

2

Assessment

The Power and Potential of Psychological Testing, Educational Measurement, and Program Evaluation Around the World

William E. Hanson, Jacqueline P. Leighton, Stewart I. Donaldson, Thomas Oakland,¹ Mark D. Terjesen, and Craig Shealy

As Ebbinghaus (1908) famously quipped, “Psychology has a long past but only a short history.” The same is true of psychological assessment. It has a remarkably long past, dating back 4,000 years in China and Greece, yet a relatively short, formally documented history. Still, psychology’s roots are grounded in assessment, constituting a core competency and day-to-day professional activity. Psychologists are leaders and experts in assessment, with specialized knowledge in research design; statistics; and, of course, psychological testing, measurement, and evaluation. Added to this leadership and expertise, psychologists also have specialized knowledge in diversity and multicultural issues as is documented throughout *Going Global* as well as advocacy (see Chapter 1, this volume), mental health and well-being (see Chapter 4), and many allied areas, which makes them major contributors on the worldwide assessment stage. Psychology, assessment, and research go hand in hand, fully informing—and perhaps even defining—psychologists’ identities as scientist-practitioners. Thus, in meeting “a world of need,” it is important to consider the full power and potential of assessment globally.

We were honored to write a chapter on assessment. We were also overwhelmed, as it is a daunting task. A typical Google Scholar or American Psychological Association (APA) PsycInfo database search yields literally hundreds of thousands of assessment hits. Moreover, there are hundreds of popular assessment books and scientific journals, including Geisinger’s (2013) three-volume

¹Tom Oakland was tragically killed during preparation of this chapter.

We thank Gauri Bhardwaj for helping reconcile references and in-text citations.

<https://doi.org/10.1037/0000311-003>

Going Global: How Psychologists Can Meet a World of Need, C. Shealy, M. Bullock, and S. Kapadia (Editors)

Copyright © 2023 by the American Psychological Association. All rights reserved.

handbook series; APA's *Psychological Assessment*; and the *International Journal of Testing (IJT)*, which is the journal of the International Test Commission (ITC; <https://www.intestcom.org/>). Today, thousands of tests circulate the globe, some clinically focused, some not. Alas, we faced many questions: Where do we start, given the field's lengthy history? What do we cover, given its broad, far-reaching size and scope? And in keeping with the book's focus on exemplary practice, what models do we highlight? We simply cannot cover it all in this relatively short chapter. So, after much deliberation, we took a bird's-eye view of assessment, a proverbial "lay of the land" perspective.

To begin, we discuss assessment broadly, focusing on its centrality to psychology and research. We also discuss contemporary national assessment practices considering the "big four" specialties: clinical, counseling, school, and industrial/organizational psychology (I/O; see also Chapter 3 on consultation and Chapter 5 on leadership). The first section discusses basic competencies, including what applied psychologists do in the United States. In subsequent sections, we discuss three important international assessment topics: psychological testing, educational measurement, and program evaluation. We also discuss cross-cultural issues in assessment, including test translation and adaptation. Throughout, we highlight challenges, opportunities, and cutting-edge exemplars, including Therapeutic Assessment (TA), formative educational testing and item development, and transformative program evaluation. To conclude, we offer practical suggestions for developing global assessment competencies and participating in the international assessment community.

ASSESSMENT-BASED RESEARCH AND PRACTICE

Assessment is a broad, overarching term and a multifaceted process. Arguably, it is the sine qua non of psychological research and practice. Without it, psychology's worldwide contributions to education, business, mental health, public policy, and other areas would be substantively diminished.

Although many people believe assessment, particularly testing, is a static, reductionist endeavor, this is not—or should not be—the case. Rather, at its best, assessment is a dynamic, excitingly rich process. Assessment reveals a great deal about people, cultures, decision making, and data processing, and it is inextricably linked to psychological research (Haynes et al., 2011). As parallel processes, assessment and research involve overlapping steps. This includes identifying problems, reviewing relevant literature and client or organizational histories, collecting and analyzing quantitative and qualitative data, measuring constructs of interest (i.e., the intended objective of the test), making reliable and valid inferences and attributions, and, in many cases, preparing formal evaluations and recommendations while answering questions that could not be answered without assessment.

Given psychologists' extensive research backgrounds and training in psychometrics, they are well qualified, highly capable, and professionally able to assess

and evaluate, provided they know their limits; understand assessment theory, research, and practice; and appreciate how their beliefs and values affect their assessment models and methods, particularly across cultures. In fact, assessment and research/evaluation are core *functional competencies*, or necessary skills, of applied psychologists, and self-assessment/reflective practice and personal/professional cultural awareness and sensitivity are core *foundational competencies*, or building blocks, of what psychologists actually do (e.g., Geisinger, 2013; Rodolfa et al., 2005). Psychologists should therefore nurture both types of competencies. In global contexts, psychologists should, for example, have the skills and wherewithal to select, administer, and interpret culturally appropriate, psychometrically sound tests and tools; provide meaningful test feedback; and use multiple data sources, while addressing basic and applied questions in an ecologically valid manner. Given that many tests are developed in Western society and adapted and modified for use globally (Muthukrishna et al., 2020), it is imperative that practitioners choose measures wisely and offer recommendations that reflect the psychometrics of the test while being culturally aware of the contexts in which recommendations are made. As described next, that is why competent psychologists understand why and how “equivalency of meaning” is a key practice and standard for the translation and usage of tests in different cultures and contexts.

Globally, assessment practices and contexts vary considerably, from mental health and psychoeducational assessments, for example, to asylum evaluations and assessments associated with migration and immigration to large-scale assessment of international programs. In almost all parts of the world, assessments are occurring. And although they occur in many specialty areas of psychology (e.g., health psychology, community psychology), this chapter focuses on the big four areas of applied psychology: clinical psychology, counseling psychology, school psychology, and I/O psychology.

