CHAPTER 19

After Admissions: What Comes Next in Higher Education?

María Elena Oliveri, Robert J. Mislevy, and Norbert Elliot

The chapters in this volume provide an international perspective on multiple areas of concern to higher education. Chapters focus on expanding the skill sets and constructs measured as part of college admissions (see Niessen & Meijer, in this volume; Kuncel, Tran, & Zhang, in this volume) and approaches to developing assessments that are sensitive to sociocognitive and sociocultural differences of the populations taking them (see Wikström & Wikström, in this volume). Efforts in these two areas aim to reduce sources of construct-irrelevant variance in assessments administered to students from diverse backgrounds in support of fair access to higher education institutions and to provide students with opportunities to develop a comprehensive (cognitive and noncognitive) skill set necessary for college success. In this chapter, we build on these areas and extend them to include a discussion about ways to better support students after admitting them to higher education institutions. The goal is to provide evidentiary data to help improve learning, decrease the percentage of remedial courses students take, and increase graduation rates.

A focus on enrollment and graduation is especially warranted as information on remedial-course enrollment and graduation rates of undergraduate students in the United States reveals challenges to remedial-course enrollment and the graduation of students from diverse ethnic groups in 2- and 4-year public institutions. As Figure 19.1 illustrates, enrollment in remedial courses is widespread and differs among subgroups. While 68 percent of Asian and 64 percent of White students were enrolled in remedial courses at public 2-year institutions, higher rates were reported among Black (78 percent) and Hispanic students (75 percent). Similar group differences are observed in 4-year institutions: While only about one-third of White (36 percent) and Asian (30 percent) students

The opinions and recommendations expressed here are those of the authors and not necessarily those of Educational Testing Service.

participated in remedial courses, a greater percentage of Black (66 percent) and Hispanic (53 percent) students were enrolled in such courses (Chen, 2016).

In terms of graduation rates, as Figure 19.2 reveals, these also differ by subgroup, with 35 percent of Asian and 29 percent of White students graduating from 2-year institutions within three years completion time and lower rates among Black (24 percent) and Hispanic (34 percent) students. Similarly, in 4-year institutions a higher percentage of Asian (73 percent) and White (64 percent) students graduated from 4-year institutions within six years completion time, with lower rates reported among Black (40 percent) and Hispanic (54 percent) students (Snyder, deBrey, & Dillow, 2019).

A comparison of Figures 19.1 and 19.2 shows that subgroups enrolled in a higher percentage of remedial courses had lower percentages of graduating students. For instance, Black and Hispanic students participated in a higher percentage of remedial courses (78 percent and 75 percent, respectively) in 2-year institutions and had the lowest graduation

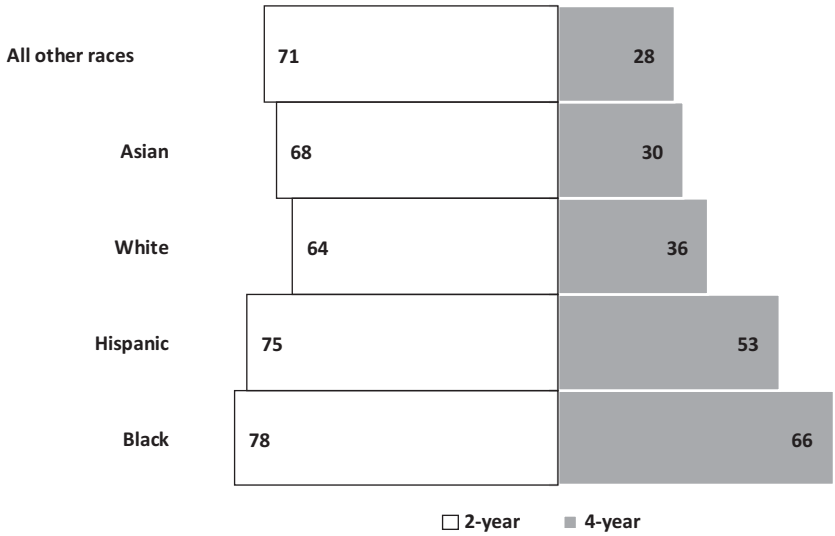

Figure 19.1 2003–2004 Remedial course enrollment in higher education institutions in the United States by race/ethnicity subgroup

Note. Numbers indicate the percentage of students among 2003–2004 beginning postsecondary students who first enrolled in 2- and 4-year higher education institutions who took remedial courses in any field by ethnic subgroup.

Source. Adapted from Chen (2016).

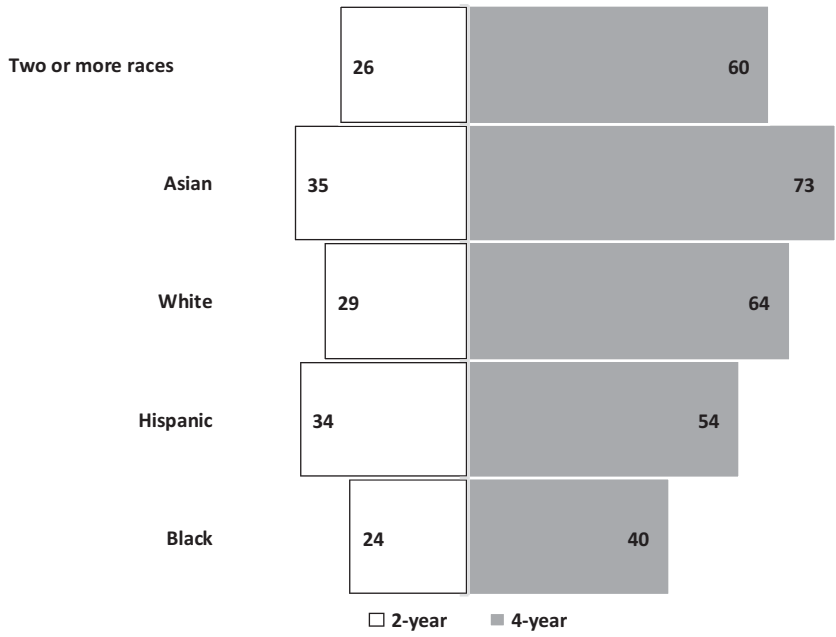

Figure 19.2 Graduation rates by higher education institution in the United States and race/ethnicity

Note. Cohort years are 2000–2013 for 2-year higher education institutions and 1996–2010 for 4-year higher education institutions. Numbers indicate the percentage of students who completed their 2-year degree within 150 percent (3-years) and 4-year degree within 150 percent (6-years) completion time. The table does not include race/ethnicity unknown or nonresident aliens.

Source. Adapted from Snyder, de Brey, & Dillow (2019).

rates (24 percent and 34 percent, respectively). Comparatively, in both 2- and 4-year institutions, White and Asian students participated in fewer remedial courses and had higher graduation rates. It appears that remediation is not having the desired effect on graduation rates. As discussed by Chen (2016), relationships between remedial coursework taken by students and their college outcomes vary by a student's level of academic preparedness. To elaborate, weakly prepared students who successfully completed all remedial courses in English/reading or mathematics may have experienced better postsecondary outcomes than did their counterparts who were weakly prepared but did not enroll in remedial courses. These same patterns did not hold for those students with moderate/strong preparation who were enrolled, perhaps incorrectly, in remedial courses.

The key, therefore, may not rest in single solutions such as remediation but, rather, in understanding obstacles comprehensively and designing strategies to overcome them in more exact ways.

