Citation: Huitt, W., & Monetti, D. (2015). Norm-based assessment. In M. Spector (Ed.), *The SAGE Encyclopedia of educational technology* (pp. 545-547). Thousand Oaks, CA: Sage.

Norm-referenced Assessment

The interaction of technology and norm-referenced assessment creates many opportunities, while raising significant issues and concerns. One of the primary issues to consider is the purpose of assessment for teaching and learning. Considering categories for assessments can clarify different purposes and help educators and policy makers to better address important questions. The most common categorization of types of assessment includes the comparison of scores to a preset standard (criterion-referenced) or to those from a group of similar individuals (norm-referenced). However, Royal Van Horn argued for the use of improvement-referenced assessment (criterion-referenced measures taken multiple times) while Peter Behuniak proposed that assessment should be consumer-referenced (for the purpose of informing stakeholders). When analyzing the efficacy of norm-referenced assessment, two critical questions become: (1) what is the benefit of norm-referenced assessment that cannot be addressed with other approaches; and (2) if norm-referenced assessment is warranted, how can technology be used to make the process more efficient and effective?

Dylan Wiliam stated that scores derived from norm-referenced testing are relatively insensitive to instruction; this is the primary challenge for their use in assessment of academic knowledge and skills. As a result, the majority of classroom- and state-based standardized assessments in the United States are criterion-referenced. There are several areas where an effort to utilize norm-referenced assessment might be justified: (1) increasing the efficiency of norm-referenced assessments when comparisons are made on academic content or skills where there is a lack of coherence of expected standards or taught content; (2) providing learners and other stakeholders information about how students compare to large groups of similar individuals; (3) providing important information for program evaluation if learner input characteristics among the comparison groups are matched; and 4) collecting data on important domains of human development and behavior that do not as yet have established standards and benchmarks.

Computer-based Testing and Computer Adaptive Testing

Two technological contributions to norm-referenced assessment are computer-based testing and computer adaptive testing. The major advantage of computer-based testing is that results can be provided to students and other interested stakeholders much quicker. An additional advantage is that the testing is supposed to be more secure because onsite administrators cannot as easily modify the results as they can with the paper/pencil versions. However, many schools do not have the computer resources for all students in a school to take the exam at the same time. Therefore, paper/pencil tests are still made available where an adequate number of computers or computer access is not available. A second advantage of the computer-based testing is the opportunity for more interactivity in the testing process. For example, the test developers can provide simulations, have students view video clips, or have students engage in activities and then answer questions rather than simply recall information they have already learned.

Computer adaptive testing (CAT) provides a different procedure for presenting the tested content to the test taker. In this process, not all items are given to any specific test taker; rather the computer adapts to the correct and incorrect answers provided by the test taker and presents different questions to different individuals. Those who advocate this procedure suggest that the test provides more valid results if most of the items answered by the test taker match his or her level of knowledge and skill rather than being too easy or too difficult. Wim van der Linden and Peter Pashley showed that a major advantage to test developers is that the processes of item selection and the estimation of item difficulty can be made more efficient because of the use of the Rasch measurement model. In the traditional test development process, items are generated and then reviewed by content experts. A test is developed and then piloted by giving the test to potential examinees. Items are then selected based on their reliability and relative difficulty levels. In an adaptive testing procedure, potential items can be used much more quickly in the development process, thereby producing savings in the developmental costs. A second advantage is that the testing procedure can be more efficient because each examinee does not answer all possible questions. Rather the items are organized into what are called testlets which Wim van der Linden described as "bundles of items related to sets of content specifications that be selected only as intact units" (p. 30). Examinees are systematically provided with testlets until it is determined that they can answer most of the questions reliably or they cannot. The test thereby converges on the true knowledge and skill of the examinee.

Use of Norm-referenced Assessment

In the United States, norm-referenced assessments are widely used for the purpose of screening high school graduates in their admittance to an institution of higher education or to specific programs within those institutions. The ACT and SAT tests are the two primary instruments used for admittance to undergraduate programs. A justification for using norm-referenced assessments for this purpose might be twofold: (1) there is a lack of coherence between content taught in high school and college and, therefore, specific content knowledge as might be assessed with criterion-referenced assessments is not a prerequisite for doing well in college-level courses, and (2) there are a limited number of openings that can be allocated in any college admission process and the college wants to select those most likely to be successful. One might raise the issue of why high stakes norm-referenced assessments are used for admission to higher education rather than measures of past academic performance such as high school cumulative grade point averages (HSCGPA). Elchanan Cohn and his colleagues demonstrated that SAT scores provide prediction of college success above and beyond HSCGPA, while Justine Radunzel and Julie Noble provide the same support for the ACT. The rationale for the use of norm-referenced assessment therefore appears to be empirically justified.

The GRE, GMAT, and the MCAT are the primary tests required for admittance to graduate or professional studies. A similar question can be raised about the use of these exams: do they provide any predictive power above and beyond what might be expected by using the past performance of undergraduate cumulative grade point average? In a widely respected meta-analytic study, Nathan Kuncel and his colleagues demonstrated that the GRE was a valid predictor of success in graduate school above and beyond what might be expected from undergraduate grades, although the subject-level tests were better predictors than were the more general verbal, quantitative, and analytic scores. Likewise, In-Sue Oh and his colleagues demonstrated that the GMAT provided increased predictive power for admittance to business

programs. What is missing in these analyses is whether individuals who score higher on these assessments become more effective, productive, or eminent in their fields after graduation.

