Inclusive Next Generation Assessment: An Examination of Approaches for Evaluating Technical Quality

Edynn Sato, Ph.D.

Reidy Interactive Lecture Series
Cambridge, Massachusetts
October 21-22, 2010

WestEd

Assessment and Accountability Comprehensive Center
AACC • A WestEd and CRESST partnership

Overview

- Research—Some Critical Points
  - Special student populations (SWDs, ELs)
  - Access
  - Validity Evidence
- Toward a Theory of Action for Inclusive Next Generation Assessment
- Technical Quality—Some Considerations
- Some Additional Considerations

What research supports: Some critical points

Special population students:

- Students with disabilities (SWDs) and English learner (EL) student populations are heterogeneous
  - e.g., EL students differ according to English language proficiency, native language proficiency, cultural proximity, time and consistency in U.S. schools, and U.S. learning and assessment experiences.
  - e.g., SWDs differ according to sensory, physical, cognitive, and/or linguistic capacities
- SWDs and EL students can learn complex, rigorous academic content

[See references at end of presentation.]

What research supports: Some critical points

Access:

- Access is arguably the most relevant threat to validity for SWDs and EL students
- There are viable, systematic ways to address the range of students' access needs such that students can fully demonstrate what they know and can do
- Accommodations, Universal Design (UD), and Universal Design for Learning (UDL) are necessary but not sufficient for addressing the range of students' access needs
- Access should be considered and addressed throughout an assessment's design, development, and implementation process

What research supports: Some critical points

Validity Evidence:

- Evidence should help to examine interactions among student characteristics, access strategies, content, and assessment formats and their effect on the assessed construct, validity of interpretations of student outcomes, and consequences should be considered
- Criteria for judging the technical quality of assessments of the general student population overlap with criteria relevant to special student populations; however the criteria do not overlap completely

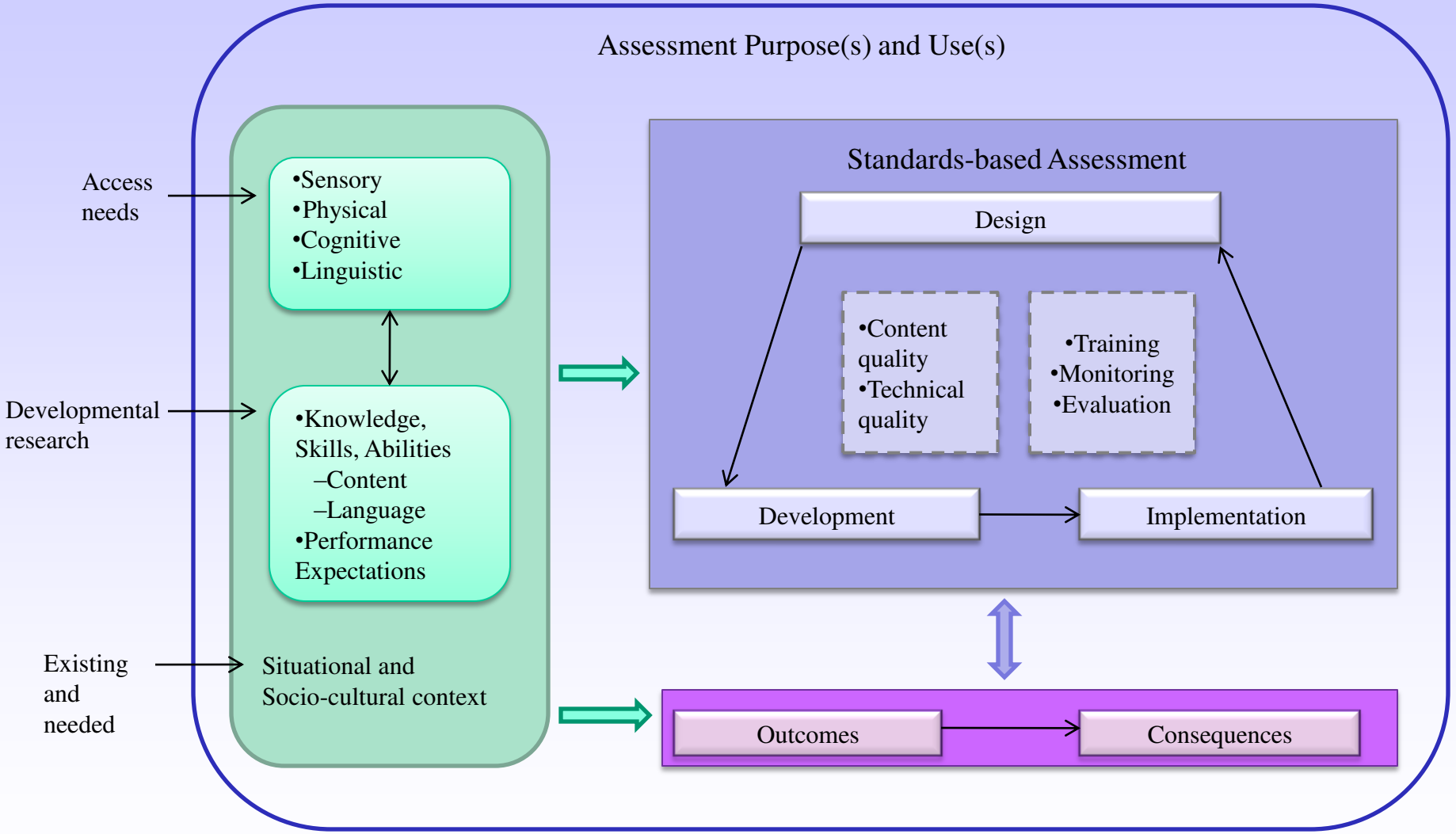
Toward a Theory of Action for Inclusive Next Generation Assessment