Instead, alternative strategies are needed. We propose that assessment for learning is an important option to explore. Consequently, we discuss the need for new assessment frameworks. While our examples are drawn from assessments primarily within the United States, we believe that the lessons learned, proposed models, and future directions we identify have value for an international audience.

As such, we provide two models: the multilevel design model (MDM) and the complementarity model (CM). When taken together, the models aim to better support diverse students' learning by improving the connection between assessments and instruction once students are admitted to higher education institutions. Thus, we suggest that students benefit when admissions, retention, and graduation processes are considered within unified frameworks rather than as unconnected atomistic events.

The goal of MDM is to minimize the unintended effects of test use when scores are employed primarily (or uniquely) to inform admissions decisions. Such effects may include the disjuncture between assessment and instruction, the use of assessments that do not provide data to inform student learning, or potential biases that may disadvantage subgroups of students identified by race, ethnicity, gender, or other discrete or intersectional categories when scores are used to inform decisions in isolation from other data sources. In such cases, assessment use may result in unintended consequences on stakeholders such as decision-makers, instructors, and students. Therefore, MDM considers both consequences of test use and the meaning of score-based interpretations together to help reduce sources of construct-irrelevant variance and support valid score-based inferences. As discussed later, these issues gain importance with shifts in the higher education climate that occur due to rising demands to assess an expanded skill set and an increasingly diverse student population.

CM aims to expand the uses and purposes of assessments from their current focus on summative uses (i.e., assessments used to evaluate student learning and skill acquisition at the end of an instructional period, such as a study unit or course) to include formative and embedded assessments. Formative assessments are used to monitor progress and provide ongoing feedback for instructors to improve their teaching and help students improve their learning. Embedded assessments are administered more frequently than formative assessments as course assignments, activities, or exercises as evidence of progress toward achieving a particular learning

outcome. Expanding the use of assessments to more frequent administrations during the course of the higher education program may also involve complementing the use of distributed assessments (assessments that are standardized for administration across settings) with locally based assessments (assessments used within specific institutions). We conjecture that expanding the types of assessments used optimizes evidentiary data and better supports student learning, retention, and graduation.

In this chapter, we further describe the MDM and CM models and discuss their use to increase retention and graduation. To situate the models, we first describe shifts in higher education related to expanding the skill set, identifying the complexities of assessing culturally and linguistically diverse students, and increasing the uses of assessments. We believe if assessments are designed and used in ways that are more closely connected to institutions' contexts and purposes, they have the potential to support student retention and graduation.

19.1 Shifts in the Higher Education Environment

19.1.1 Implications of an Expanded Skill Set

Changes in higher education's landscape have implications for expanding the skill set that needs to be assessed. This expansion means that not only do the cognitive and academic skills needed for college success need to be assessed but that noncognitive skills must also be evaluated. Examples of noncognitive skills are collaboration (in the interpersonal domain) and persistence (in the intrapersonal domain). Both types of skills are important for degree completion and college success (National Research Council, 2012).

The skill set needs expanding because of an "increasingly global economy, elevated use of technology, and shifts in the types of economies and industries dominating the national and global economies" (Oliveri & Markle, 2017, p. 1). An expanded skill set is also needed, in light of research on classroom learning, student diversity, and the skills required for earning credit. For instance, teamwork may be necessary to complete classroom assignments and group projects or to summarize a particular student's contribution to projects (Oliveri, Lawless, & Molloy, 2017). The same is true of self-efficacy as a way to examine the independent and interactive effects of race and social class when evaluating interventions to close science, technology, engineering, and mathematics achievement gaps (Harackiewicz et al., 2016).

In broader terms, Hesse et al. (2015) note that higher education institutions do not tend formally to teach or assess collaborative skills. Results from employer surveys suggest that recent graduates are poorly prepared to meet workplace demands due to a lack of relevant noncognitive skills (Hart Research Associates, 2010). Coley, Goodman, and Sands (2015) warn that reacting too slowly in expanding students' skill sets may adversely affect a nation's ability to respond to increasing international demands, with negative effects on employability and economic prosperity.

Assessing this expanded skill set, important to both academic and workplace settings, presents challenges to testing organizations. The challenges include the need to develop new assessment frameworks and tasks that better integrate cognitive and noncognitive constructs to meet shifting demands of the twenty-first century (Bereiter & Scardamalia, 2012).

19.1.2 Implications of Population Changes

Additional changes to assessments are needed due to the increased diversity of the student body attending higher education institutions. Mislevy (2018) argues for a sociocognitive approach to assessment design when assessing diverse populations to reduce sources of construct-irrelevant variance in assessments and obtain accurate score-based inferences. A sociocognitive approach to educational assessment and measurement views capabilities as emerging from the interplay of cognitive processes within persons and social and cultural processes across persons in a complex adaptive system. A sociocognitive approach to test development and score-based interpretations describes ways to remain attentive to key elements of task design and construct representation and the type of resources and knowledge culturally and linguistically diverse populations might bring to the assessment (O'Sullivan & Weir, 2011; Weir, 2005). To allow for valid and fair score-based interpretations when assessing populations from diverse backgrounds, the purpose of this approach is to guide decisions (e.g., the types of vocabulary, items, or situations) that can be included in an assessment without creating unnecessary complexity in the assessment (construct-irrelevant variance) or adapting tasks in accordance with test-takers' supporting capabilities.

A sociocognitive approach to testing is called for given the projected demographic shifts in the United States and internationally. Hussar and Bailey (2017) foresee an increasing number of students from various racial and ethnic backgrounds enrolling in higher education over the next 40 years. Internationally, Altbach, Reisberg, and Rumbley (2009) and Kelly,

Moores, and Moogan (2012) describe that increasing numbers of specific subgroups are attending higher education institutions, such as Asian international students.

Shifts in the demographic composition of potential students have implications for assessment design, conceptualization, and fairness (Dorans & Cook, 2016). One set of implications includes increased problems with the standardization of content or language and calls for developing tests that are more sensitive to the diverse backgrounds of test-takers. Consequently, professional organizations highlight the need for fairness considerations in assessment design and conceptualization. For example, the *Standards for Educational and Psychological Testing* include fairness as part of the section on foundational measurement, signaling a growing need "to support appropriate testing experiences for all individuals" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 5).

The International Test Commission (2018) proposes that considerations for fairly assessing culturally and linguistically diverse learners span an assessment's lifecycle, starting from conceptualization and design and extending to uses and score interpretation. For instance, administering the same test to diverse individuals may lead to incorrect inferences about individuals if tests contain language, item formats, or tasks that are differentially familiar to test-takers. Validity threats may emerge from the use of so-called exported assessments, those developed for domestic use and then administered in other countries in the same or a different language (Oliveri & Lawless, 2018). As the use of exported higher education assessments increases, attention to design principles from a sociocognitive perspective are necessary for identifying ways to minimize irrelevant sources of score variance that may emerge due to differences in the opportunity to learn, curricular exposure, or familiarity with cultural references used, which may all present additional validity threats for these assessments.

19.1.3 Uses of Higher Education Assessment Scores

Changes in the skill set measured by assessments and demographic shifts also have implications on how information gleaned from tests is used to monitor student learning or placement decisions regarding which courses students are ready to pursue. In *Who Gets In? Strategies for Fair and Effective College Admissions*, Zwick (2017) described admissions policies as more than a set of rules, and proposed principles for informing decisions. These principles include acknowledging that there is no universal

definition of merit. Therefore, the selection of which students to admit may vary, depending on the predictors evaluated. Such evaluation may expand to using high school grades to measure noncognitive constructs (e.g., tenacity and commitment), being transparent in the use of admissions rules to increase fairness and access to colleges for diverse student groups, or using a combination of criteria to inform admissions.

