Chapter 5
The Assessment Landscape in the United States: From Then to the Future

Eva L. Baker and Harold F. O'Neil Jr.

Assessment in Elementary and Secondary Schools

What is the state of assessment in US elementary and secondary schools? Certainly, in the current scene, times have changed from periods when testing occupied center stage of salient topics reported in media, and the foci of public policy, research, and practice. Why has testing contention faded from its historic centrality in US educational public policy discourse? Pointed questions have all but vanished about preemption of teachers' role by the use of external tests? Because in the past two decades, new teachers were hired into a test-driven accountability environment, many now know nothing different. Gareis (2017) points out that new teachers have also experienced test-based accountability from the vantage point of students. The day has been won by the advocates of testing with little if any substantiation that accountability provisions work, that is, that they have been shown to improve schools.

E. L. Baker (✉)
UCLA Graduate School of Education & Information Studies, Los Angeles, CA, USA

National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA, USA
e-mail: baker@cse.ucla.edu

H. F. O'Neil Jr.
University of Southern California, Los Angeles, CA, USA

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_5

Elementary and Secondary School Assessment

We focus on US Federal policy history leading up to the most current Federal provisions. Recall first that in the United States education is a function principally under the authority of each of the 50 states rather than a systematic Federal responsibility. However, the Federal government has considerable influence in that it can allocate marginal funds to states for particular programs with these funds providing disproportionate incentives.

US Federal Policy History on Assessment More than 50 years ago, the Federal government stepped in to the educational arena to address inequitable educational practices and lagging outcomes for poor and minority students. This disparity was verified by James Coleman in his landmark study of equality of educational opportunity (1966). Watershed legislation was passed by Congress in the first Elementary and Secondary Education Act (ESEA 1965) influenced as much by politics as research. This law notably introduced testing provisions to assure that underperforming and economically challenged students made demonstrated progress in academic performance and stands as the first major, national effort in accountability. This and subsequent reauthorizations of the law prescribed the use of commercialized standardized tests focused on general constructs such as reading and mathematics ability (see, for instance, the summary in Baker et al. 2016).

Almost 30 years later, in 1994, a significant revision of ESEA occurred, i.e., the Improving America's Schools Act (IASA 1994). IASA notably shifted in testing requirements from only disadvantaged students to include *all* students. IASA development had been strongly influenced by a series of high-profile, bipartisan reviews and recommendations: for example, *A Nation at Risk* (NCEE 1983) identified the reduction of US achievement; the *National Education Goals* (1999) articulated by the National Governor's Association specified goals in early childhood, literacy, teaching, and adult learning and including a fanciful goal that the United States would be "best in the world" in mathematics and science by the year 2000 (www2.ed.gov) and *Raising the Standards for American Education* (NCREST 1992). This letter review was issued by the National Council on Education Standards and Testing, a council of national and State legislators and measurement experts (the lead author participated).

In addition to broadening the focus from those economically deprived and minority students, IASA supported a set national standard to frame the assessments borrowing from practices in the United Kingdom. Standards-based assessments were supposed to be influenced by instruction, but in fact, it changed little of actual testing practice. Commercial assessments were marketed as standards-based, but in many cases on, relied on existing item pools and on traditional approaches to item development and psychometrics. Commercial publishers simply reformatted reporting from score averages to results expressed in terms of numbers of students reaching given thresholds (performance standards).

Alignment to curriculum and instruction was advocated, but took place *post* hoc rather than as in a coherent design (Baker 2005). IASA also provided for the use of noncognitive measures in accountability reports, but first ventures used archival information, such as school absences, rather than actual assessments of students' intrapersonal skills such as effort or self-efficacy.