Assessment Specialists: The Big Four and Beyond

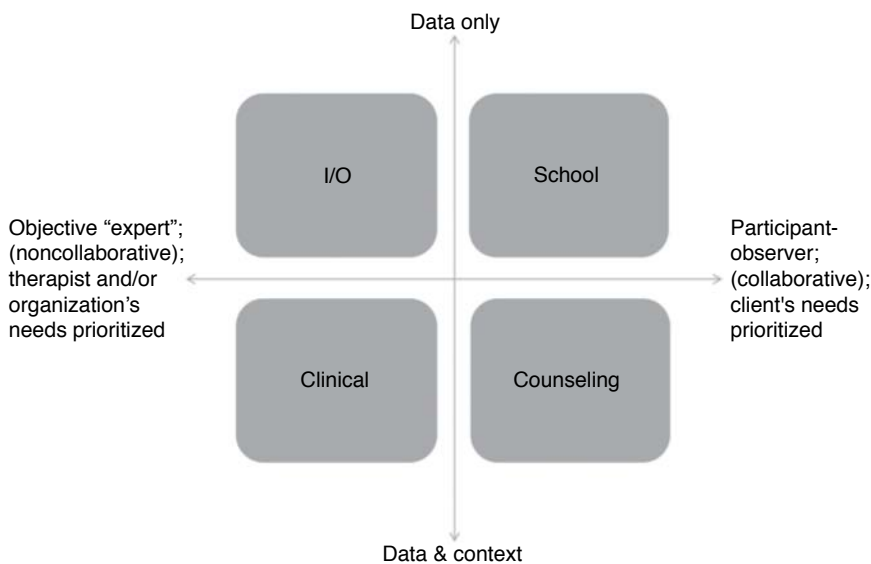
Assessment competencies and skills are particularly important for clinical, counseling, school, and I/O psychologists, affectionately known as the big four. Clinical and counseling psychologists work mostly with adults, typically in applied settings (e.g., private practices, hospitals, community agencies). School psychologists work mostly with children, adolescents, and parents and teachers, typically in school settings. I/O psychologists work mostly with adults in business settings. Many psychologists work across specialty areas (Health Service Psychology Education Collaborative, 2013). They also work across settings and age groups, providing a variety of assessment services. In the mental health field, for example, clinical, counseling, and school psychologists use multiple assessment methods, including behavioral observations, interviews, group and individual testing, historical background data, and corroborative consultations with others. Haverkamp (2012, 2013) classified clinical, counseling, school, and I/O psychologists along two axes: the extent of client–clinician

involvement/collaboration in testing (x axis) and the extent to which contextual data are considered in interpreting test results (y axis). As Figure 2.1 illustrates, psychologists vary across specialties.

Typically, mental health assessments describe current functioning, facilitate clinical hypothesis testing, assist in diagnosis and treatment planning, track/monitor client progress, manage risk and legal liability, and facilitate positive therapeutic change (Meyer et al., 2001). In non-mental-health capacities, I/O psychologists facilitate organizational and human resource decision making (e.g., employee selection, leadership development; see Table 2.1). Other specialty areas including, but not limited to, educational psychology, forensic psychology, and neuropsychology also engage in assessment research and practice. Examples include achievement testing, competency-to-stand-trial evaluations, and cognitive testing. To get a sense of the overall scope of psychological tests, readers are directed to the Buros Center for Testing (<https://buros.org/>), a prominent 90-year-old assessment and testing organization that classifies tests into 18 categories.

Assessment and psychological testing are not without criticism, especially early in their history. For example, in 1918 Yerkes ushered in intelligence testing en masse. The multiple-choice Army Alpha test determined “mental fitness” of literate World War I army draftees and, in turn, their prospective jobs (Gould, 1981). At the time, group-administered, multiple-choice standardized intelligence tests were new, efficient, and distinct from Binet’s individualized testing of French children. From 1917 to 1940 in the United States, the ease and

FIGURE 2.1. Global Assessment Competencies and Practices



Note. Based on Haverkamp (2012, 2013).

TABLE 2.1. Common Psychological Tests Used Among Clinical, Counseling, School, and Industrial/Organizational Psychologists

| Big 4 applied specialty areas | Psychological areas/constructs assessed | Commonly used tests |
|-------------------------------|--|---|
| Clinical | Neuropsychological, Intellectual & Personality Functioning/ Psychopathology | WAIS-IV, WMS, MMPI-3, MCMI-IV, Rorschach Ink Blot Test, WRAT-4 |
| Counseling | Intellectual & Personality Functioning, Career Interests, Strengths & Well Being, & Diversity Issues (e.g., ethnic identity) | WAIS-IV, MMPI-3, 16PF-5th ed., SII, SDS, Rokeach Values Survey, QoL |
| School | Intellectual, Academic Achievement, Social-Emotional Assessment, & Curriculum-Based Assessment | WISC-V, WIAT-III, BASC-3 |
| Industrial/ Organizational | Job Analysis, Leadership, & Employee Selection & Development | Cognitive/Mental Ability Tests, KSAOs, LMX-7 |

Note. These tests are used in highly diverse settings, including child/school, psychiatric, community mental health centers, rehabilitation hospitals, university clinics, and forensic units. BASC-3 = Behavior Assessment Scale for Children-3; KSAO = knowledge, skills, abilities, and other; LMX-7 = Leader-Member Exchange-7 Questionnaire; MCMI-IV = Millon Clinical Multiaxial Inventory-IV; MMPI-3 = Minnesota Multiphasic Personality Inventory-3; QoL = Quality of Life measure; SDS = Self-Directed Search; SII = Strong Interest Inventory; 16PF-5th Ed. = 16 Personality Factors Inventory-5th Edition; WAIS-IV = Wechsler Adult Intelligence Scale-IV; WIAT-III = Wechsler Individual Achievement Test-III; WISC-V = Wechsler Intelligence Scale for Children-V; WMS = Wechsler Memory Scale; WRAT-4 = Wide Range Achievement Test-4.

speed of intelligence testing helped open the door, unfortunately, to a eugenics revival (Marshall & Robertson, 2019; Reddy, 2008). Although lessons were learned, such as the need to exercise sound and ethical test use, misuses should not be forgotten, especially as they relate to international testing practices. For example, in the late 1920s, dubious forms of “mental proficiency” testing were used in the Canadian provinces of British Columbia and Alberta to justify the sterilization of Indigenous women under the Sexual Sterilization Act (Province of Alberta, 1928). That such profound harm was perpetrated with the aid of psychological tests is a reminder that “scientific” tools can be abused to justify human rights violations.

Fortunately, testing has come a long way over the past century, evolving and advancing significantly, thanks mostly to advancements in human ethics and to psychologists’ pioneering measurement concepts, such as *reliability* to quantify the consistency of test scores and *validity* to substantiate their interpretation, as well as improved norming and standardization procedures. That said, testing is not infallible; some amount of error always exists. The *Standards for Educational and Psychological Testing*, published simultaneously by the American Educational Research Association (AERA), the APA, and the National Council on Measurement in Education (NCME; 2014), is testament to present-day standards of test precision, rigor, and fairness, including validity evidence based on test/score content, response processes, internal structure, relations to other variables, and consequences.

INTERNATIONAL TEST USE AND DEVELOPMENT

Three factors affect test use internationally: (a) professional development, (b) presence of testing companies, and (c) a country's gross domestic product (GDP; Hu & Oakland, 1991). In a survey of 76 countries, Oakland et al. (2016) identified 606 tests used somewhat prominently in each country. These tests stood out and were used more than others. Intelligence tests (e.g., Wechsler scales, Ravens Progressive Matrices) and achievement tests were used most frequently, followed by measures of language, personality, neuropsychological qualities, school readiness, social-emotional, motor, and adaptive behavior. Although no group differences were found regarding intelligence testing, some countries (e.g., Argentina, Portugal, Latvia) used achievement, language, and motor skill tests more than other countries (e.g., Bulgaria, Cuba, Estonia). These same countries also used diagnostic, admissions, personal guidance, and career/vocational tests more. Finally, test use differed by age, with 6- to 12-year-olds tested most, followed by ages 12 to 18, 3 to 6, 19 and older, and finally 0 to 3.