A non-universal definition of merit is particularly relevant in the assessment of diverse populations because candidates may have different backgrounds or educational and life experiences. Such differences challenge the paradigm of comparing individuals on a set of common criteria and using a ranking system to inform admissions decisions in which only some individuals are admitted. We believe that while the principles described by Zwick (2017) are important to admissions, similar principles may be applied to students with diverse backgrounds throughout their higher education studies to improve retention and graduation rates, as we elaborate later. Concomitant with the expansion of uses of assessments is a growing need for arguments and evidence to support score interpretations, from which implications for assessment design, construction, and the interpretation of claims are derived. Many assessment models focus on psychometrics and technical accuracy and do not explicitly include approaches using assessment data to inform score-based decisions affecting stakeholders.

Now situated with reference to shifts in higher education related to expanding skill sets and addressing the needs of culturally and linguistically diverse students, we turn to applying an MDM model for designing and using assessments to guide decisions and efforts in a structured manner.

19.2 A Multilevel Design Model for Assessment Development and Use

We start our description of the MDM by providing a taxonomy of assessment purposes, institution type, and stakeholders (see Figure 19.3). The face of the cube lists test purposes we described earlier (e.g., informing classroom instruction), which go beyond the use of tests to inform admissions decisions to the use of tests to also provide information that is useful to inform decisions throughout students' higher education studies in support of student retention and graduation. The cube also displays the stakeholders, which may require information at different grain sizes. For instance, policymakers may require information at a coarser level to make high-level policy decisions that offer meaningful strategies for monitoring

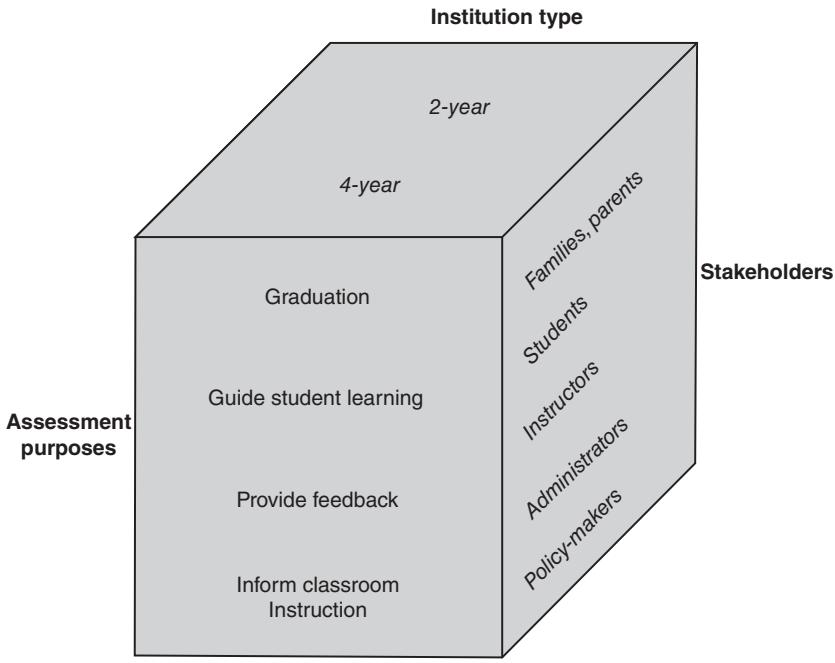

Figure 19.3 Taxonomy of assessment purposes, institution types, and stakeholders

educational improvements and outcomes for the performance of higher education institutions (Haertel & Herman, 2005). Administrators may require student-level data to inform decisions and institution-specific data to enable them to monitor student learning and progress within fields and programs within their institution. Individual students and instructors also require different types and levels of data (Oliveri & Wendler, 2017).

The diversity of assessment aims, institutional types, and stakeholders' needs presents additional challenges to developers to provide meaningful results. In this fluid environment, one challenge is the need to collaborate across multidisciplinary teams to design assessments. Evidence-centered design (ECD) (Riconscente, Mislevy, & Corrigan, 2016) can play a central role in guiding multidisciplinary teams of experts (i.e., assessment developers, cognitive psychologists, scientists, statisticians) and stakeholders (i.e., policymakers, administrators, instructors) to jointly identify and meet the goals of assessment use. This framework may involve developing a common language, mental models, artifacts, and best practice to capture the connected thinking underlying test design. ECD tools and concepts

are not sufficient, however, as they can capture design elements and rationales, but do not encompass the social system in which an assessment will function. The resources, constraints, purposes, and stakeholder perspectives composing the social system also need to be considered.

19.2.1 Model Overview

The MDM goals are to understand the following aspects of assessments: (a) the social system within which they operate, including the viable space they have for design and related constraints; (b) the purposes and constraints they must satisfy in operation; and (c) the effects that score use will have on the stakeholders. Explicit articulation of the goals and system within which assessments operate is needed to support the development of the more complex assessments attuned to today's economy; to help mitigate the negative consequences associated with the primary uses of summative assessments and the measurement of a narrower set of traditional constructs; and improve the use, meaning, and impact of score-based assessments.

Figure 19.4 illustrates the MDM for assessment design, development, and use. The MDM has three layers of components: (a) consequences, (b) logic, and (c) construction. The consequences layer articulates key properties and objectives with an eye toward possible unintended effects that may reduce an assessment's utility. The logic layer is where the assessment design is conceptualized in relation to the desired Theory of Action (ToA) specified in the consequences layer. The construction layer specifies the machinery used in assessment development. The goals are to operationalize the argument specified in the logic layer and the ToA specified in the consequences layer. We use double-sided arrows to indicate the interactive nature across layers.

19.2.1.1 Consequences

The consequences layer is located at the top of the model. We discuss consequences in relation to the ToA. Bennett's (2010) application of ToAs to assessment highlights the importance of explicitly identifying the assessment components, the claims that will be made from the results, the action mechanisms designed to lead to the desired effects, and the identification of potential unintended negative consequences and what will be done to mitigate them. Bennett suggests that a ToA is needed when assessments are viewed as instruments of change, so that designers do not focus only on the instruments' technical adequacy but also consider consequences of using assessments.

