (2010). "Capturing Quality in Formative Assessment Practice: Measurement Challenges (CRESST Report)." Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.
- Huebner, E. S. (1991). "Initial Development of the Student's Life Satisfaction Scale." *School Psychology International* 12(3): 231–40.
- INDEX Schools. (March 2013). "Mission Skills Assessment." Presented at the annual NAIS Conference, Philadelphia, PA. <http://annualconference.nais.org/sites/2013/Pages/default.aspx>
- Jackson, P. W., and S. Messick. (1965). "The Person, the Product, and the Response: Conceptual Problems in the Assessment of Creativity." *Journal of Personality* 33(3): 309–29.
- Johnson, W. L., and S. B. Zaker. (2012). "The Power of Social Simulation for Chinese Language Teaching." *Proceedings of the 7th International Conference and Workshops on Technology and Chinese Language Teaching in the 21st Century*. [http://www.tacticallanguage.com/files/TCLT7\\_Presentation\\_Johnson\\_Zakar\\_May2012.pdf](http://www.tacticallanguage.com/files/TCLT7_Presentation_Johnson_Zakar_May2012.pdf).
- Jussim, L., J. Eccles, and S. Madon. (1996). "Social Perception, Social Stereotypes, and Teacher Expectations: Accuracy and the Quest for the Powerful Self-Fulfilling Prophecy." *Advances in Experimental Social Psychology* 28: 281–388.
- Kamens, D. H., and C. L. McNeely. (2010). "Globalization and the Growth of International Educational Testing and National Assessment." *Comparative Education Review* 54(1): 5–25.
- Kane, M. T. (2001). "Current Concerns in Validity Theory." *Journal of Educational Measurement* 38(4): 319–42.
- . (2006). "Validation." In *Educational Measurement*, edited by R. Brennan, 4th ed., 17–64. Westport, CT: American Council on Education/Praeger.



- . (2012). “Validating Score Interpretations and Uses.” *Language Testing* 29(1): 3–17.
- . (2013). “Validation as a Pragmatic, Scientific Activity.” *Journal of Educational Measurement* 50(1): 115–22.
- Kell, M. (2010). “International Testing: Measuring Global Standards or Reinforcing Inequalities.” Retrieved from <http://repository.ied.edu.hk/dspace/handle/2260.2/11902>.
- Klein, S. P., J. Jovanovic, B. M. Stecher, D. McCaffrey, R. J. Shavelson, E. Haertel ... K. Comfort. (1997). “Gender and Racial/Ethnic Differences on Performance Assessments in Science.” *Educational Evaluation and Policy Analysis* 19(2): 83–97.
- Koenig, J. A. (2011). *Assessing 21st Century Skills: Summary of a Workshop*. Washington, DC: National Academies Press.
- Koretz, D. (1998). “Large-Scale Portfolio Assessments in the U.S.: Evidence Pertaining to the Quality of Measurement.” *Assessment in Education* 5(3): 309–34.
- . (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press.
- Koretz, D., R. L. Linn, S. B. Dunbar, and L. A. Shepard. (1991). “The Effect of High-Stakes Testing on Achievement: Preliminary Findings about Generalization across Tests.” U.S. Department of Education, Office of Educational Research and Improvement, Educational Resources Information Center. Retrieved from <http://www.colorado.edu/UCB/AcademicAffairs/education/faculty/lorrieshepard/PDF/High%20Stakes%20Testing.pdf>.
- Koretz, D., B. Stecher, S. Klein, and D. McCaffrey. (1994). “The Vermont Portfolio Assessment Program: Findings and implications.” *Educational Measurement: Issues and Practice* 13(3): 5–16.
- Kyllonen, P. C. (2008). *The Research behind the ETS® Personal Potential Index (PPI)*. Princeton, NJ: ETS. Retrieved from [http://www.ets.org/Media/Products/PPI/10411\\_PPI\\_bkgrd\\_report\\_RD4.pdf](http://www.ets.org/Media/Products/PPI/10411_PPI_bkgrd_report_RD4.pdf).
- . (2012). “Measurement of 21st Century Competencies within the Common Core State Standards.” Retrieved from [http://www.usc.edu/programs/cerpp/docs/Kyllonen\\_21st\\_Cent\\_Skills\\_and\\_CCSS.pdf](http://www.usc.edu/programs/cerpp/docs/Kyllonen_21st_Cent_Skills_and_CCSS.pdf).
- Lai, E. R. (2011). “Collaboration: A Literature Review.” Pearson. Retrieved from <http://ed.pearsonassessments.com/hai/images/tmrs/Collaboration-Review.pdf>.
- Landine, J., and J. Stewart. (1998). “Relationship between Metacognition, Motivation, Locus of Control, Self-Efficacy, and Academic Achievement.” *Canadian Journal of Counselling* 32(3): 200–12.
- Lane, S., C. S. Parke, and C. A. Stone. (2002). “The Impact of a State Performance-Based Assessment and Accountability Program on Mathematics Instruction and Student Learning: Evidence from Survey Data and School Performance.” *Educational Assessment* 8(4): 279–315.
- Lin, Y.-G., W. J. McKeachie, and Y. C. Kim. (2001). “College Student Intrinsic and/or Extrinsic Motivation and Learning.” *Learning and Individual Differences* 13(3): 251–8.