In Oakland et al.'s (2016) survey, 11 professional groups were identified as using tests. Those who used tests most included, from high to low frequency, school and educational psychologists, clinical and counseling psychologists, general psychologists, school counselors, university researchers/professors, general education teachers, and special education teachers. Physicians, speech language therapists, nurses, and I/O psychologists rarely used tests in this sample. Professionals who have less academic training and experience were more likely to use outdated or obsolete tests. For example, some countries continue to use older versions of the Wechsler scales, some dating back to 1949, instead of the current version of the Wechsler Intelligence Scale for Children (WISC).

Recent survey results highlight several important international test use patterns. Going forward, it is important to consider variables that affect test use, and the role culture, language, politics, and economics play in the process. Much more research needs to be conducted and culturally appropriate practice guidelines developed. For example, training professionals to administer and interpret testing in a standardized manner may be more easily accessible to individuals in some countries than others. Furthermore, practitioners may find some tests do not take population characteristics or other cultural variables in their country into account, which may affect test norms, usage, and interpretation. Finally, the economic aspect of purchasing tests in countries with lower GDPs may preclude institutions and professionals from procuring updated versions of tests.

Importantly, since 1990, international test use has changed considerably, particularly regarding children, adolescents, and emerging adults. Several prominent programs and studies are underway currently, including the Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study, and Progress in International Reading Literacy Study (see the TIMSS and PIRLS International Study Center at <https://timss.bc.edu/>). These programs rely heavily on testing, primarily to assess educational

outcomes (Mateo et al., 2012; Pokropek et al., 2021; Rindermann, 2007). For practical reasons, such as the development, implementation, and refinement of policies, the evaluation and improvement of programs or systems, and the strategic allocation of resources, many countries need to know where their citizens are across multiple metrics. Consider the educational sector as one example. National educational attainment is associated with advances in science and technology and believed to make integration into the worldwide economy possible. Without assessment to appraise where the populace is at an educational level, from basic literacy to the mastery of sophisticated knowledge and skills, there is no way for a country to know where to go because it doesn't know where it is. This stark observation regarding education—just one of many societal sectors—illustrates the fundamental and ineluctable need to base policy and planning decisions on sound assessment data (e.g., Nishitani, 2020).

Many factors affect test development, including most prominently commercial marketing companies. For example, except for South Africa, companies that develop and market tests are uncommon in Sub-Saharan Africa. Test companies are also uncommon in the 22 Arab countries, a large region in which the testing industry is more or less absent. Most tests are developed in a few countries, including Canada, France, Germany, Italy, Mexico, Romania, Spain, and the United States (Oakland et al., 2013).

International Challenges, Opportunities, and Cutting-Edge Assessments

Many educational and psychological assessments have been adapted globally (Iliescu, 2017; Leong et al., 2016). To consider just a few examples, the WISC has been translated and adapted into multiple languages, with norms having been established for several countries, including France, Pakistan, Mexico, and Vietnam. A Spanish language version of the WISC also exists and has been adapted for international use. Similarly, many achievement batteries have been adapted and standardized globally. As an example, the Woodcock–Johnson Test of Achievement has been adapted and standardized for use in Australia and Brazil. Any adaptation and subsequent use of a measure needs to consider the varying degrees of cultural differences and linguistic variables that affect test performance. It is important, for example, for psychologists who administer these tests to determine whether differences in language or cultural background contribute to test performance difficulties over and above the presence of a true type of low-level outcome.

If tests developed in one country or culture are to be used in a different country and culture, important issues to consider include equivalency of meaning (e.g., constructs being assessed, item content) and underlying epistemologies (e.g., assessment and testing processes, interpretation of results). In their explication of a global psychology, Moghaddam and colleagues (2007) captured these concerns. In particular, outside the U.S. and Western contexts,

“alternative psychologies” include countries seeking to “tackle questions left unaddressed by mainstream U.S. psychology” while generating “locally relevant knowledge rather than refashion imported knowledge” (Moghaddam et al., 2007, pp. 180–181). Even so, if basic principles of reliability and validity are followed, and cultural sensitivity and “equivalency of meaning” are maintained, there should not be regionally defined distinctions (i.e., this is a test “from the West”). From this perspective, a “good” assessment measure is just that, whatever its origins, because it adheres to best practices of development, validation, implementation, and adaptation.

On that point, it is important to understand that translation of an extant measure into another culture and context may be necessary but by no means sufficient. For example, the goal really is not translation but, as noted, equivalency through a process of adapting a measure into a form in which the underlying meaning processes are consistent across cultures and contexts. Resources and organizations such as the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, n.d.), the ITC, and the International Organization for Cooperation in Evaluation (IOCE; <https://ioce.net/>) provide key guidance in this regard. Additionally, two issues of the *IJT* (global norming: Bartram, 2008; advances in test adaptation research: Gregoire & Hambleton, 2009) further promote an understanding of how such rapprochement may be pursued conceptually, methodologically, and practically, demonstrating, for example, that there may be more within- than between-group differences around the world (e.g., Wandschneider et al., 2015). Next, we highlight a cutting-edge example of culturally appropriate assessment and testing: collaborative/therapeutic assessment.

Collaborative/Therapeutic Assessment: An Exemplary Approach to Psychological Testing

Regardless of the test used, developed, or adapted, careful attention should be paid to matters of usage.² In the mental health realm, for example, Poston and Hanson (2010) concluded that “assessment procedures—when combined with personalized, collaborative, and highly involving test feedback—have positive, clinically meaningful effects” (p. 1033). To that end, collaborative/therapeutic assessment (C/TA) represents a potentially beneficial and empowering paradigm shift in assessment internationally, from a traditional authoritarian, therapist-centric, information-gathering approach to a flexible, client-centered, therapeutic approach—one that is highly collaborative, empathic, and culturally sensitive (Finn et al., 2012). C/TA uses a collaborative, semistructured, theoretically based assessment approach in which psychological testing helps clients understand and overcome persistent problems in living (Finn et al., 2012). Specifically, C/TA helps clients develop more coherent, accurate, compassionate, and useful “stories” about themselves and the world. Client narratives

²Contributions to this section were made by Stephen E. Finn.

are changed through a collaborative process in which the client and assessment specialist are participant-observers. Because the core values of C/TA are collaboration, respect, humility, compassion, openness, and curiosity, C/TA is ideally suited to international contexts. Thus, C/TA can be transformative in nature, particularly at the individual level, which suggests it is a commendable and ecologically valid approach to assessment research and practice.

Perhaps not surprisingly, C/TA is practiced in many countries, including France, Italy, Sweden, the Netherlands, Japan, Finland, Argentina, and Mexico. To take any number of examples in this regard, professional deliberations among C/TA scholars and practitioners have included assessors' moderating their expression of positive emotion in Sweden, introducing Italian psychotherapists to the potential benefits of assessment consultation, speaking plainly and bluntly to clients in the Netherlands, and dealing with the highly abstract and philosophical ambience of psychology in France. Such adjustments are standard TA practice. Although TA typically uses a semistructured format, each assessment is tailored to the unique client and their context. Additional information, worldwide training opportunities, and international centers are described on the C/TA website (<https://www.therapeuticassessment.com>).