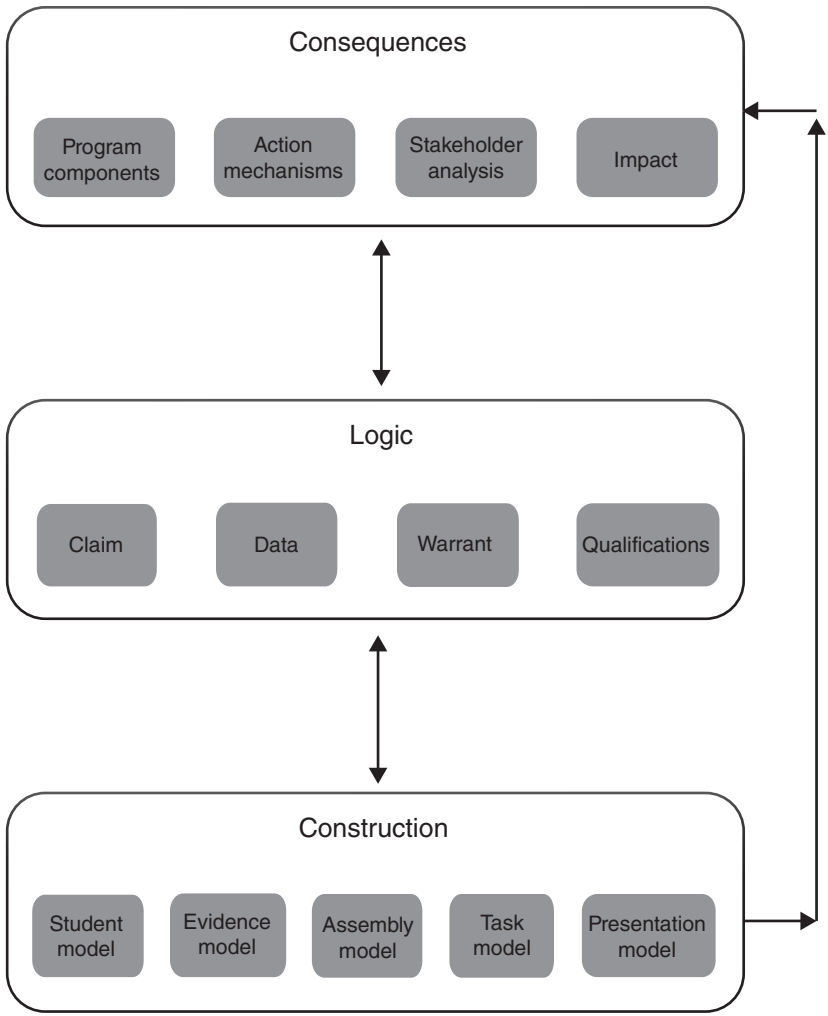

Figure 19.4 A multilevel design model for assessment development and use

Consistent with the ToA, we suggest that explicitly articulating beliefs underlying consequences involves identifying the components of an assessment, action mechanisms, stakeholder analysis, and impact. The program components include the items and scores obtained from the assessment, services designed for test-takers, and services designed for score users. The action mechanisms capture the types of decisions, behaviors, and solutions

expected from different stakeholders using scores for decision-making. Stakeholder analysis components include anticipating interpretation and use of scores by varied populations, from policymakers to parents. Finally, impact includes intended and unintended direct and indirect consequences of the assessment (Oliveri, Rutkowski, & Rutkowski, 2018). To help minimize short- and long-term unintended effects of decisions informed by scores, we suggest the use of various types of data sources, as well as considerations that include the degree to which the assessments cover the appropriate constructs and acknowledge test-taker diversity in interpretive arguments. Otherwise, potential unintended effects may occur, including over remediation (see Figure 19.1), as well as retaining and graduating a particular student subgroup of students more frequently than another (see Figure 19.2).

19.2.1.2 Logic

The next layer in the model is logic, which Toulmin (2003) describes as a claim, data, warrant, and qualification. With modification, Toulmin's model is useful for constructing validity arguments (Bachman & Palmer, 2010; Kane, 2006). An example of an assessment-use claim may be the use of scores from a writing test to inform higher education decisions, such as readiness to take courses that are more advanced. The data may involve providing evidence, which may range from a test score to the use of additional variables beyond test scores to inform inferences on students' readiness to pursue more advanced college courses.

The warrant supporting the claim may be that "the test" is designed to assess written proficiency, which might be measured using proficiency indicators such rhetorical knowledge, critical thinking, writing processes, and knowledge of conventions (Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project, 2011). The test may be designed to resemble the types of written compositions students need to carry out in the more advanced college courses. The qualifications (or qualifiers) may range from weak to strong. Qualifiers are strong, and call for more caution in interpretation, if there is more distance between the evidence (e.g., writing assessed through multiple-choice questions focusing on knowledge of conventions) and the skills students need (e.g., rhetorical knowledge) to have to complete their courses. Conversely warrants are stronger when the skills assessed are closely connected to course content (e.g., writing assessed by demonstrating proficiency in the use of rhetorical structures in various academic and workplace genres across a more representative construct domain).

Rebuttals to claims may arise when test-takers are tested using tasks that are unfamiliar, or irrelevant to the construct or domain assessed. Their use introduces inferential errors for construct-irrelevant reasons. Considering these issues during assessment design is important to identify validity and fairness issues early on and develop more-appropriate interpretive inferences and validity arguments for all students (International Test Commission, 2018; Mislevy, 2018; Oliveri & Lawless, 2018). To identify the types of intended and unintended consequences that might arise when developing assessments, it is important to work through the consequences and logic layers of the MDM prior to test construction. Examples of unintended consequences potentially leading to unintended outcomes may include: (a) under-representing the constructs needed for college success by only assessing cognitive constructs and leaving out noncognitive constructs, (b) emphasizing elements of the writing construct that focus solely on easily capturable features such as conventions that, in turn, suggest narrow views of writing, (c) failing to identify tasks that authentically align with the assessment's language-use domain, and (d) developing test items that use technology that is unfamiliar to test-takers.

19.2.1.3 Construction

Construction is the final layer of the model. As indicated, there is an interactive connection between the consequences and logic layers. This connection reflects the necessary linkage between the key elements of the construction and the other two layers, such as the inferential limits of data use and considerations for the populations that comprise the test-taker population. The five elements that belong to the construction layer are consistent with an ECD model and include conceptual models that describe technical specifications of the assessment and considerations relative to the student, evidence, assembly, task, and presentation components of the assessment. The student model identifies variables for the knowledge, skills, or other attributes of the construct measured by the assessment. The evidence model provides information about how the student model variables should be updated given student performance, as captured in the form of work products. The assembly model describes how the student, evidence, and task models work together to form the assessment. The task model describes how to structure the kinds of situations that allow evidence to be obtained for analysis. The presentation model describes how the tasks appear in various settings, thus providing a test specification for organizing the material to be presented and captured (Mislevy, Almond, & Lukas, 2004; Mislevy & Haertel, 2006).

19.2.1.4 Reflection on the MDM

The MDM includes useful considerations for the design, development, and uses of tests appropriate in higher education and are particularly needed when developing assessments for complex constructs (e.g., critical thinking, collaboration, or interactive communication). To maximize the desired, intended consequences from their use and to minimize undesirable and unintended consequences, developing assessments requires collaborations across stakeholder groups to design constructed-response tasks and link them to the assessment-based claims. Moreover, effective assessment development may involve the construction of a variety of linked tools and tests to produce a comprehensive assessment system capable of providing multiple and varied forms of evidence to support higher education decisions. Such considerations would apply to construct conceptualization, task development, and interpretive materials to guide the decision-maker. As we explain next, some of these types of assessments already exist.

19.3 A Complementarity Design Model for Assessment Development and Use

Breland et al. (2002) suggest that a single assessment may be insufficient to address all stakeholders' needs, particularly as the assessment purposes increase due to the desire to assess an expanded skill set for more diverse populations. The complementarity model we are about to describe is best seen as within the family of integrated assessment systems (IAS). An IAS may enable a more meaningful alignment of data from higher education assessments that serve different purposes and share common goals. An IAS's goal is to provide additional information about students in local contexts (e.g., within higher education institutions) to support learning by providing students and instructors with ongoing feedback. An IAS may include familiar, large-scale assessments (e.g., ACT®, SAT®, SweSAT, TOEFL®) used by multiple institutions to provide information across locations, users, and populations and compare individuals from different backgrounds on common items to inform higher education decisions. An IAS may also include locally administered assessments, such as intelligent tutoring systems or curriculum-based, computer-delivered tasks.