- Lombardi, A, M. Seburn, and D. Conley (2011). "Development and initial validation of a measure of academic behaviors associated with college and career readiness." *Journal of Career Assessment* 19(4): 375–91.
- Longo, C. (2010). "Fostering Creativity or Teaching to the Test? Implications of State Testing on the Delivery of Science Instruction." *The Clearing House* 83(2): 54–57.
- Mansilla, V. B., and H. Gardner. (1998). *What Are the Qualities of Understanding? Teaching for Understanding: Linking Research with Practice*, 161–96. San Francisco: Jossey-Bass Publishers.
- Masters, G. and B. McBryde. (1994). *An Investigation of the Comparability of Teachers' Assessments of Student Folios*. Brisbane, Australia: Tertiary Entrance Procedures Authority.
- McNeil, L. (2000). *Contradictions of School Reform: Educational Costs of Standardized Testing*. Oxford, UK: Routledge.
- Messick, S. (1994). "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Researcher* 23(2): 13–23.
- Metcalf, S. J., J. A. Chen, A. M. Kamarainen, K. M. Frumin, T. L. Vickrey, T. A. Grotzer, and C. J. Dede. (2013). "Shifts in Student Motivation during Usage of a MultiUser Virtual Environment for Ecosystem Science." *Proceedings of the National Association for Research in Science Teaching (NARST) Annual Conference*. <http://www.narst.org/annualconference/2013conference.cfm>.
- Metcalf, S. J., A. M. Kamarainen, M. S. Tutwiler, T. A. Grotzer, and C. J. Dede. (2011). "Ecosystem Science Learning via Multi-User Virtual Environments." *International Journal of Gaming and Computer-Mediated Simulations* 3(1): 86–90.
- Mohr, J. J., R. J. Fisher, and J. R. Nevin. (1996). "Collaborative Communication in Interfirm Relationships: Moderating Effects of Integration and Control." *Journal of Marketing* 60(3): 103–15.
- Nabi, G., R. Holden, and A. Walmsley. (2010). "From Student to Entrepreneur: Towards a Model of Graduate Entrepreneurial Career-Making." *Journal of Education and Work* 23(5): 389–415.
- Nichols, S. L., and D. C. Berliner. (2007). *Collateral Damage: How high-stakes Testing Corrupts America's Schools*. Cambridge, MA: Harvard Education Press.
- Organisation for Economic Co-operation and Development. (2013). *Draft Collaborative Problem Solving Framework*. <http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf>.
- Papay, J. P., R. J. Murnane, and J. B. Willett. (2010). "The Consequences of High School Exit Examinations for Low-Performing Urban Students: Evidence from Massachusetts." *Educational Evaluation and Policy Analysis* 32(1): 5–23.
- Paris, S. G., and P. Winograd. (1990). "How Metacognition Can Promote Academic Learning and Instruction." *Dimensions of Thinking and Cognitive Instruction* 1: 15–51.
- Parkhurst, H. B. (1999). "Confusion, Lack of Consensus, and the Definition of Creativity as a Construct." *Journal of Creative Behavior* 33(1): 1–21.