ESEA legislation was again changed leading to the contentious "No Child Left Behind" (NCLB 2001). This legislation significantly modified IASA, with a mix of desirable and ultimately less useful provisions. For example, NCLB asked states for a plan to document Adequate Yearly Progress (AYP). These growth curves were to show how the yearly indicators of student achievement would eventually result in all students reaching the desired proficiency standards by 2014. Many states successfully gamed their design of AYP projections to start with small annual increments but sharp (and unrealistic) increases in later years in order to reach the 2014 proficient targets. AYP heightened attention to test-driven instruction. AYP results were to be disaggregated by subgroup to show disparities in performance to be overcome. If only one subgroup failed to make the AYP goal, the school was deemed failing. It isn't hard to estimate that more diverse schools, with more reportable subgroups, had increased probability of failure and these schools were often low-income schools.

On the positive side, NCLB included stronger equity provisions, requiring that a high percentage (95%) of students from identifiable groups, such as disadvantaged, English learning, participate in mandatory testing in order to combat inflated achievement results of schools and districts that encouraged poorer performing students to stay at home on testing dates. Like AYP, schools could be labeled "failing" if they did not meet participation levels. However, because most subgroups were each required to meet 95% participation, probabilistically more diverse schools were certain to fail over a 5-year period, just by chance. NCLB, like some of its predecessors, had consequences for failing schools. It required sanctions by states for districts and schools that persistently underperformed, after attempts at school improvement had been made.

Another set of education requirement later modifying NCLB was the Race to the Top legislation signed into law in 2009 as part of the American Recovery and Reinvestment Act (ARRA 2009), which was the Federal response to the worldwide financial downturn. The Race to the Top law created a competition among states for educational reform awards totaling more than $4 billion dollars. The RTT law included traditional ESEA provisions, such as developing and adopting common standards, requiring procedures for teacher and principal evaluation, supporting transition to standards, and developing longitudinal data systems based on individual performance. It also rewarded the development and expansion of charter schools.

The law prescribed, in addition to accountability and professional evaluation, the use of student data to improve instruction, a provision that implies different outcome measures. The absolute priority required for funding was that the state describes a comprehensive approach to reform where assessment was a strong component. Over three rounds of competition, 18 states (or 36%) were awarded amounts ranging from 500 million to 17 million dollars. However, the competition was

roundly criticized. EPI, for example, published a critique of a report by Shavelson et al., criticizing the use of student test scores to evaluate teachers and value-added analyses (Baker et al. 2010).

As part of the educational reform in this time period, and in recognition of the central role of assessment, the Federal government with ARRA funds held a competition for consortia of states to develop standards-based assessments to use in required standards-based assessment and to promote college and career readiness. These consortia were to prepare and to motivate others to develop more innovative approaches to testing. Two major awards were given, one to the Smarter Balanced Assessment Consortium (SBAC) and the other to Partnership for Assessment of Readiness for College and Careers (PARCC), each basing their assessments on the Common Core Standards in Math and English Language Arts (2010), projects that had been supported by private foundations and businesses. Starting with virtually all states participating, the consortia lost members over the next few years as states defected from the common core standards and chose less ambitious, less costly approaches. In any case, the innovation intended for the consortia has been marginal, perhaps because both developers and users were state policy makers already under various political and economic pressure. Another damper on innovation was that their contractors were drawn from the usual pool of standardized test purveyors.

Present-Day Federal Assessment Policy The most recent reauthorization of the ESEA is Every Student Succeeds Act (ESSA 2015). External testing for accountability unsurprisingly remains a part of required educational practice. The barebones of its provisions call for annual testing in reading and math in grades 3–8 plus once in secondary school. Science is to be tested at least once in each of the grade spans of 3–6, 6–9, and 10–12. Testing of English language proficiency is to occur annually in elementary and secondary grades for all those who qualify for English language development programs and to continue until students are transitioned to fluent English learner status. Assessment for students with special needs is also reflected in the ESSA law.

For the most part, assessments for all students are to be standards-based, but no longer emphasize common core standards (which fell into political disrepute). However, no criteria for assessment quality were provided. Some "flexibility" or variations in assessment are enabled in the legislation, for example, encouraging the development of innovative assessments and allowing the use of existing commercial tests at the high school level (operationally meaning the use of the SAT or ACT, college admission tests rather than measures of course based achievement). Graduation rates of 67% are required for each subgroup. Nonacademic measures are to be included in accountability reports, for example, student engagement, climate, safety, and postsecondary readiness. As before, archival information can be used, and the law now allows some process indicators such as the provision of resources, such as enrichment.