GLOBAL IMPACT OF EDUCATIONAL MEASUREMENT AND ASSESSMENT

Measurement is "at the heart of every science for, without measurement, meaningful research, both basic and applied, is impossible" (Haynes et al., 2011, p. xi). Psychologists and allied professionals across the globe create metrics for assessing student achievement within, as well as between, countries for comparison purposes. From South America to North America, Europe, Asia, and Australasia, psychometricians are employed by government ministries to evaluate learning outcomes in K–12 students and guide them into postsecondary institutions. For example, since 1988, students in Chile have been administered the *Sistema Nacional de Medición de la Calidad de la Educación* (SIMCE) assessment in certain grades (Meckes & Carrasco, 2010). The SIMCE is a government test designed to ensure that students learn the Chilean-designed curriculum in reading, math, and science, among other content areas. Further, since 2003, the *Prueba de Selección Universitaria* is used to guide Grade 12 students entering the country's universities (Mateo et al., 2012).

Testing is not only administered to K–12 students but also used to monitor and enhance student experiences at postsecondary institutions. For example, the European Higher Education Area, made up of 49 countries, is focused on attracting students by ensuring the compatibility of an integrated network of postsecondary institutions. In 1999, the Bologna Process was catalytic in the genesis of the European Higher Education Area to fulfill the goals of enhancing (a) the appeal of attending postsecondary institutions in Europe; (b) the educational experience, including improved mobility and employability, based on a

system of accessible, understandable, and user-friendly undergraduate and postgraduate studies and degrees; and (c) the quality of their postsecondary education (Neave & Veiga, 2013). Likewise, the Aurora Network of European universities (<https://alliance.aurora-network.global/>) has adopted a multi-faceted Aurora Competence Framework, which relies on mixed-methods assessment to evaluate student learning processes and outcomes. This ambitious undertaking has many challenges, including assessing student outcomes or “qualification frameworks” and creating metrics to ensure the comparability of experiences at distinct institutions.

Although psychologists are often responsible for designing and administering tests to evaluate K–12 educational systems within countries, increasingly they also facilitate international tests to draw comparisons between countries. In these cases, international tests are not without controversy. For example, PISA provides country-by-country rankings of 15-year-old students. Large-scale tests such as PISA are controversial, in part, because there is debate about the nature of the schools that participate in the assessment and, thus, the sample of students compared across countries (e.g., such questions include whether participating students are from public or private schools and how much they represent the country’s educational system).

The cost of administering any kind of large-scale test also receives significant scrutiny. In the United States and some other countries, large-scale test administrations cost less than 1% of total government expenditures for publicly funded projects (Clarke, 2012). However, the situation is different in other countries, where costs are significantly higher. In these countries, there is intense pressure to institute full-scale assessment systems into educational infrastructures. *Assessment systems* are “groups of policies, structures, practices, and tools for generating and using information on student learning” (Clarke, 2012, p. 1), and large-scale testing programs, including national and international assessments, are important system-related tools. Using cross-country growth regressions, consistent interdependence between acquisition of knowledge and skills and GDP growth has been demonstrated (Hanushek & Woessmann, 2012). “Benchmarking” a country’s learning outcomes is vital to acquiring and maintaining a competitive edge in the global workforce. Along similar lines, Japan recently implemented a nationwide process of assessing the impact of international education in general and Collaborative Online International Learning or COIL in particular (e.g., the Institute for Innovative Global Education [<https://www.kansai-u.ac.jp/Kokusai/IIGE/>]).

Assessment systems also measure 21st-century skills. An example of this newfound focus is the Partnership for 21st Century Skills, sponsored by 32 organizations, including Apple Inc., Cisco Systems, Educational Testing Service, Microsoft Corporation, the U.S. National Education Association, and the Walt Disney Company. Such skills include foreign languages, creativity, and career and life skills, among others (e.g., Global Competency Standards for Students and Teachers). Binkley et al. (2010) outlined requirements for these assessments to be successful, such as aligning assessments with 21st-century

knowledge and skills, transparency of constructs designed to capture them, use of performance-based items to enhance score validity, technical precision of score reliability, selection of appropriate measurement models, and assurance of fairness to diverse student groups.

Locally and globally, educational measurement faces many challenges and opportunities. One major challenge is properly capitalizing on technology advancements and digital environments to improve testing because traditional paper-and-pencil tests are becoming obsolete in many parts of the world. Another challenge is securing test items and testing environments from potential cheating as high-stakes assessment becomes digitized and more susceptible to hacking. A long-standing challenge with increased globalization is test translation from one language to another, normally from English to another language, and test equivalence (Leong et al., 2016). Although there are several challenges, there are also opportunities, such as cutting-edge assessments, including computer-based and adaptive testing (Mills et al., 2002; see also Gershon, 2005); cognitive diagnostic testing (Embretson, 2009); the assessment of international, multicultural, and transformative learning (Wandschneider et al., 2015); and serious games and stealth assessments (Mislevy, 2010; Shute & Ventura, 2013), which reflect new waves of assessment design for summative and formative purposes.

Formative Educational Testing Advancements Around the World

To thrive amidst the profound global and economic changes in the 21st century, countries require skilled workers who can think, reason, solve complex problems, adapt to novel situations, communicate, and collaborate.³ Providing teachers and students with timely, detailed feedback is the norm, above and beyond a single final test score and performance evaluation. Thus, formative educational testing principles guide today's global assessment practices. Formative principles include any assessment-related activities—including administering tests more frequently—that provide process-altering feedback to teaching and other educational processes and, in turn, to improve desired outcomes.

With an increase in test usage comes requiring tests to be created more efficiently and economically, factors that are especially relevant in an international context. Fortunately, this requirement coincides with changes in educational technology, particularly computerized testing. Computerized testing supports and promotes formative assessment. For instance, computers permit on-demand testing, which allows students to take tests at any time and as often as they choose. Further, test items are scored instantaneously, thereby providing students with immediate feedback, and computers support the development of innovative item formats that allow educators to measure complex performances as well as a broader variety of knowledge and skills (van der Linden & Glas, 2010). The use of technology will provide greater access to testing for

³Contributions to this section were made by Mark J. Gierl.

students in remote settings and provides greater measurement precision. Of course, the drawbacks include access to device issues and the potential for the technology to unduly influence that format of the assessment.

Computerized testing can be expected to provide educators with local data as to student performance as well as to track and predict student performance over time and identify students who may warrant further intervention. As an example, computerized testing of reading performance at multiple points in kindergarten can be used to predict reading trajectories from first to fourth grade. In higher education, complementary research demonstrates that we can predict interest in, and satisfaction from, study abroad or multicultural education before a student has even taken his or her first class in university (Wandschneider et al., 2015). Despite these types of benefits, computerized testing can be challenging, particularly regarding large-scale item development (Downing & Haladyna, 2006) and automatic item generation (Gierl et al., 2021).