- Pellegrino, J. W., and M. L. Hilton. (2013). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st century*. Washington, DC: National Academies Press.
- Peredo, A. M., and M. McLean, M. (2006). "Social Entrepreneurship: A Critical Review of the Concept." *Journal of World Business* 41(1): 56–65.
- Perie, M., S. Marion, and B. Gong. (2009). "Moving toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments." *Educational Measurement, Issues and Practice* 28(3): 5–13.
- Queensland Studies Authority. (2010). "School-Based Assessment: The Queensland System." Queensland, Australia. [http://www.qsa.qld.edu.au/downloads/approach/school-based\\_assess\\_qld\\_sys.pdf](http://www.qsa.qld.edu.au/downloads/approach/school-based_assess_qld_sys.pdf).
- Quellmalz, E. S., M. J. Timms, M. D. Silbergitt, and B. C. Buckley. (2012). "Science Assessments for All: Integrating Science Simulations into Balanced State Science Assessment Systems." *Journal of Research in Science Teaching* 49(3): 363–93.
- Rosenthal, R., and L. Jacobson. (1968). *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development*. St. Louis, MO: Rinehart and Winston.
- Ruiz-Primo, M. A., and R. J. Shavelson. (1996). "Rhetoric and Reality in Science Performance Assessments: An Update." *Journal of Research in Science Teaching* 33(10): 1045–63.
- Runco, M. A. (1996). "Personal Creativity: Definition and Developmental Issues." *New Directions for Child and Adolescent Development* 72: 3–30.
- Ryan, R. M., and E. L. Deci. (2000a). "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions." *Contemporary Educational Psychology* 25(1): 54–67.
- . (2000b). "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." *American Psychologist* 55(1): 68–78.
- Saavedra, A. R., and V. D. Opfer. (2012). *Teaching and Learning 21st Century Skills: Lessons from the Learning Sciences*. New York: Asia Society.
- Sanders, W. L., and S. P. Horn. (1998). "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research." *Journal of Personnel Evaluation in Education* 12(3): 247–56.
- Sawyer, R. K. (2006). "Educating for Innovation." *Thinking Skills and Creativity* 1(1): 41–48.
- Schunk, D. H., and B. J. Zimmerman. (2012). *Motivation and Self-Regulated Learning: Theory, Research, and Applications*. Oxford, UK: Routledge.
- Schwartz, H., L. S. Hamilton, B. M. Stecher, and J. L. Steele. (2011). *Expanded Measures of School Performance*. Santa Monica, CA: RAND Corporation.
- Secondary School Admission Test Board. (2013). "Think Tank on the Future of Assessment." Princeton, NJ.
- Shallcross, D. J. (1981). *Teaching Creative Behavior: How to Teach Creativity to Children of All Ages*. Englewood Cliffs, NJ: Prentice-Hall.