Parents are given the explicit ability to "opt out" of testing, but 95% rates of assessment participation are still required for each identifiable subgroup. However, in contrast to NCLB, falling below this percentage does not result in sanctions.

Moreover, expectations for making adequate yearly progress, the nature of interventions for underperforming schools, and reporting options are left to the devices of the various states. Disaggregating results by subgroup is still required, and new subgroups for reporting were added, including students with one or more parent in the military, homeless students, and students in foster care, the last three a chilling assessment of changes in American society. All told, ESSA represents a weakened form of accountability testing, and it is in large measure a compromised wrought by a Democratic administration and a Republican congress.

National Assessment of Educational Progress In addition to Federal statutes that influenced state assessment development, the Federal government itself manages additional accountability measures. Most important is the National Assessment of Educational Progress (NAEP), fully funded by the Federal government and administered by the by the National Center for Education Statistics (NCES), and is currently managed by the Educational Testing Service (ETS). NAEP also has a long history. Since the 1960s, NAEP has sought to provide overall and disaggregated estimates of the performance of students across the United States. It is administered on a sampling basis, with mathematics, reading, science, and writing domains regularly assessed for grades 4 and 8. There are aperiodic measures of other domains, such as Civics and Art. Although there is an extended trend line of performance reaching back several decades, for the last 20 years or so, each assessment has been generated using a domain-specific framework developed by educators and experts in assessment and contact domains. These frameworks and resulting assessments are intended to reflect more contemporary views of content and skills as currently taught in schools. Results from the NAEP administrations typically offered on a 2-year cycle, are provided overall, disaggregated by subgroups, such as gender, race/ethnicity, disabilities, proxies for socioeconomic level, and identified English learners. Although at its inception NAEP's purpose was to lead in assessment innovation, that goal has had only sporadic attention recently, focused on technology options.

Originally, great pains were taken so that NAEP would not be used to compare state and local jurisdictions and lead to undue focus on the test content during instruction. Instead, as advocates of accountability grew more insistent, NAEP changed its sampling and reporting protocols so that results could be compared not only on broad geographic bases and across administration cycles but to a more explicit comparative framework. NAEP changed from gross regional reporting, e.g., West, South to reports of state by state achievement. Later, 25 large city schools were also sampled sufficiently to provide usable data so that their achievement could be compared with one another as well with states.

Across the states, from cycle to cycle, there is variation in achievement, with high-performing states having both high levels of achievement on NAEP and high graduation rates. But the interpretation of NAEP data is conflicted. For example, from 1960 to present, the overall trajectory of NAEP is largely flat, meaning that no major improvement has been found. However, there has been discernible growth in some aspects of NAEP achievement. Over last 20 years, there has been more

noticeable at particular percentile levels, e.g., 25th. There is also variation in US states. The best performing states have between 80% and 87% of students scoring at or above the basic level (the lowest standard) and between 40% and 53% scoring at or above proficient in 4th grade math.

It should be noted that some measurement experts have had difficulty with NAEP achievement levels (or performance standards) thinking that they were set too high. There have been sunnier interpretations of overall data; for instance, we have seen noticeable improvement, in closing the gap between Latino and White students. But overall, not much has happened. Particularly disappointing given the federal and state resources invested in education during this timeframe.