PROGRAM EVALUATION AND GLOBAL ASSESSMENT PRACTICE

Over the years, psychologists' expertise in assessment has been valued by scholars, practitioners, and leaders from a wide range of disciplines and professions outside of psychology. Program evaluation is one such area. Because psychologists design and evaluate social, educational, organizational, and community programs—which aspire to improve the human condition and make the world a better place—we can gather and analyze data to answer a wide range of questions, including the following: Was the program implemented well? Did the program achieve its goals? What works for whom, where, why, for what, and when (Gargani & Donaldson, 2011)? What was the overall impact of the program? Was the program cost effective? How can the program be improved?

Donaldson (2007, 2021) described common steps that evaluators go through to determine which questions to answer in a program evaluation: (a) determining the theory behind the program, (b) developing and prioritizing evaluation questions, and (c) designing evaluations to answer evaluation questions with credible and actionable evidence (Donaldson et al., 2015). More than 75 universities worldwide provide degrees and certificates in program evaluation, and more than 150 professional organizations provide evaluators with networking opportunities and professional development (LaVelle & Donaldson, 2021). A great deal of research shows that careers in program evaluation can be lucrative, fulfilling, and socially meaningful.

Psychologists evaluate programs in many countries and contexts. To consider just a few examples, applied social psychologists evaluate education, public health, drug abuse prevention, and a wide variety of workplace improvement programs and policies in the United States and abroad (Donaldson, 2021; Donaldson & Chen, 2021). Developmental psychologists enhance developmentally appropriate preschool services and education efforts. Health psychologists

determine the effectiveness of obesity prevention and health promotion efforts. Organizational psychologists design and evaluate leadership development programs, policies, and problems. Clinical, counseling, and school psychologists design and evaluate mental health programs and services. In fact, across specialty areas, psychologists design, evaluate, and implement international programs focused on improving human welfare.

The quest for credible, actionable evidence fuels demand for program evaluation services. Approximately 150 national and regional program evaluation associations exist today. Furthermore, groups such as Eval Partners, the IOCE, and the International Evaluation Academy facilitate the development of the global evaluation community. Although, historically, evaluation methods relied largely on quantitative methods, they are now more diverse, with additional design choices and methods, including mixed methods.

Importantly, psychologists share many values and interests with evaluators. The APA's (2017) *Ethical Principles of Psychologists and Code of Conduct* emphasizes that psychologists should be socially responsible, concerned for others' welfare, respectful of people's rights and dignity, and adhere to scientific standards and professional ethics as they apply their assessment and evaluation expertise. Psychologists are well positioned to engage in such activity because they possess substantive knowledge about the phenomena under investigation as well as how to acquire such knowledge, as well as strong critical thinking skills and solid foundations in matters of methodology, design, and data analysis (Donaldson & Crano, 2011; Mark et al., 2011). That said, a successful career in program evaluation requires more than knowledge of psychology and social science research methods, especially in a global context. When juxtaposed with self-other awareness and cross-cultural competence, program evaluation knowledge, skills, and theory enhance dramatically the program offerings built into education and training in psychology.

Despite the great resonance of program evaluation for the professional identity of psychologists, more needs to be done vis-à-vis education and training to prepare psychologists to assume these roles internationally (Donaldson et al., 2013). Ofir and Kumar (2013) outlined some of the common challenges faced by Western evaluators commissioned to evaluate programs internationally. For example, to do such work, program evaluators should use varied lenses to frame evaluation questions and designs. Such an approach involves being aware of the following issues: getting good and sustained results are more complicated and unpredictable; deep cultural differences often persist; there are usually capacity and other serious constraints; and evaluators must address significant privilege and power differentials (e.g., Donaldson et al., 2013).

Given such complexities, mixed-methods research (MMR) is becoming increasingly popular for international program evaluation, particularly in the United Kingdom, Sri Lanka, Germany, Japan, and South Africa (see Mixed Methods International Research Association at <https://mmira.wildapricot.org/>). Essentially, MMR involves collecting, analyzing, and integrating or "mixing" quantitative and qualitative data in a single or multiphase study. Although

complex, given the mixing of methodologies/paradigms, such an approach often enriches study findings above and beyond the use of quantitative or qualitative studies alone—and ultimately leads to better understandings, insights, and problem solving (Hanson et al., 2005; Mertens et al., 2010; Tashakkori & Teddlie, 2010; Wandschneider et al., 2015).

Transformative Program Evaluation

Transformative program evaluators view assessment and evaluation as contributing to positive social change, especially in marginalized communities (Mertens & Wilson, 2019).⁴ To ensure that human rights and social justice are prioritized, these evaluators use MMR to understand the complexity of culturally diverse communities. Transformative evaluators question the assumption that evaluation questions can be predetermined and instead support the engagement of the full range of stakeholders in ways that are culturally respectful to reach better understandings of the culture, context, and challenges to develop appropriate questions for study. Thus, evaluators do critical preliminary qualitative work—prior to developing questions—to understand the community, build relationships, and ensure that appropriate groups are included and supported in respectful ways. Overall, transformative evaluators strive to identify the nature of power relationships and the implications of those relationships in terms of whose voices are given privilege in the planning of the evaluation and the program.

A transformative evaluation would use both quantitative and qualitative data from interviews, observations, and “hanging out” with the targeted population, as well as epidemiological data disaggregated by relevant dimensions of diversity. These evaluations may help promote an understanding of such issues as power relationships, the effects of poverty, and other issues that need to be incorporated into plans for an intervention intended to create desired changes (Chilisa, 2012). In short, program evaluation designs that use MMR have considerable potential to enhance evidence-based practices designed to provide appropriate, effective, and sustainable community support (Mertens & Hesse Biber, 2013). In other words, this work can increase the potential for systematic inquiry and enhancing human rights and social justice globally.

DEVELOPING GLOBAL ASSESSMENT COMPETENCIES

As assessment expands globally, a scientific acculturation process occurs (Berry, 2013). Whether you are interested in psychological testing, therapeutic assessment, educational measurement, formative testing, program evaluation, MMR, or some other aspect of assessment, you are encouraged to learn from past mistakes, controversies, and challenges, and use lessons to enhance future

⁴Contributions to this section were made by Donna Mertens.

interactions and collaborative projects. This learning may be facilitated—and cross-cultural integrity (and associated competencies) developed and maintained—by embracing the following exemplary practices (outlined in Exhibit 2.1).

Become Familiar With Academic Resources and Professional Opportunities

You should familiarize yourself with both relevant coursework and with tests, readings, conferences, and online resources. In addition to obtaining excellent training and firsthand assessment, measurement, and evaluation experience, you are encouraged to acquaint yourself with the ITC and IOCE. These organizations lead the way internationally in terms of exemplary assessment and evaluation practices.