- Shavelson, R. J., G. P. Baxter, and X. Gao, X. (1993). "Sampling Variability of Performance Assessments." *Journal of Educational Measurement* 30(3): 215–32.
- Smith, M. L. (1991). "Put to the Test: The Effects of External Testing on Teachers." *Educational Researcher* 20(5): 8–11.
- Stecher, B. (1998). "The Local Benefits and Burdens of Large-Scale Portfolio Assessment." *Assessment in Education* 5(3): 335–51.
- . (2002). "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practices." In *Making Sense of Test-Based Accountability in Education*, edited by L. Hamilton, B. Stecher, and S. Klein, 79–100. Santa Monica, CA: RAND Corporation.
- Stecher, B. M., and J. L. Herman. (1997). "Using Portfolios for Large-Scale Assessment." In *The Handbook of Classroom Assessment: Learning, Achievement and Adjustment*, edited by G. D. Pyle. Orlando: Academic Press.
- Stecher, B. M., and S. P. Klein. (1997). "The Cost of Science Performance Assessments in Large-Scale Testing Programs." *Educational Evaluation and Policy Analysis* 19(1): 1–14.
- Sternberg, R. J. (2010). "Teach Creativity, Not Memorization." *The Chronicle of Higher Education*. Retrieved from [http://www.yosoyartista.net/bobbyo\\_files/2010-11-07-8BA990.pdf](http://www.yosoyartista.net/bobbyo_files/2010-11-07-8BA990.pdf).
- Stevenson, H. W., S.-Y. Lee, and J. W. Stigler. (1986). "Mathematics Achievement of Chinese, Japanese, and American Children." *Science* 231(4739): 693–99.
- Stevenson, H. W., S.-Y. Lee, C. Chen, J. W. Stigler, C.-C. Hsu, S. Kitamura, and G. Hatano. (1990). "Contexts of Achievement: A Study of American, Chinese, and Japanese Children." *Monographs of the Society for Research in Child Development* 55: i–119.
- Stigler, J. W., and H. W. Stevenson. (1991). "How Asian Teachers Polish Each Lesson to Perfection." *American Educator* 15(1): 12–20.
- Stone, D. N., E. L. Deci, and R. M. Ryan. (2009). "Beyond Talk: Creating Autonomous Motivation through Self-Determination Theory." *Journal of General Management* 34(3): 75.
- Tan, E. (January 2013). "Assessment in Singapore: Assessing Creativity, Critical Thinking and Other Skills for Innovation." *Proceedings of the OECD Educating for Innovation Workshop*. [http://www.oecd.org/edu/cei/07%20Eugenia%20Tan\\_Singapore.pdf](http://www.oecd.org/edu/cei/07%20Eugenia%20Tan_Singapore.pdf).
- Tanveer, M. A., O. Shafique, S. Akbar, and S. Rizvi. (2013). "Intention of Business Graduate and Undergraduate to become Entrepreneur: A Study from Pakistan." Retrieved from [http://www.textroad.com/pdf/JBASR/J.%20Basic.%20Appl.%20Sci.%20Res.,%203\(1\)718-725,%202013.pdf](http://www.textroad.com/pdf/JBASR/J.%20Basic.%20Appl.%20Sci.%20Res.,%203(1)718-725,%202013.pdf).
- Trompenaars, F., and C. Hampden-Turner. (1998). *Riding the Waves of Culture*. New York: McGraw-Hill.
- Vrugt, A., and F. J. Oort. (2008). "Metacognition, Achievement Goals, Study Strategies and Academic Achievement: Pathways to Achievement." *Metacognition and Learning* 3(2): 123–46.

———. (2010). *The Global Achievement Gap: Why Even Our Best Schools Don't Teach the New Survival Skills Our Children Need—and What We Can Do about It*. Phoenix, AZ: Basic Books.

Walton, G. M., and G. L. Cohen. (2011). "A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students." *Science* 331(6023): 1447–51.

Walumbwa, F. O., B. J. Avolio, W. L. Gardner, T. S. Wernsing, and S. J. Peterson. (2008). "Authentic Leadership: Development and Validation of a Theory-Based Measure." *Journal of Management* 34(1): 89–126.

Wilson, L. D. (2007). "High-Stakes Testing." *Second Handbook of Research on Mathematics Teaching and Learning*. Charlotte, NC: IAP (Information Age Publishing), 1099.

Yeager, D. S., and G. M. Walton. (2011). "Social-Psychological Interventions in Education They're Not Magic." *Review of Educational Research* 81(2): 267–301.

Yeager, D., G. Walton, and G. L. Cohen. (2013). "Addressing Achievement Gaps with Psychological Interventions." *Phi Delta Kappan* 94(5): 62–65.

Zimmerman, B. J. (1990). "Self-Regulated Learning and Academic Achievement: An Overview." *Educational Psychologist* 25(1): 3–17.

———. (2001). "Theories of Self-Regulated Learning and Academic Achievement: An Overview and Analysis." In *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives*, 2, 1–37.



---

Global Cities  
Education Network

