Developing a Framework for the International Benchmarking of Performance Standards

Andrew Wiley and Susan Davis-Becker

Paper presented at the annual meeting at the
National Council on Measurement in Education
Saturday April 18, 2015

Validity in standard setting

- Kane (1994, 2001)
  - Procedural
  - Internal
  - External
  - No perfect criterion
  - Encouraged the collection of evidence that the performance standards are reasonable and defensible

Validity evidence for standard setting

- Characteristics of panelists
- Training of panelists
- Multiple iterations of ratings, and change in panelists' ratings
- Internal consistency of panelists' ratings
- Consistency across panels
- Panelists' understanding of the process and comfort with the passing scores being recommended

Buckendahl, et al (2009)

- Created an overall framework for the evaluation of NAEP standard setting
- Illustrated the data and information necessary for the evaluation of each aspect of Kane's validity framework

Procedural validity

Criterion	Brief Explanation
Participants	Qualifications, competence, & representativeness of panelists; sufficient number of panelists
Standard setting method(s)	Degree to which methods used are logical, defensible, & congruent with testing purpose
Panelist training	Degree to which panelists were properly oriented, prepared, and trained
Clarity of goals/tasks	Degree to which standard setting purposes, goals, and tasks were clearly articulated
Data collection	Data were gathered as intended
Implementation	Method implemented as intended
Panelist confidence	Panelists understood tasks / had confidence in ratings
Sufficient documentation	Documentation of the entire process so (a) it is understood and (b) can be replicated

Internal Validity

Criterion	Brief Explanation
Inter-panelist consistency	Reasonable standard deviations and ranges of cut scores across panelists
Decreasing variability across rounds	The variability across panelists' cut scores decreases across rounds—evidence of emerging consensus
Standard error of cut score	Estimate of degree to which cut scores would change if study were replicated
Consistency across independent panels	Estimate of degree to which cut scores would change if different panelists were used
Consistency across panelist subgroups	Estimate of degree to which cut scores would change if specific types of panelists were used
Consistency across item formats	Estimate of the consistency of cut scores across item formats
Borderline students performance on specific items	How consistent is the performance of the hypothetical borderline students' performance with the performance of students near the cut scores

External Validity

Criterion	Brief Explanation
Consistency across other student classification data	Comparison with other test score data
Mean differences across proficiency groups on external criteria	Passing rates of other external criterion
Reasonableness	Degree to which cut scores produce results that are within a sensible range of expectations

Value of International assessments

- Hanushek and Woessman (2009, 2011, 2012)
  - Used TIMSS and PISA
  - Demonstrated positive correlation between education achievement and economic growth
- Baker (2007)
  - Looked at the 12 nations who took the first international mathematics test in 1964
  - Looked at the per capita gross domestic produce
  - the higher a nation’s test score 40 years ago, the worse its economic performance

Framework for using int’l assessment data

- 5 essential features
  - Purpose or intent of the assessments
  - Test content
  - Examinee population
  - Administration model
  - Scoring model

Intent or Purpose

	CCSS Assessment (standard setting focus)	PISA	TIMSS
Purpose	CCSS Measurement	To assess students' preparedness to learn in today's knowledge society	Designed to allow for comparisons across school systems
	School and teacher accountability	Not used for accountability purposes	Designed to provide information to help systems adopt successful practices
	Tracking student performance		Not used for accountability purposes

Test Content

	CCSS Assessment (standard setting focus)	**PISA**	**TIMSS**
Content	Directly tied to the CCSS	Assessment of content knowledge in a real world context	More traditional classroom materials
	Variety of item types, including technology enhanced items	Tests a broad range of mathematical concepts (not a specific curriculum); closer to literacy	Covers both math and science content

Examinee population

	CCSS Assessment (standard setting focus)	PISA	TIMSS
Examinees	Approximately 98% of students are required to complete the test during their high school career	Age based sampling - Most students are 15 to 16 years old	Grade based sampling - Focuses of students in 4th and 8th grades
		Approximately 30 countries in common	
		An additional 30 countries unique to one or the other	

Test Administration

	CCSS Assessment (standard setting focus)	PISA	TIMSS
Test Administration	Administered every year at every grade level (3-8, HS)	Administered every three years	Occurs every four years
	CAT administration	One content focus per administration	
		Two hours, CBT	

Test Scoring

	CCSS Assessment **(standard setting focus)**	**PISA**	**TIMSS**
Scoring	Mean scores reported and performance in classified into four categories	Report a mean score with a mean of 500 and a SD of 100	Report a mean score with a mean of 500 and a SD of 100

Recommendations

- Identifying appropriate validity evidence to support performance standards can be challenging
- International benchmarks are an appealing option given the larger goals of many educational programs
- Appropriate use must first be evaluated against 5 key characteristics of focal program

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R. K., Zenisky, A. L., & Wells, C. S. (2009). *Evaluation of the National Assessment of Educational Progress: Final Report*. Washington, D.C.: U.S. Department of Education.

Egan, L., Beattie, K., Byrd, P., DeCandia, M. (2014, June). *Harnessing untappend potential: Using international and national assessment data to inform the transition to next general assessment systems*. Symposium conducted at the National Conference on Student Assessment, New Orleans, LA.

Hanushek, E. & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, 17(4), 267-321.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Thank you!

- Questions/comments?

- Awiley999@gmail.com