Likewise, a key resource and publication outlet for assessment professionals is the *IJT*. First published in 2001, it is the flagship journal of the ITC. The journal is thriving, with contributions from hundreds of scholars and dozens of countries around the world. Along with other top assessment and measurement journals, such as *Applied Psychological Measurement* and *Educational and Psychological Measurement*, *IJT* is highly diverse, in both content and international scholarly affiliations.

From graduate students to established professionals, attending conferences is another useful step. In program evaluation, for example, four of the largest conferences are the American Evaluation Association (AEA), Canadian Evaluation Society (CES), European Evaluation Society (EES), and Australasian Evaluation Society (AES). The AEA is an international professional association of evaluators devoted to the application and exploration of program evaluation, personnel evaluation, technology, and many other forms of evaluation, with approximately 8,000 members representing all 50 states and more than 70 countries. AEA, CES, EES, and AES are equally devoted to the advancement of program evaluation. These organizations bring academics, practitioners, and the public sector together, thus creating a forum for bridge building and collaboration.

EXHIBIT 2.1

Global Assessment Competencies Around the World and Our Top Nine Most Important Skills

Become familiar with academic resources and professional opportunities
 Know assessment theory
 Understand test translation issues
 Cultivate a “cultural self”
 Collaborate with key stakeholders
 Consider positives and assess/evaluate what is right with people
 Think transformatively
 Collect feedback and improve
 Own your epistemologies

Training and employment opportunities can be found through various sources, both locally and abroad. Within the United States, for example, current and future psychologists interested in such work may wish to review the activities of APA Division 5, Evaluation, Measurement, and Statistics; Section IX on Assessment of APA's (n.d.) Division 12 (<https://apadiv12secix.com>), Clinical Psychology; as well as other divisions (e.g., 52, International; 13, Consulting; 14, Industrial and Organizational). Another productive way to become involved in international assessment is to engage with large-scale testing programs, such as the Educational Testing Service.

Know Assessment Theory

It is essential that competent developers and users of psychological tests and measures understand theories of assessment and evaluation, from different types of validity, reliability, and attendant statistical procedures to what is meant by concepts such as "true score" and "item response" theory, "operationalization of constructs," and "response set" and "confounding variables" such as "social desirability," as well as countervailing perspectives such as "dustbowl empiricism." Without sufficient theoretical knowledge, test developers and users may inadvertently engage in assessment research and practice that is conceptually inadequate and inappropriately conducted (e.g., AERA, APA, & NCME, n.d.; Alkin, 2012; Geisinger, 2013; IOCE; ITC; Wandschneider et al., 2015). Relatedly, test users are encouraged to carefully consider their assessment philosophies and underlying assumptions: Who am I as an assessor? What are my core beliefs and values? What are my primary goals of assessment? What is my primary role? And perhaps most important, What is my preferred assessment paradigm, and why? Wiggins (2005) provided a nice, easily accessible overview of assessment paradigms, including empirical, multivariate, interpersonal, personological, and psychodynamic. Paying close attention to assessment theory and paradigms enhances one's intentionality, competence, and ultimately one's positive results when testing and evaluating people around the world.

Understand Test Translation Issues

As noted previously, it may be possible to adapt, evaluate, and apply tests and measures developed in one country and culture to another. However, as indicated earlier (e.g., see "International Challenges, Opportunities, and Cutting Edge Assessments"), appropriate guidelines must be understood and followed. Such guidelines offer practical guidance when engaging in the complex process of test translation from one country and culture to another, which involves—as noted earlier—much more than translating an instrument from one language to another. Determining equivalency of meaning across languages and cultures is of paramount importance.

Cultivate a "Cultural Self"

Professionals engaged in transnational assessment work are encouraged to cultivate their cultural self while attending closely to personal beliefs, values,

and worldviews and to better understand how these qualities affect their assessment and evaluation practices (Wandschneider et al., 2015). Such individuals acknowledge imposed and derived etics (i.e., wholesale application of Western concepts and measures to non-Western people and systems/organizations) as well as culturally appropriate emics (e.g., local, ingroup perspectives), value diversity and curiosity, are open-minded/committed to intellectual renewal, and immerse themselves in Indigenous cultures and traditions (Flores & Obasi, 2003).

Collaborate With Key Stakeholders

Collaboration is essential, in part because assessor–assessee power (and/or value) differences may be magnified in international contexts. Terjesen (2015) promoted the ABCs of developing global partnerships:

- **Adopt a shared vision.** That is, identify the key stakeholders and ascertain what they perceive to be the most important questions they want answered and what the implication of these results may be.
- **Be patient.** Like anything important, developing partnerships takes time, and there may be many detours throughout the process. Some of these detours can be anticipated, and having knowledge of the context that the partners are coming from may be helpful.
- **Collaboration, communication, and celebration.** Like any good partnership, the key is fostering effective collaboration through communication. Technology certainly has made possible collaborative global relationships that may have been a challenge 20 years ago, but even still, it is important to establish the best methodologies and patterns for collaboration. Finally, it is important to celebrate success throughout the partnership to provide intermediary reinforcement for the partners on their efforts at varied stages of the process.

Consider Positives: Assess and Evaluate What Is Right With People

Assessment and evaluation often focus on psychopathology and what is wrong, or “not working,” with a person or group/organization. These are important foci, but when working in an international context, a focus on positive human strengths and attributes (e.g., hopefulness, problem solving, subjective well-being) is also important (Flores & Obasi, 2003). Furthermore, if your research examines assessment results cross-nationally, we recommend caution in interpretation of differences: Difference does not mean deficiency.

Think Transformatively

Assessment and evaluation can be transformative, especially when done collaboratively, empathically, and in culturally sensitive ways. Such practices

are worth contemplating, if not modeling. Assessment and evaluation can assist the work of change agents that lead to the greater good and to advocate for others (e.g., the Summit Series at <https://summitx.org/>).

Collect Feedback and Improve

As you develop assessment- and evaluation-related experience and expertise, it is not okay to guess, or assume, you are “on track.” It is a foundational competency to practice what you preach. Thus, it is incumbent upon the international assessor to constantly collect feedback to determine whether—and to what degree—they are on track (Miller et al., 2020). Guessing and assuming run counter to ethical principles, standards of practice, and general scientist-practitioner sensibilities—a point that should become evident to anyone engaged in assessment work internationally. Additionally, assessment- and evaluation-feedback can be transformative in and of itself, especially when approached collaboratively and humanistically (Hanson et al., in press).

Own Your Epistemologies

The question of pure “objectivity” constitutes one of the more epistemologically contentious debates in assessment and evaluation. This fundamental concluding point is implicit, if not explicit, in all preceding points and relevant to the entire chapter because our underlying assessment-related beliefs and values are culturally mediated (Shealy, 2016). Even though we strive for objectivity and all of its worthy concomitants in assessment research and evaluation, it behooves us to remember three final *Going Global* take-home points. First, clinicians may not be less biased than the clients we evaluate. Second, worthy standards of objectivity are more aspirational than fully realizable. Third, the privileging of data over engagement may have unintended negative consequences. Through ongoing reflection, we hope the local and global assessments we render—and the care we provide—evidence humility and integrity in the “realities” we apprehend and facilitate growth, healing, and meaning making. Such a process involves an active and ongoing commitment to ascertaining whether one’s assessment-based conclusions and recommendations are ecologically valid, in that they not only resonate with clients but also are convincing to relatively impartial colleagues.