An IAS differs from a classic selection paradigm that uses an outcome (e.g., first-year grade point average) and predictor variables (e.g., admissions test scores) assumed to be linearly related in the full population of applicants based on an instructional program. In the United States, the classic selection paradigm was a starting point for many institutions, as it

provided the foundation for the use of admissions tests. An IAS also differs from a placement paradigm, which allows multiple treatments (e.g., grades from different courses) to seek optimal placement of individuals (Cleary, 1968; Novick & Petersen, 1976). Instead, an IAS uses data from different assessments to evaluate students' course-taking patterns to provide them with feedback within and across courses. Such feedback may be used to evaluate students' performances more frequently to better inform their course placement decisions, such as which courses to take next, based on how others with similar course and covariate backgrounds fared.

Bayesian inference networks may be used to link data from different assessments. They may be built around an institution's population and data (Braun & Jones, 1984). The goal is to analyze how students with different data patterns fare under various placement decisions through time while allowing for potentially missing data as students may not have available information on all variables.

Making decisions more rapidly, with the student as decision-maker (to increase students' involvement and agency in the decisions made), are perhaps best when interwoven through courses. With shorter feedback cycles, the traditional notions of placement and instruction through a course become blurred. The notion of placement blends into the notion of supported, individualized, instruction – made possible by learning frameworks of modules/experiences that can be tailored to individuals in relation to content, timing, and intensity. The idea of individualized feedback based on localized assessment is not new. The Individually Prescribed Instruction project, grounded in behavioral psychology and using criterion-referenced tests, was implemented in schools in the Pittsburgh, Pennsylvania area of the United States in the 1960s and 1970s (Glaser & Nitko, 1970). More recently, Pane et al. (2017) examined pedagogies based on personalized learning in 40 institutions dedicated to personalized learning-based instruction and concluded that there is evidence that implementation of personalized learning practices may have positive effects on achievement. As we illustrate through examples of CMs, such systems may incorporate and balance the strengths of formal and informal assessments by capitalizing on an array of conceptual, methodological, and technological developments.

In addition, these papers challenge the testing industry to develop assessment systems that can capture evidence of student learning at multiple time points, from different sources (i.e., inside and outside of school settings) and different types (i.e., quantitative and qualitative), and that allow for resonance with the teaching, learning, and assessment processes.

19.3.1 Model Overview

In agreement with the statement of Breland et al. (2002) that a single test is insufficient to meet all stakeholders' goals, the CM to which we now turn provides an expanded space for assessment uses to support not only admissions decisions but also student retention and graduation. Figure 19.5 shows the CM. Four types of assessments fill the figure's quadrants. The selected assessments are administered either by colleges or in high school for college credit. The assessments are independent of each other and do not act as a system; the aim of Figure 19.5 is to identify elements of complementarity in order to achieve an IAS. In the figure, summative and formative assessments are on the y-axis; distributed and local assessments are on the x-axis. In the figure's center is "complementarity" to show that the assessments can serve various purposes through their balance of local utility, portability, formative, and summative types of score uses. The assessments were selected for their desirable features to meet stakeholders' needs, such as providing interactive experiences that more closely align with skills students need in their future careers and expanding opportunities for students to learn in more meaningful and contextualized learning situations.

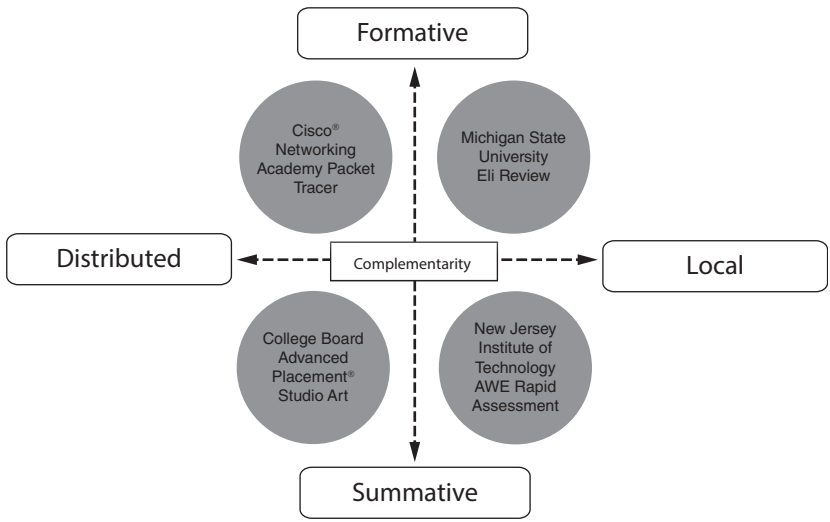

Figure 19.5 A complementarity model for assessment development and use
Note. AWE = automated writing evaluation.

19.3.1.1 Cisco® Networking Academy (CNA)

The CNA illustrated in the upper-left quadrant shows the complementary relationship between formative and distributed assessments with its use of a simulation tool called Packet Tracer software (Cisco Systems Inc., 2010). The CNA, designed to help students develop skills in beginning network engineering, emphasizes both formative and distributed aspects of assessment: It provides learning materials, assessment tools, and a social network connecting local CNA classrooms globally (Frezzo, Behrens, & Mislevy, 2009). The software's environment enables students and teachers to construct, configure, troubleshoot, and share computer network simulations. Web-based-delivered exercises and embedded assessments provide students with unlimited opportunities to engage in the interactive experiences central to learning how to think and act like network engineers.

Packet Tracer tasks are open-ended in operation but standardized with respect to interfaces, standards, and evaluation methods; thus, the use of Packet Tracer is distributed throughout the CNA community. Students can engage with the tasks individually, at times and places that the instructors or the students themselves determine; this use is local and formative. Extended tasks are used in course assessments and for learning (often by students working collaboratively, again locally and formatively, but as applied with distributed substance and standards). Shorter facsimiles of Packet Tracer-like tasks appear in the Cisco Certified Network Associate (CCNA) Exam that many CNA students take; a use that is distributed and summative but retains the environment, substance, and standards of the formative use of Packet Tracer in learning environments. CNA leverages technology and elements of standardization to create assessments that advance learning, allow for considerable local adaptation, and provide evidence of a common set of valued skills for subsequent college or career decisions. Successful completion of the sequence of four CNA courses, for example, provides admissions officers, placement counselors, and employers with information about a student's networking experience. A CCNA certificate provides even stronger evidence of proficiency, whether it was developed in a CNA course or otherwise, due to the layering of standardized and monitored testing conditions on top of the evidence gathering in the common networking domain.

19.3.1.2 Advanced Placement® Studio Art

First introduced in 1955, the College Board's Advanced Placement (AP®) examinations emphasize distributed and summative elements (see the lower-left quadrant of Figure 19.5). Although AP assessments are not

administered in higher education, we include them in this chapter because they have implications in higher education as AP scores are used to award course credit or to allow students to place out of certain college requirements and thus allow them to take more advanced courses earlier in their college career (College Board, n.d.).

In most AP subject areas, students prepare for a common end-of-course examination based on a course-specific syllabus and other curriculum requirements. Thus, these examinations are both distributed and summative. Most students participate in AP as part of local high school classes, although the examinations are open to all. Instructional methods, curricula, and local assessments are customizable for local and formative uses within the boundaries set by the College Board. Scores from the examinations provide valid and credible information to college admissions officers and departmental counselors about students' capabilities in particular subjects (College Board, n.d.).