There are numerous explanations for the NAEP data, some more apologetic than others. Here is a brief reprise of them: (1) The frameworks guiding NAEP development have only loose connections with state and district curriculum emphases, so that instruction is similarly loosely connected. At best, NAEP could be considered as measures of transfer from school learning. (2) Students taking NAEP have little incentive to perform well, as the test doesn't "count" in any noticeable way. Without incentives, performance can lag, especially when there is a surfeit of testing for students. (3) The educational system has absorbed during this time period many new types of students, with myriad background and language issues, all of which could impact NAEP performance. (4) The lower socioeconomic level of US society has dropped below the wealthiest sectors. Students are coming from struggling families who may have less opportunity to help their children learn. (5) Technology may not be supporting learning. Screen time is up, with 2017 data suggesting that students between 8- and 10-year-old spent an average of 6 h a day, up about 50% from 2015 and those 11–14 in front of screens an average of 9 h a day. With time split among mobile devices, computers, and television, much of the screen time is spent in social interaction, such as texting not much time for reading is left. (6) The numbers for out of school reading are not easily compared, but overall, the United States is one of the lower countries in number of minutes read a day, and the results diminish with age. So, it is likely that children are not seeing much modelling of intellectual activity at home. (7) NAEP is not high stakes for students, teachers, or principals; thus there is little teaching to the test. So, in combination with other findings, it is reasonable to propose that the NAEP findings actually represent US achievement.

Assessment in Workforce

Workforce readiness is a concern of educators at the K-12 and postsecondary levels. It also has a long history. For example, in 1991, Secretary's Commission on Achieving Necessary Skills (SCANS) identified twentieth century (now twenty-first-century skills desired by employers). These involved key cognitive demands, personal responsibility, and interpersonal skills. Although there is a history of specifying workforce requirements to inform schooling, two major innovations

should be noted. First, workforce expectations entered specifically into the K-12 curriculum as 20th and cognitive readiness (O'Neil et al. 2014). These skills included domain independent skills like problem-solving and communication, personal behaviors like timeliness, and readiness to learn and interpersonal skills like teamwork (Pellegrino and Hilton 2012). Personal requirements like abstaining from drug use are also considered. These general topics seemed to recur in periodic surveys of employers (NACE 2014). In higher education, annual surveys have reported on the perceptions of college students about their job readiness. Only about 4 in 10 feel well prepared for their careers after graduation and feel unprepared in transitional skills, like resume writing and interview preparation, as well as areas like problem-solving. Gender gaps also occur with men feeling better prepared than women despite that data favor women in graduation rates and access to graduate and professional education.

Trends like job insecurity and wage stagnation provide incentives for students and institutions to focus more heavily on career preparation rather than foundational skills, which are usually the focus of college and university. For liberal arts, a recent study (Jaschik 2016) reported almost 9% loss of major in the humanities, and universities regularly report starting salaries of graduates by their major course of study. The large average burdens of student loans (currently just under $40,000 inhibit university students accepting low-paying, entry-level jobs https://studentloanhero.com/student-loan-debt-statistics).

One interesting resource is the analysis of needed competencies that is widely available. The US Department of Labor has sponsored website (O∗Net) that provides a list of potential occupations supported with each illustrated by the set of skills required. This site also shares updates on the potential availability of occupations for job seekers of all ages. We expect an increase in assessments relevant to workforce skills at both the pre-collegiate and higher education levels.

An Innovative Workforce Assessment System: Training Assessment Framework (TAF)

We now briefly report the CRESST design and findings of a multiyear Navy assessment project that applies assessment lessons learned from the K-12 and higher education settings, as well as insights from the science of assessment. A new system for assessment is under development by CRESST for the Navy Education and Training Command (NETC), the group responsible in the Navy for preparing young enlisted US sailors for more 70 career paths.

It is based on previous CRESST models for assessment design (Bewley et al. 2009) implementation and validation. The model has a number of important attributes. First, it is designed to serve multiple purposes, including certification of end-of-course competencies. Second, the system will support aggregation of performance and achievement results across disparate courses and jobs. A third

purpose was to provide feedback to instructors and curriculum designers. A fourth and central purpose of the project was to provide an additional indicator to validate the job classification decision made at the point of recruitment. The decision is based predominantly on the ASVAB (https://www.military.com/join-armed-forces/asvab).