CONCLUSION

Assessment is a broad, globally relevant, and much-needed area of inquiry and practice—with tremendous power and potential. Without question, it has value and usefulness around the world. As should be evident, psychologists provide worldwide leadership in testing, educational measurement, and program evaluation, among many other applied areas. A prevailing theme is that

when done competently and collaboratively, with keen alertness and responsiveness to cultural relevance, assessment and evaluation transform people's lives. Pragmatically, assessment is not only unavoidable at multiple levels of analysis (from individuals and communities to societies and nations) but also essential for understanding better who we are and how we may be more responsive to each other and our increasingly interconnected world. As such, it is time to expand and revise our understanding of assessment by contemplating the full spectrum of implications and applications in education, clinical practice, government, and policy making. To the extent that psychologists work across borders, disciplines, and specialty areas, they can make meaningful and measurable differences in the lives of many. As these important collaborations occur, let us first "do no harm." In so doing, the legacy of assessment—and its capacity to change lives for the better—will be appreciated and affirmed, both locally and internationally.

REFERENCES

- Alkin, M. C. (Ed.). (2012). *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.). Sage.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (n.d.). *Open access files*. <https://www.testingstandards.net/open-access-files.html>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct* (2002, Amended June 1, 2010, and January 1, 2017). <https://www.apa.org/ethics/code/index.aspx>
- Bartram, D. (Ed.). (2008). Global norming [Special issue]. *International Journal of Testing*, 8(4). <https://doi.org/10.1080/15305050802434980>
- Berry, J. W. (2013). Global psychology. *South African Journal of Psychology*, 43(4), 391–401. <https://doi.org/10.1177/0081246313504517>
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2010). *Defining 21st century skills: The Assessment and Teaching of 21st Century Skills (ATCS 21) project* [White paper, draft 1]. University of Melbourne, Cisco, Intel, and Microsoft. https://oei.org.ar/ibertic/evaluacion/sites/default/files/biblioteca/24_defining-21st-century-skills.pdf
- Chilisa, B. (2012). *Indigenous research methodologies*. Sage Publications.
- Clarke, M. (2012, April 19). *What matters most for student assessment systems: A framework paper* (SABER Working Paper No. 68235). World Bank. <https://documents.worldbank.org/curated/en/2012/04/16238771/matters-most-student-assessment-systems-framework-paper>
- Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications*. Psychology Press. <https://doi.org/10.4324/9780203809730>
- Donaldson, S. I. (2021). *Introduction to theory-driven program evaluation: Culturally responsive and strengths-focused applications*. Routledge. <https://doi.org/10.4324/9780429353277>
- Donaldson, S. I., Azzam, T. A., & Conner, R. (2013). *Emerging practices in international development evaluation*. Information Age.
- Donaldson, S. I., & Chen, C. (Eds.). (2021). *Positive organizational psychology interventions: Design and evaluation*. Wiley-Blackwell.
- Donaldson, S. I., Christie, C. A., & Mark, M. M. (2015). *Credible and actionable evidence: The foundation for rigorous and influential evaluations*. Sage Publications. <https://doi.org/10.4135/9781483385839>

- Donaldson, S. I., & Crano, W. C. (2011). Theory-driven evaluation science and applied social psychology: Exploring the intersection. In M. M. Mark, S. I. Donaldson, & B. Campbell (Eds.), *Social psychology and evaluation* (pp. 141–160). Guilford Press.
- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Lawrence Erlbaum Associates.
- Ebbinghaus, H. (1908). *Psychology: An elementary textbook*. Heath.
- Embretson, S. E. (Ed.). (2009). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.
- Finn, S. E., Fischer, C. T., & Handler, L. (Eds.). (2012). *Collaborative/therapeutic assessment: A casebook and guide*. Wiley.
- Flores, L. Y., & Obasi, E. M. (2003). Positive psychological assessment in an increasingly diverse world. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 41–54). American Psychological Association. <https://doi.org/10.1037/10612-003>
- Gargani, J., & Donaldson, S. I. (2011). What works for whom, where, why, for what, and when? Using evaluation evidence to take action in local contexts. *New Directions for Evaluation*, 2011(130), 17–30. <https://doi.org/10.1002/ev.362>
- Geisinger, K. F. (Ed.). (2013). *APA handbook of testing and assessment in psychology*. American Psychological Association.
- Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, 6(1), 109–127.
- Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge. <https://doi.org/10.4324/9781003025634>
- Gould, S. J. (1981). *The mismeasure of man*. W. W. Norton.
- Gregoire, J., & Hambleton, R. K. (Eds.). (2009). Advances in test adaptation research [Special issue]. *International Journal of Testing*, 9(2). <https://doi.org/10.1080/15305050902880678>
- Hanson, W. E., Creswell, J. W., Plano Clark, V. L., Petska, K. S., & Creswell, J. D. (2005). Mixed methods research designs in counseling psychology. *Journal of Counseling Psychology*, 52(2), 224–235. <https://doi.org/10.1037/0022-0167.52.2.224>
- Hanson, W. E., Zhou, H., Armstrong, D. L., & Liwski, N. T. (in press). A humanistic approach to mental health assessment, evaluation, and measurement-based care. In J. J. W. Andrews, S. R. Shaw, J. F. Domene, & C. McMorris (Eds.), *School, clinical, and counselling psychology: Mental health assessment, prevention, and intervention*. Springer.
- Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes and causation. *Journal of Economic Growth*, 17(4), 267–321. <https://doi.org/10.1007/s10887-012-9081-x>
- Haverkamp, B. E. (2012). The counseling relationship. In E. Altmeier & J. C. Hansen (Eds.), *Oxford handbook of counseling psychology* (pp. 32–70). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780195342314.001.0001>
- Haverkamp, B. E. (2013). Education and training in assessment for professional psychology: Engaging the “reluctant student.” In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 2. Testing and assessment in clinical and counseling psychology* (pp. 63–82). American Psychological Association. <https://doi.org/10.1037/14048-005>
- Haynes, S. N., Smith, G. T., & Hunsley, J. D. (2011). *Scientific foundations of clinical assessment*. Routledge. <https://doi.org/10.4324/9780203829172>
- Health Service Psychology Education Collaborative. (2013). Professional psychology in health care services: A blueprint for education and training. *American Psychologist*, 68(6), 411–426. <https://doi.org/10.1037/a0033265>
- Hu, S., & Oakland, T. (1991). Global and regional perspectives on testing children: An empirical study. *International Journal of Psychology*, 26(3), 329–344. <https://doi.org/10.1080/00207599108246857>