AP Studio Art courses differ from other AP subject assessments because they represent more localized approaches. If AP Studio Art appeared separately in the CM, it would be closer to the "local" quadrant than would the other subjects. There is no culminating written examination for Studio Art courses. Instead, students submit portfolios of works they created throughout the school year for central evaluation. AP Studio Art offers locally situated experiences for students that require them to produce a certain number and type of artifacts evaluated using a centralized rating. These aspects make central evaluation possible in support of their high-stakes distributed use. The course design balances support for localized student learning and provides information that helps college personnel evaluate students' accomplishments (Myford & Mislevy, 1995).

19.3.1.3 Eli Review

Construct-specific digital ecologies are now being used to improve students' writing (Hart-Davidson & Meeks, forthcoming). Within a web-based environment, instructors and students provide specific and explicit feedback on student writing using rubrics tailored to assignments they create individually or shared within or across institutions. In turn, this feedback leads to improved instructor and peer reviews enabling writing program administrators to make evidence-based curricular changes by assessing student learning at the class, student, and program levels. It also enables the identification and subsequent support of students at risk, to improve retention and graduation rates. Platforms such as Eli Review, created at Michigan State University, as well as Peerceptiv, and Peergrade.io, are examples of such systems that focus on student feedback.

These systems exemplify formative and local uses of higher education assessments. They also illustrate distributed uses, as they are part of a family of other web-based assessments. They illustrate the complementarity between local and distributed views that can occur when information is gathered across sites within a digital environment. While assignments and course grades are within the domain of summative–instructor judgment, formative feedback is associated with both curricular change and positive outcomes for students. As such, resonance is established between summative grading and formative feedback to better support admitted students' learning.

19.3.1.4 Criterion®

The use of automated writing evaluation (AWE) technologies has led to a comprehensive body of knowledge associated with both summative judgment and local, formative feedback in various educational settings (Shermis & Burstein, 2013; Shermis et al., 2016). An example is Criterion (see the lower-right quadrant of Figure 19.5), which is an online writing evaluation service and a machine-scored, web-based writing tool that helps students plan, write, and revise their writing (Burstein, Chodorow, & Leacock, 2004). It represents local and distributed uses and can be useful in providing a rapid assessment of students placed in remediation classes to reduce unwarranted remediation by identifying students in need of additional instructional support.

The use of AWE in rapid assessment (i.e., to collect information about students' knowledge or skills prior to designing an intervention) reveals how local adaptation can be linked to scores often associated with summative assessments in a formative way. The results of a study with undergraduate students (N = 1,482) at the New Jersey Institute of Technology (NJIT), a public technological research university in the United States, indicates that Criterion offered a defined writing construct congruent with established models (Klobucar et al., 2013). It achieved acceptance among students and instructors and showed no statistically significant differences between ethnic and racial groups of sufficient sample size. It also correlated at acceptable levels with other writing measures, performed in a stable fashion, and enabled instructors to identify at-risk students to increase their course success. NJIT utilized the automated system to gain real-time information about student writing performance and aligned its local efforts with trends calling for decreased remediation of qualified students in basic skill areas of English. Rather than use scores to remediate students, the institution adopted the philosophy that an admitted student

was a qualified student and used the AWE technology to leverage resources to students enrolled in credit-bearing courses. This novel AWE use demonstrates how a system that was designed to be distributed can be tailored to a local setting in valid, reliable, and fair ways.

Although it would have been impossible to have read hundreds of essays twice in the first week of class to produce evidence of inter-rater reliability without exhausting the instructional staff, the AWE produced scores in real-time. Because maintaining diversity is key to the mission of the institution, the scores could be examined before use to ensure that group differential impact was not evident. Therefore, while AWE rapid assessment is depicted in the lower-right quadrant of Figure 19.5, it is important to see how summative scores can be used formatively. It is equally important to see how assessments that are designed to be distributed can, with attention paid to categories of evidence, be used locally to advance opportunities for students by removing barriers and increasing opportunity to learn.

We highlight that Burstein et al. (2018) modified the AWE scoring engine within Criterion to develop the Writing Mentor™ application. Writing Mentor is a Google Docs add-on designed to help students improve their writing by obtaining real-time, formative feedback. It provides feedback using natural-language processing approaches and linguistic resources according to a defined model of the writing construct that includes the use of sources, claims, and evidence; coherence; and knowledge of English conventions. Because this new technology provides individual feedback in relation to content, timing, and intensity, Writing Mentor may be described as a form of personalized-learning experience.

19.3.2 Reflection on the CM

The above examples illustrated how assessments can be used to meet various purposes, expand the construct being measured, and leverage additional data sources to inform varied decisions, such as course placement, remediation, and support for student learning with ongoing feedback. Kane and Mislevy (2017) also suggest augmenting information from summative tests because although they provide much-needed information about achievement, they provide limited information in terms of how test-takers perform various tasks. Therefore, summative assessments are of limited value to inform instructional planning or support student learning. They propose the use of intelligent tutoring systems or formative assessments that use a cognitive diagnosis model connected to curricular

outcomes and instructional units as an optimal way to select instructional options and design curricula. Ercikan and Pellegrino (2017) also describe the possibilities associated with integrating process-model interpretations (e.g., understanding test-taking behavior such as eye movements, mouse clicks, and time on task) to draw additional inferences about student learning that are extractable from digital assessments. Currently, such information is only starting to be used for test validation, but its use in formative assessments and digital-learning environments – both lower-stakes applications – is a topic of interest in the learning analytics community. These approaches may help minimize the types of unintended consequences of using summative assessments, which may provide decontextualized information about students or may not comprehensively measure the constructs needed for college success.

19.4 Looking Forward

According to Holland (2008) in "The First Four Generations of Test Theory," the first three generations of testing involve acquiring increasingly sophisticated understandings and methods in the field of applied statistics. The fourth generation, now emerging, broadens its understandings to assessment as integrated into social systems, in many forms and roles to create and use assessments in a complex world of test-takers, teachers, policymakers, and institutions. The models we provided are our attempt to contribute to the use and score-based interpretations of assessments used for diverse purposes beyond admissions to placement, course grading, and providing formative feedback to students to support learning. We draw on: (a) a sociocognitive perspective for learning and the reconceived roles assessment can play (Mislevy, 2018); (b) a sociocultural perspective on learning and assessment in increasingly diverse populations (Oliveri, Lawless, & Mislevy, 2019); (c) a sociopolitical perspective on educational systems (Feuer, 2013); and (d) a philosophical position of assessment design and use as applied ethics (Elliot, 2016; Mislevy & Elliot, forthcoming). We used these integrative perspectives to bring out the deeply interrelated nature of forms of assessment and the consequential basis of our actions. We suggested how we might do more than design assessments for sequestered selection problems; as well, we have demonstrated how coordinated assessment practices that encompass selection, placement, and within-course guidance can be used together to optimize both students' and higher education's educational goals.