- Extends knowledge and research of validity approaches (e.g., Messick's Unified View of Validity, Kane's Argument-based Approach, Mislevy's Evidence-centered Assessment Design)
- Integrates emerging research on access and the assessment of special student populations
- While focused on special student populations, generalizes to all students—necessarily moves beyond UD, UDL, and targeted accommodations

Toward a Theory of Action for Inclusive Next Generation Assessment

- Supports an integrated, systemic approach to assessment design and development
- Applies to formative and summative assessments
- Necessary for computer-delivered assessments and innovative accessible tasks
- Serves as a "road map" for guiding the design, development, and implementation of inclusive next generation assessment
- Allows for multiple points of entry

Toward a Theory of Action for Inclusive Next Generation Assessment



Technical Quality: Some Considerations Related to Inclusive Assessment

Given two assessment tasks developed to measure a particular construct, to what degree do they both measure the that construct? (construct comparability)

- Alignment
  - Array of representations of the construct
    - Different semiotic representations
    - "Default, alternate, and supplemental content"
    - Content versus language
  - Effect of access strategy/strategies on the construct
    - Presentation, engagement, response, format/medium
    - Sensory, physical, cognitive, and/or linguistic
  - Effect of context on the construct
    - Situational
    - Socio-cultural

Technical Quality: Some Considerations Related to Inclusive Assessment

Do students from different subgroups attach the same meaning to the construct as a whole? (construct equivalence)

- Dimensionality
- Construct invariance
  - Non-monotonic increase

Technical Quality: Some Considerations Related to Inclusive Assessment

Do equivalent test scores have the same meaning for all students, regardless of their group membership? (measurement invariance)

- Selection bias
  - Different sample sizes of focal and reference groups
  - Different proficiency distributions between the focal and reference groups
- Dimensionality
- Variation at item level versus total score
- Flexibility versus standardization of task/administration

Some Additional Considerations

Inferences and consequences…

- Flexibility and standardization
- Targeted and non-targeted constructs
- Underestimation
- Under-representation
- Differential boost
  - How much and for whom?
- Incremental validity
  - How should this relate to possible marginal changes to the assessed construct and/or limited test equivalence?

References

- Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74:* pp. 1-28.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for Educational and Psychological Testing. Washington, DC: AERA.
- Butler, F.A. & Stevens, R. (2001). Standardized assessment of the content knowledge of English Language Learners K-12: Current trends and old dilemmas. Language Testing 2001, 18(4), 409-427.
- Chapelle, C.A., Enright, M.K., Jamieson, J. (2010). Does and argument-based approach to validity make a difference? Educational Measurement: Issues and Practice, 29 (1), 3-13.
- Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment, 3*(7). Available from http://www.jtla.org
- Gong, B. & Marion, S. (2006). Dealing with flexibility in assessments for students with significant cognitive disabilities (synthesis Report 60). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 2010, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis60.html.
- Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis MN: University of Minnesota, National Center on Educational Outcomes.
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (Volume 4, pp. 17-64). Westport, CT: American Council on Education and Praeger.
- Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*, 3-16.
- Kopriva, R. (2000). Ensuring accuracy in testing for English language learners. Washington, DC: Council of Chief State School Officers.
- Kopriva, R. & Carr, T.G. (2008). It's about time: Matching English learners and the ways they take tests by using an online tool to properly address individual needs. Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Diego, CA.

References

- Marion, S.F. & Pellegrino, J.W. (2006). A validity framework for evaluating the technical quality of alternate assessments. Educational Measurement: Issues and Practice, 25(4), 47-57.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103.). Washington, DC: American Council on Education.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*. Retrieved August 2010, from http://www.education.umd.edu/EDMS/mislevy/papers/BriefIntroECD.pdf
- Mislevy, R. J. & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 4,* 6-20.
- Pearson, P. & Garavaglia,D. (2003). *Improving the Information Value of Performance Items in Large Scale Assessments: NAEP Validity Studies. Working Paper Series*. ED Pubs.
- Rabinowitz, S., & Sato, E. (2005). *A technical review of high-stakes assessments for English language learners.* San Francisco: WestEd.
- Rivera, C. & Collum, E. (2004). *An analysis of state assessment policies addressing the accommodation of English Language Learners.* Issue paper commissioned for the National Assessement Governing Board Conference on Increasing the Participation of SD and LEP Students in NAEP. Arlington, VA: George Washington University.
- Rose, D. & Meyer, A. (2000). Universal Design for Learning. Journal of Special Education Technology, 15 (1), 67-70.
- Russell, M. (in press). Accessible test design. Boston, MA: Boston College.
- Russell, M., Hoffmann, T., & Higgins, J. (2009). NimbleTools: A universally designed test delivery system. *Teaching Exceptional Children*. 42(2), 6-12.
- Sireci, S.G., Li, S., & Scarpati, S. (2002). *The effects of test accommodations on test performance: A review of the literature*. CEA Research Report 485. Amhearst, MA: School of Education, University of Massachusetts.
- Thompson, S. J., Johnstone, C. J. & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Winter, P.C., Kopriva, R. J., Chen, C-S., & Emick, J.E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences* 16 (2006) 267–276.

For more information contact:

Edynn Sato

esato@wested.org