The project team will develop reports (both graphical and in text) tailored to different audiences and commit to develop innovative technical approaches to validate inferences for the full range of purposes. The Training Assessment Framework design document (Baker and Choi 2018) lists user needs, cognitive demands, domain knowledge acquisition, and task specifications to delimit parameters to guide assessment development. These parameters will allow future of automated design systems. This framework also address validity, e.g., development comparisons of expert and novice performance.

Two courses for two different Navy jobs or ratings were the proof of concept for the Training Assessment Framework Model to determine whether the process was generalizable. The very different Navy jobs were damage control (ship protection) and fire control (a radar-technician-like job). The former involved dangerous situations, team environments, and problem-solving and procedural knowledge. The second is a technical electronics job that involved operations and maintenance skills. Navy subject matter expects in these areas helped determine relevant content for assessment. This step was followed by extensive development of ontologies documenting content for each rating (Baker and Choi 2018).

Three major types of formats comprised the examination. The first was the generally familiar selected response format that used multiple choice, drag and drop, and hotspots to show answers. The second involved performance assessments using scenario-based simulations asking for search and procedural knowledge on integrated tasks likely to be confronted by the sailor on the job. The third was a knowledge map (O'Neil and Chung 2011) that asked sailors to create nodes and links to reveal them understand of the hierarchy, structural and functional relationships among important elements of the job knowledge domain. The maps are scored against expert maps. Trainees were also asked to complete affective scales measuring domain-specific self-efficacy and anxiety so as to allow us to understand the degree to which different test formats influenced affective responses.

We were constrained to use a technology platform (iPad) and to limit testing time. The examinations were administered on a pre-and post-instructional basis to trainees' subject matter experts (SMEs) to determine a benchmark for validity. The report of this work (Baker and Choi 2018) documents the design options, describes the data, and presents innovative psychometric and validity solutions to performance and knowledge mapping items.

In the future, the project will scale-up with additional Navy jobs. Our commitment to develop assessments with a heavy emphasis on their technical quality (e.g., validity evidence) will have relevant implications for civilian workforce assessment and for K-12 systems as well.

Critical Discussion

The accretion of requirements for K-12 assessment in federal law, for the most part, has been decisive in the management of schools. Because of the historical concern with accountability, terms framing instruction as "test driven" or "evidence based" are frequent, with almost no discussion of the quality of the evidence. Earlier warning about the limitations of instruction geared to what could be measured seems to have evaporated as a major concern. Alignment studies may address standards but more attention is given mapping instruction to actual test specifications. Despite professional and academic attention to appropriate requirements for design, development, and validity, there is little evidence that validity is taken seriously, despite widespread recognition of the Standards for Validity (AERA, APA, NCME 2014, 1999, 1985). For instance, the notion of policy capture, that is, what experts think performance should be related to standard setting for achievement, has little in the way of empirical validity, and it is an essential part of criterion-referenced reporting models (Baker 2012). Moreover, there is rare validity attention to the full range of purposes for tests, e.g., certification, accountability, or instructional improvement, and none is easily accessed on the public-facing websites by testing groups.

To end this tour, let's consider some trends in society that are influencing assessment. The first, which we have informally observed, is there is less attention and reliance on technical quality indicators derived from high-quality research. Moreover there is a general retreat from scientific bases of quality mirrors everyday life, where stars, likes, and other populist indicators of quality have seemingly replaced reliance on technical expertise and evidence. Although most commercial tests received some level of analysis, many focus solely on content validity; that is, are important content and skills measured and are they representative of desired content? A second general purpose motivated by the legal system is the verification of fairness and equity; that is, are all student given a fair opportunity to succeed?

Second, psychometrics has continued to evolve and indicators will move toward findings that can be directly interpretable for improvement of learning, as opposed to those procedures that require transformations that take findings somewhat far afield from their original design. One positive development has been the exploration of feature analysis as a validity and design check on assessment. Here cognitive demands, task requirements, and domain content are qualitatively rated on existing or newly developed items, and then relationships of item features and performance on the test are examined (Baker and Cai 2014; Baker 2015; Madni et al. 2018).