At the international level, the Program for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) are somewhat useful for comparisons of academic content and skills among countries. These tests are seen as an audit of a country's educational system, assessing only a relatively small number of learners in a limited number of schools, districts, and states. The results can be used to make statements about the effectiveness of the combination of non-formal, informal, and formal learning experiences of groups of learners.

Because norm-referenced tests address content and skills beyond what is taught in schools, increasing the efficiency and/or effectiveness of administering these tests can provide some information about the sociocultural context as a whole (which includes schooling). Again, however, it provides less information about how teaching and instruction in schools might be improved. The fact that scores on norm-referenced assessments of learners' academic content knowledge and skills add predictive validity beyond grades in formal learning environments demonstrates that the non-formal and informal learning experiences are important for predicting success in the formal learning environments of undergraduate and graduate schooling.

Potential Uses of Norm-referenced Assessments

Previously, the point was made that the collection of data on important domains of human development offers a great deal of advantages for norm-referenced assessment. Criterion-referenced assessment would not be appropriate for these domains because they do not as yet have established standards and benchmarks. There are a number of research and theoretical orientations that are worthy of consideration. For example, Howard Gardner, Daniel Goleman, and Robert Sternberg and his associates have all concluded that cognitive intelligence, and by implication norm-referenced assessments of academic knowledge and skills, at best predict one-third of the variance in adult success. Up to two-thirds of that variance is related to development of other factors such as emotional, social, and self-regulation Each of these researchers has developed or identified norm-referenced instruments or procedures that can be used to assess factors beyond the traditional academic knowledge and skills. Other researchers who have done similar work include Ed Diener, Martin Seligman, and the Collaborative for Academic, Social, and Emotional Learning. Ed Diener and his colleagues developed measures related to an individual's wellbeing: positive and negative experiences, positive thinking, and psychological well-being. Likewise, Martin Seligman worked with a variety of colleagues to develop measures of the five components of his PERMA theory of wellbeing: Positive Emotion, Engagement, Positive Relationships, Meaning and Purpose, and Accomplishment. Similarly, the Collaborative for Academic, Social, and Emotional Learning (CASEL) developed a number of instruments to measure the five components that provide a foundation for personal development: Self-awareness, Social awareness, Self-management, Relationship skills, and Responsible decision making. The lack of inclusion of these assessments into the overall assessment process of children, youth, and young adults omits factors that account for a majority of the variance related to adult success. This is certainly an area where technology-based norm-referenced assessment can add value to the teaching and learning process. William Huitt provided an overview of the research that supports a focus on domains beyond that of student academic achievement.

Summary and Conclusion

In conclusion, norm-referenced testing has an important role to play in developing a deeper understanding of teaching and learning. This analysis discussed the different categories of standardized assessment, summarized the role that norm-referenced assessment plays in K-12, undergraduate, and graduate education, looked at how norm-referenced assessment is utilized in international comparisons, and how the testing experience is fundamentally altered through the use of technology. However, what is most important about the efforts to expand the focus of schooling beyond the traditional academic knowledge and skills assessed by most norm-reference tests is that they address many of the non-academic competencies that the Partnership for 21st Century Skills and others advocate as important for adult success in the 21st century. While it is certainly important to be more efficient and effective in the development and delivery of traditional norm-reference assessments, an argument can be made the expansion of assessments to include data on attributes that are perhaps twice as predictive of life success as is academic knowledge.

William G. Huitt, PhD
David M. Monetti, PhD

Further Readings

Behuniak, P. (2002). Consumer-referenced testing. *Phi Delta Kappan, 84(*3), 199-207.

Cohn, E., Cohn, S., Balch, D., & Bradley, J., Jr., (2001). *The effect of SAT scores, high-school GPA and other student characteristics on success in college*. Proceedings of the Annual Meeting of the American Statistical Association, August 5-9. Retrieved from http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00187.pdf

Huitt, W. (2011, July). *A holistic view of education and schooling: Guiding students to develop capacities, acquire virtues, and provide service*. Revision of paper presented at the 12thAnnual International Conference sponsored by the Athens Institute for Education and Research (ATINER), May 24-27, Athens, Greece. Retrieved from http://www.edpsycinteractive.org/papers/holistic-view-of-schooling-rev.pdf

Kuncel, N., Hezlett, S., & Ones, D. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*(1), 162-181.

Oh, I-S., Schmidt, F., Shaffer, J., & Huy, L. (2008). The Graduate Management Admission Test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. *Academy of Management Learning & Education, 7*(4), 563-570.

Radunzel, J., & Noble, J. (2012). Predicting long-term college success through degree completion using ACT® composite score, ACT benchmarks, and high school grade point average. ACT Research Report Series (5). Iowa City, IA: ACT, Inc.. Retrieved from http://www.eric.ed.gov/PDFS/ED542027.pdf

Van Horn, R. (2003). Computer adaptive tests and computer-based tests. *Phi Delta Kappan, 84(*8), 567, 630-631.

van der Linden, W. (2000). Constrained adaptive testing with shadow tests. In W. van der Linden & C. Glas (Eds.). *Computerized adaptive testing: Theory and practice* (pp. 27-52). Hinghma, MA: Kluwer Academic Publishers.

van der Linden, W., & Pashley, P. (2000). Item selection and ability estimation in adaptive testing. In W. van der Linden & C. Glas (Eds.). *Computerized adaptive testing: Theory and practice* (pp. 1-25). Hinghma, MA: Kluwer Academic Publishers.

Wiliam, D. (2008). International comparisons and sensitivity to instruction. *Assessment in Education: Principles, Policy & Practice, 15*(3), 253-257.