- Iliescu, D. (2017). *Adapting tests in linguistic and cultural situations*. Cambridge University Press. <https://doi.org/10.1017/9781316273203>
- LaVelle, J., & Donaldson, S. I. (2021). Opportunities and challenges ahead for university-based evaluator education programs, faculty, and students. *The American Journal of Evaluation, 42*(3), 428–438. <https://doi.org/10.1177/1098214020937808>
- Leong, F., Bartram, D., Cheung, F., Geisinger, K. F., & Iliescu, D. (Eds.). (2016). *The ITC international handbook of testing and assessment*. Oxford University Press.
- Mark, M. M., Donaldson, S. I., & Campbell, B. (Eds.). (2011). *Social psychology and evaluation*. Guilford Press.
- Marshall, T., & Robertson, G. (2019, June 7). Eugenics in Canada. In E. Yarhi (Ed.), *The Canadian Encyclopedia*. <https://www.thecanadianencyclopedia.ca/en/article/eugenics>
- Mateo, J., Escofet, A., Martínez-Olmo, F., Ventura, J., & Vlachopoulos, D. (2012). Evaluation tools in the European Higher Education Area (EHEA): An assessment for evaluating the competences of the Final Year Project in the social sciences. *European Journal of Education, 47*(3), 435–447. <https://doi.org/10.1111/j.1465-3435.2012.01536.x>
- Meckes, L., & Carrasco, R. (2010). Two decades of SIMCE: An overview of the National Assessment System in Chile. *Assessment in Education: Principles, Policy & Practice, 17*(2), 233–248. <https://doi.org/10.1080/09695941003696214>
- Mertens, D., Bledsoe, K., Sullivan, M., & Wilson, A. (2010). Utilization of mixed methods for transformative purposes. In A. Tashakkori & C. Teddlie (Eds.), *SAGE handbook of mixed methods in social & behavioral research* (pp. 193–214). Sage Publications. <https://doi.org/10.4135/9781506335193.n8>
- Mertens, D. M., & Hesse-Biber, S. (2013). Mixed methods and credibility of evidence in evaluation. *New Directions for Evaluation, 2013*(138), 5–13. <https://doi.org/10.1002/ev.20053>
- Mertens, D. M., & Wilson, A. T. (2019). *Program evaluation theory and practice: A comprehensive approach* (2nd ed.). Guilford Press.
- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*(2), 128–165. <https://doi.org/10.1037/0003-066X.56.2.128>
- Miller, S. D., Hubble, M. A., & Chow, D. (2020). *Better results: Using deliberate practice to improve therapeutic effectiveness*. American Psychological Association. <https://doi.org/10.1037/0000191-000>
- Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.). (2002). *Computer-based testing: Building the foundation for future assessments*. Taylor & Francis.
- Mislevy, R. J. (2010). Design under constraints: The case of large-scale assessment systems. *Measurement: Interdisciplinary Research and Perspectives, 8*(4), 199–203. <https://doi.org/10.1080/15366367.2010.526452>
- Moghaddam, F. M., Erneling, C. E., Montero, M., & Lee, N. (2007). Toward a conceptual foundation for a global psychology. In M. J. Stevens & U. P. Gielen (Eds.), *Toward a global psychology: Theory, research, intervention, and pedagogy* (pp. 179–206). Lawrence Erlbaum Associates.
- Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McNerney, J., & Thue, B. (2020). Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science, 31*(6), 678–701. <https://doi.org/10.1177/0956797620916782>
- Neave, G., & Veiga, A. (2013). The bologna process: Inception, ‘take up’ and familiarity. *Higher Education, 66*(1), 59–77. <https://doi.org/10.1007/s10734-012-9590-8>
- Nishitani, H. (2020). *English education reform-based on EBPM (Evidence-Based Policy Making)*. <https://www.openaccessgovernment.org/english-education-reform/79823/>
- Oakland, T., Douglas, S., & Kane, H. (2016). Top ten standardized tests used internationally with children and youth by school psychologists in 64 countries: A 24-year

- follow-up study. *Journal of Psychoeducational Assessment*, 34(2), 166–176. <https://doi.org/10.1177/0734282915595303>
- Oakland, T., Wechsler, S. M., & Maree, K. (2013). Test use with children across cultures: A view from three countries. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 231–257). American Psychological Association. <https://doi.org/10.1037/14049-011>
- Ofir, Z., & Kumar, A. K. S. (2013). Evaluation in developing countries: What makes it different? In S. I. Donaldson, T. A. Azzam, & R. Conner (Eds.), *Emerging practices in international development evaluation* (pp. 11–24). Information Age Publishing.
- Pokropek, A., Marks, G. N., & Borgonovi, F. (2021). How much do students' scores in PISA reflect general intelligence and how much do they reflect specific abilities? *Journal of Educational Psychology*. Advance online publication. <https://doi.org/10.1037/edu0000687>
- Poston, J. M., & Hanson, W. E. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological Assessment*, 22(2), 203–212. <https://doi.org/10.1037/a0018679>
- Province of Alberta. (1928). *The Sexual Sterilization Act of Alberta*. Statutes of the Province of Alberta. <https://eugenicsarchive.ca/database/documents/5172e81ceed5c6000000001d>
- Reddy, A. (2008). The eugenic origins of IQ testing: Implications for post-Atkins litigation. *DePaul Law Review*, 57(3), Rev. 667. <https://via.library.depaul.edu/law-review/vol57/iss3/5>
- Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, 21(5), 667–706. <https://doi.org/10.1002/per.634>
- Rodolfa, E., Bent, R., Eisman, E., Nelson, P., Rehm, L., & Ritchie, P. (2005). A cube model for competency development: Implications for psychology educators and regulators. *Professional Psychology, Research and Practice*, 36(4), 347–354. <https://doi.org/10.1037/0735-7028.36.4.347>
- Shealy, C. N. (Ed.). (2016). *Making sense of beliefs and values: Theory, research, and practice*. Springer.
- Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. The MIT Press. <https://doi.org/10.7551/mitpress/9589.001.0001>
- Tashakkori, A., & Teddlie, C. (2010). *Sage handbook of mixed methods in social and behavioral research*. Sage Publications. <https://doi.org/10.4135/9781506335193>
- Terjesen, M. (2015, March). *Developing global partnerships for conducting research in schools: Closing the science-practice-culture gap* [Paper presentation]. Cultivating the Globally Sustainable Self Summit Series, Harrisonburg, VA, United States.
- van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer.
- Wandschneider, E., Pysarchik, D. T., Sternberger, L. G., Ma, W., Acheson, K., Baltensperger, B., Good, R. T., Brubaker, B., Baldwin, T., Nishitani, H., Wang, F., Reisweber, J., & Hart, V. (2015). The Forum BEVI Project: Applications and implications for international, multicultural, and transformative learning. *Frontiers: The Interdisciplinary Journal of Study Abroad*, 25, 150–228.
- Wiggins, J. S. (2005). *Paradigms of personality assessment*. Guilford Press.