Finally, there is the impact of technology as a pervasive element in modern life, no less in testing. Technology has been used in ways to make assessment more efficient (computer-adapted testing) and more palatable, in game and simulation-based tests, in automating scoring (Burstein 2003) and in supporting automated design. The energy of the startup community in testing, with their race to market, has once again served to reduce commitment to careful validation studies, seen by the startup community as not to add much value and slows access to the market.

Unless we can continue to develop infrastructure tools, such as automated ontology extraction, automated scoring and reporting, and simulated students for accelerated data collection, we fear that the demand for data and the propensity to develop empirical findings to drawn validity inferences will die a relatively quick death. However, we look enthusiastically to the future where we believe assessment, technical quality, and technology when properly combined (Baker et al. 2016) will continue to drive greater utility, motivation, consequences, and innovation in assessment and learning.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5.

Baker, E. L. (2005). Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 315–335). Mahwah: Erlbaum.

Baker, E. L. (2012). Standards for educational and psychological testing. In J. A. Banks (Ed.), *Encyclopedia of diversity in education* (Vol. 4, pp. 2076–2081). Thousand Oaks: SAGE.

Baker, E. L. (2015, May). *Feature analysis: Improving the validity of competency tests.* Presentation at the colloquium on assessment and accountability implications of competency and personalized learning systems. Boulder, CO.

Baker, E., & Cai, L. (2014, August). *CRESST analysis of the quality of a state assessment* (Technical slide report). Los Angeles: CRESST.

Baker, E. L., & Choi, K. C. (2018). *Training Assessment Framework* (Report to funder). Los Angeles: University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Barton, P. E., Darling-Hammond, D., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010, August). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper #278). Washington, DC: Economic Policy Institute. Available at http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf

Baker, E. L., Chung, G., & Cai, L. (2016). Assessment gaze, refraction, and blur: The course of achievement testing in the last hundred years. *Review of Research in Education* (Centennial Issue), *40*, 94–142.

Bewley, W. L., Chung, G. K. W. K., Delacruz, G. C., & Baker, E. L. (2009). Assessment models and tools for virtual environment training. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 300–313). Westport: Praeger Security International.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach* (pp. 113–121). Mahwah: Lawrence Erlbaum Associates.

Coleman, J. S. (1966). *Equality of educational opportunity*. Oxford: U.S. Department of Health, Education.

Elementary and Secondary Education Act of 1965 as amended, 20 U.S.C. §241 (1974).

Every Student Succeeds Act of 2015, Pub. L. No. 114-95.

Gareis, C. R. (2017). *Assessment leadership: Leveraging performance-based assessments for deeper learning*. Presented at the 21st Annual School-University Research Network Leadership Conference at the College of William & Mary, Williamsburg VA.

Improving America's Schools Act of 1994, Pub. L. No. 108, Stat. 3518.

Jaschik, S. (2016, March). The shrinking humanities major. *Inside Higher Ed*. https://www.insidehighered.com/news/2016/03/14/study-shows-87-decline-humanities-bachelors-degrees-2-years

Madni, A., Kao, J. C., Rivera, N. M., Baker, E. L., & Cai, L. (2018). *Exploring career-readiness features in high school test items through cognitive laboratory interviews* (CRESST Report 857). Los Angeles: University of California.

National Association of Colleges and Employers (NACE). (2014). *NACE 2013–2014 career services benchmark survey for colleges and universities*. Bethlehem: Author.

National Commission on Excellence in Education. (1983). *Nation at risk*. Washington, DC: Department of Education.

National Council on Education Standards and Testing. (1992, January 24). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People*. Washington, DC: Government Printing Office.

National Education Goals Panel. (1999). *The national education goals report: Building a nation of learners, 1999*. Washington, DC: U.S. Government Printing Office.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, §115, Stat. 1425 (2002).

O'Neil, H. F., & Chung, G. K. W. K. (2011, April). *Use of knowledge mapping in computer-based assessment*. In Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

O'Neil, H. F., Perez, R. S., & Baker, E. L. (Eds.). (2014). *Teaching and measuring cognitive readiness*. New York: Springer.

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