We ask how institutions and sponsors of assessments would behave differently if they used an alternative (formative assessment) paradigm in

which assessments are used to inform student learning in lower stakes environments, as our proposed models suggest. It is clear that changes are already occurring in the United States. For instance, the Idaho State Board of Education (2015) began working with Compete College America, a nonprofit organization whose mission is to eliminate achievement gaps by providing an equality of opportunity for college completion. Since 2014, institutions have provided for-credit options for underprepared students in various ways, ranging from offering co-requisite models (concurrently delivered remedial instruction) to emporium models (computer-lab-delivered instruction). Idaho's example illustrates a range of responses when admissions and placement are considered as complementary processes. In some cases, however, budgets are cut to reduce remediation without equal commitment to providing instruction for students in need. And, in other cases, necessary instruction is provided without the benefit of distributed assessments, which would allow precise information on student ability to be provided to administrators and teachers.

Across these cases, we suggest that the use of the MDM would yield benefits for varied stakeholders, from students to assessment sponsors. The model highlights the need to account for consequences and fairness in the initial stages of assessment design, without forfeiting the emphasis on the logic of the interpretation and use-arguments and the evidential basis for test construction. Similarly, the CM offers value by bridging diverse types of assessments under a common goal. In the proposed expanded assessment space, measurement innovation is more likely to be realized in a way that better suits local needs by augmenting the available data-based information to enrich connections between high school and higher education institutions to better support instructional guidance.

For assessment developers, the future would be based on designing, validating, and using varied forms of assessments and providing information that could result in both selection and subsequent success. For students, the educational environment could be structured to provide a continuum of actionable pedagogies aligned to academic and workplace demands (see Burrus, Way, Bobek, Stoeffler, & O'Connor, in this volume). We also envision the possibility of eradicating the term "remedial," as new forms of assessment with rapid feedback cycles would help students know in detail which skills and what level of complexity are needed for success. Even if the proposed benefits of the models are conjecture, we believe there is reason to think that change is possible. We close by providing three innovations that may expand higher education assessments to better meet stakeholders' needs.

19.4.1 Progress in Measuring Noncognitive Skills

Currently, educational measurement has advanced in several ways that will continue to influence higher education assessments. One example is the inclusion of noncognitive skills (e.g., intrapersonal and interpersonal domains) into assessments, which would be relevant to higher education admissions (see Niessen & Meijer, in this volume) as well as workplace success (see Burrus, Way, Bobek, Stoeffler, & O'Connor, in this volume). Although the quality of current measures and their use in consequential decisions have been open to doubt, new advances in how to better assess these types of skills in lower-stakes situations continue (Kuncel, Tran, & Zhang, in this volume; Oliveri, McCaffrey, Ezzo, & Holtzman, 2017).

19.4.2 Innovation in Task Design

Another example of innovation is task design. Oliveri (2018) describes the development of assessment prototypes using scenario-based tasks of twenty-first-century (e.g., communication and collaboration) skills. The research prototype is designed formatively to assess students' ability to communicate and collaborate in workplace-relevant contexts. Because the prototype was developed to be aligned with the Occupational Network's content model (Occupational Information Network Resource Center, 2017), its tasks align with workplace activities (Oliveri & McCulla, forthcoming). The tasks also align with college curricula potentially to inform classroom instruction in support of student learning. This example illustrates the complementary relationship between formative assessment and the use of culminating tasks and summative grades that may be integrated within an IAS.

19.4.3 Technological Advancements

Technological advances also open up possibilities for assessments by using tasks that are adaptive, immersive, interactive, and customizable to students' backgrounds and capabilities, and that can be delivered in test centers, online, or integrated into learning systems. Such tasks have a potential to enhance the skills measured by distributed assessments, used locally by higher education institutions, and embedded in learning environments for shorter feedback cycles. The assessment of writing using AWE systems exemplifies these innovations, as they provide rapid feedback to students and instructors to help them identify what kinds of supports are

needed. Such computer-assisted feedback allows instructors and their students to make decisions more frequently and provide learning experiences that more closely match students' needs, thus supporting retention.

Advances in data science, such as using optimization algorithms from operations research, allow institutions to choose from a range of predictor variables and specify a targeted balance across a range of desirable, and often competing, outcomes (see Zwick, in this volume). Moreover, statistical models such as Bayesian modeling may "borrow strength" across institutions to improve score-based decisions, which can enable smaller institutions to build admissions and placement models tailored to their populations and decision environments while leveraging information from empirical patterns in other institutions. The creation of shared databases across institutions may allow for improving higher education decisions by rendering innovative methodologies feasible.

19.4.4 Reflection on Beginning Again

We began this chapter with an overview of changes in the higher education environment, as to the constructs assessed and the composition of test-taker populations. In response, we offered an MDM to guide assessment design, development, and use that provides data at various levels (student, institutional, and state) to meet stakeholders' needs. We also described a CM that integrates data from assessments used for various purposes to improve uses of data from assessments in ways that support admitted students' learning. We described existing assessments that illustrate a complementary relationship across assessment uses. We also suggested that a similar design approach, combined with our evidence model, could be used to inform future assessment design and development efforts for use in informing higher education score-based decisions. We hope that our models will prove useful in the national and international assessment contexts described in this volume.

REFERENCES

Altbach, P. G., Reisberg, L., & Rumbley, L. E. (2009). *Trends in global higher education: Tracking an academic revolution. Report prepared for the UNESCO 2009 World Conference on Higher Education*. Paris: United Nations Educational, Scientific and Culture Organization. Retrieved from http://unesdoc.unesco.org/images/0018/001831/183168e.pdf.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for*

educational and psychological testing. Washington, DC: American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use the real world*. Oxford: Oxford University Press.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, 8, 70–91. https://doi.org/10.1080/15366367.2010.508686.

Bereiter, C., & Scardamalia, M. (2012). *What will it mean to be an educated person in mid-21st century?* Princeton, NJ: The Gordon Commission on the Future of Assessment in Education. Retrieved from https://gordoncommissionblog.wordpress.com/commissioned-papers/what-will-it-mean-to-be-an-educated-person-in-mid-21st-century/.

Braun, H. I., & Jones, D. H. (1984). *Use of empirical Bayes methods in the study of the validity of academic predictors of graduate school performance* (ETS RR-84-48). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00074.x.

Breland, H. M., Maxey, J., Gernand, R., Cumming, T., & Trapani, C. (2002). *Trends in college admission 2000: A report of a national survey of undergraduate admissions policies, practices, and procedures*. Princeton, NJ: ACT, Inc.; Association for Institutional Research; The College Board; Educational Testing Service; and National Association for College Admission Counseling. Retrieved from www.ets.org/research/policy_research_reports/publications/report/2002/cnrr.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion Online service. *AI Magazine*, 25(3), 27–36.

Burstein, J., Elliot, N., Beigman Klebanov, B., Madnani, M., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing Mentor™: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, 2, 285–313. Retrieved from https://wac.colostate.edu/docs/jwa/vol2/bursteinetal.pdf.

Chen, X. (2016). *Remedial coursetaking at U.S. public 2- and 4-year institutions: Scope, experiences, and outcomes* (NCES Report No. 2016-405). Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016405.

Cisco Systems Inc. (2010). Cisco packet tracer data sheet. Retrieved from www.cisco.com/c/dam/en_us/training-events/netacad/course_catalog/docs/Cisco_PacketTracer_DS.pdf.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x.

Coley, R. J., Goodman, M. J., & Sands, A. M. (2015). *America's skills challenge: Millennials and the future*. Princeton, NJ: Educational Testing Service. Retrieved from www.ets.org/s/research/30079/asc-millennials-and-the-future.pdf.

College Board. (n.d.). AP central. Retrieved from https://apcentral.collegeboard.org/.

Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project. (2011). Framework for success in postsecondary writing. Retrieved from www.nwp.org/img/resources/framework_for_success.pdf.

Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement*. New York, NY: Routledge. https://doi.org/10.4324/9781315774527.

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, 9(1). Retrieved from http://journalofwritingassessment.org/article.php?article=98.

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. New York, NY: Taylor & Francis. https://doi.org/10.4324/9781315708591.

Feuer, M. (2013). Validity issues in international large-scale assessments: "Truth" and "consequences." In M. Chatterji (Ed.). *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 197–216). Bingley: Emerald Group.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2009). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology*, 19, 105–114. https://doi.org/10.1007/s10956–009-9192-0.

Glaser, R., & Nitko, A. (1970). *Measurement in learning and instruction*. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Eds.). *Uses and misuses of data for educational accountability and improvement. The 104th Yearbook of the National Society for the Study of Education, Part II* (pp. 1–34). Malden, MA: Blackwell. https://doi.org/10.1111/j.1744-7984.2005.00023.x.

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*, 745-765. http://dx.doi.org/10.1037/pspp0000075.

Hart Research Associates. (2010). *Raising the bar: Employers' views on college learning in the wake of the economic downturn*. Washington, DC: Association of American Colleges and Universities. Retrieved from www.aacu.org/sites/default/files/files/LEAP/2009_EmployerSurvey.pdf.

Hart-Davidson, B., & Meeks, R. (forthcoming). Behavioral indicators of writing improvement: Feedback analytics for peer learning. In D. Kelly-Riley and N. Elliot (Eds.). *Improving outcomes: Disciplinary Writing, local assessment, and the aim of fairness*. New York, NY: Modern Language Association.

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.). *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht: Springer.

Holland, P. W. (2008, March). The first four generations of test theory. Paper presented at the Association of Test Publishers on Innovations in Testing, Dallas, Texas.

Hussar, W. J., & Bailey, T. M. (2017). *Projections of education statistics to 2025* (NCES Report No. 2017-019). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://files.eric.ed.gov/fulltext/ED576296.pdf.

Idaho State Board of Education. (2015). Governing policies and procedures: Section III: postsecondary affairs, subsection S: Remedial education. Retrieved from https://boardofed.idaho.gov/board-policies-rules/board-policies/higher-education-affairs-section-iii/iii-s-development-and-remedial-education/.

International Test Commission. (2018). ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations. Retrieved from www.intestcom.org/files/guideline_diverse_populations.pdf.

Kane, M. T. (2006). Validation. In R. J. Brennan (Ed.). *Educational measurement* (4th ed.) (pp. 18–64). Westport, CT: Praeger.

Kane, M. T., & Mislevy, R. J. (2017). Validating score interpretation based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.). *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 11–24). New York, NY: Routledge. https://doi.org/10.4324/9781315708591-2.

Kelly, P., Moores, J., & Moogan, Y. (2012). Culture shock and higher education performance: Implications for teaching. *Higher Education Quarterly*, 66, 24–46. https://doi.org/10.1111/j.1468-2273.2011.00505.x.

Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing*, 18(1), 62–84. https://doi.org/10.1016/j.asw.2012.10.001.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. London: Routledge. https://doi.org/10.4324/9781315871691.

Mislevy, R. J., Almond, R. G., & Lukas, J. (2004). *A brief introduction to evidence-centered design* (CSE Technical Report No. 632). Los Angeles, CA: The National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for Studies in Education, UCLA.

Mislevy, R. J., & Elliot, N. (forthcoming). Ethics, psychometrics, and writing assessment: A conceptual model. In J. Duffy & L. P. Agnew (Eds.). *Rewriting Plato's legacy: Ethics, rhetoric, and writing studies*. Logan, UT: Utah State University Press.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x.

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (CSE Technical Report No. 402). Los Angeles, CA: The National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for Studies in Education, UCLA.

National Research Council. 2012. *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13398.

Novick, M. R., & Petersen, N. S. (1976). Towards equalizing educational and employment opportunity. *Journal of Educational Measurement*, 13, 77–88. https://doi.org/10.1111/j.1745-3984.1976.tb00183.x.

The Occupational Information Network Resource Center. (2017). *O*NET 22.2 Database* [Data file and code book]. Retrieved from www.onetcenter.org/database.html.

Oliveri, M. E. (2018, April). Kitchen Design: A research prototype to assess communication at work. In O. Troitschanskaia (Chair), *Assessing student learning outcomes in higher education*. Symposium conducted at the meeting of the American Educational Research Association, New York, NY.

Oliveri, M. E., & Lawless, R. R. (2018). *The validity of inferences from locally developed assessments administered globally.* (ETS RR-18-35). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12221.

Oliveri, M. E., Lawless, R. R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing*, 19(1), 1–31. https://doi: 10.1080/15305058.2018.1543308.

Oliveri, M. E., Lawless, R. R., Molloy, H. (2017). *A literature review of collaborative problem solving for college and workforce readiness.* (ETS RR-17-06; ETS GRE RR-17-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12133.

Oliveri, M. E., & Markle, R. (2017). *Continuing a culture of evidence: Expanding skills in higher education* (ETS RR-17-09). Princeton, NJ: Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/ets2.12137.

Oliveri, M. E., McCaffrey, D., Ezzo, C., & Holtzman, S. (2017). A multilevel factor analysis of third-party evaluations of noncognitive constructs used in admissions decision-making. *Applied Measurement in Education*, 30, 297–313. http://dx.doi.org/10.1080/08957347.2017.1353989.

Oliveri, M. E., & McCulla, L. (forthcoming). *Using the occupational network database to assess and improve English language communication for the workplace* (Research Report Series). Princeton, NJ: Educational Testing Service.

Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). *Bridging validity and evaluation to match international large-scale assessment claims and country aims* (ETS RR-18-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12214.

Oliveri, M. E. & Wendler, C. (2017, April). Enhancing the validity argument of assessments: Identifying, understanding, and mitigating unintended

consequences of test use. Professional development workshop presented at American Educational Research Association, San Antonio, Texas.
O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.). *Language testing: Theories and practices* (pp. 13–32). Basingstoke: Palgrave Macmillan.
Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S., & Pane, J. D. (2017). *Informing progress: Insights on personalized learning implementation and effects*. Santa Monica, CA: RAND Corporation. Retrieved from http://rand.org/t/RR2042.
Riconscente, M. M., Mislevy, R. J., & Corrigan, S. (2016). Evidence-centered design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.). *Handbook of test development* (2nd ed.) (pp. 40–63). New York, NY: Routledge.
Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge. https://doi.org/10.4324/9780203122761.
Shermis, M. D., Burstein, J., Elliot, N., Miel, S., & Foltz, P. W. (2016). Automated writing evaluation: An expanding body of knowledge. In C. A. McArthur, S. Graham, & J. Fitzgerald (Eds.). *Handbook of writing research* (2nd ed.) (pp. 395–409). New York, NY: Guilford.
Snyder, T. D., de Brey, C., & Dillow, S. A. (2019). *Digest of education statistics 2017* (NCES 2018-070). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, US Department of Education. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018070.
Toulmin, S. E. (2003). *The uses of argument*. Cambridge: Cambridge University Press. (Original work published 1958). Retrieved from https://doi.org/10.1017/CBO9780511840005.
Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan. https://doi.org/10.1057/9780230514577.
Zwick, R. (2017). *Who gets in? Strategies for fair and effective college admissions*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674977648.