Improving assessment systems by clarifying testing purposes

Andrew Ho, *Harvard Graduate School of Education* August, 2022

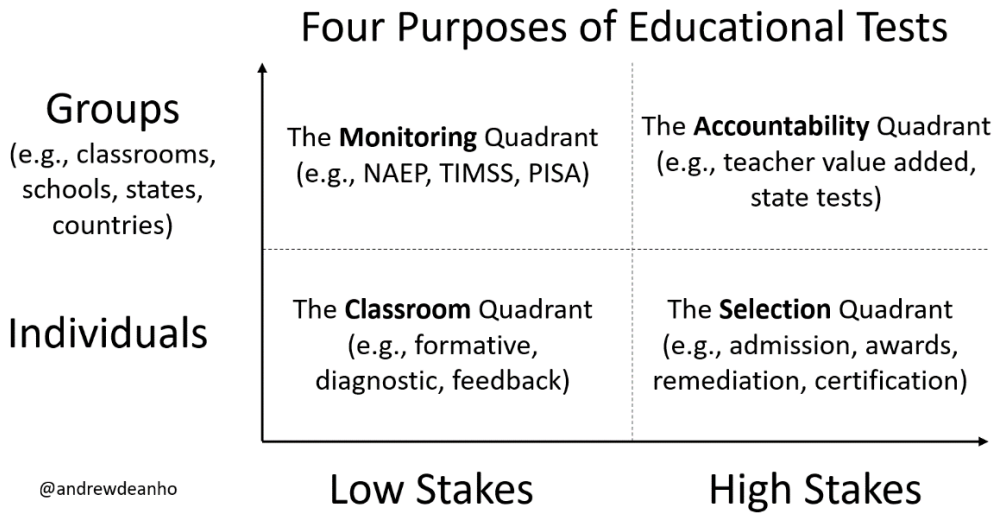
Description:

Modern assessments serve multiple conflicting purposes, including supporting instruction, informing parents, monitoring progress, and evaluating schools. I introduce a "four quadrants" framework for the purposes that tests serve in educational systems. The framework highlights opportunities to separate and clarify the multiple conflicting purposes that educational tests support. I explain how conflicting purposes can corrupt each other and threaten good teaching and learning. Using examples from formative assessment, interim benchmark assessment, and testing through the COVID-19 pandemic, I argue that separating purposes and negotiating tradeoffs can improve the how policymakers, teachers, parents, and the public use test scores.

Framing Questions:

1) What are the four purposes of educational testing that can conflict and corrupt each other?
2) What are three examples of tradeoffs among the four purposes that illustrate the challenge of serving multiple purposes simultaneously?
3) What is one strategy for clarifying and thereby advancing the purposes that tests can serve?

Many modern debates about educational testing can be clarified by specifying the intended purposes that test scores serve. I provide a mnemonic: The 3 Ws, for: Who is using Which scores for What purpose (Ho, 2022a)? I then provide a simple “4 Quadrants” taxonomy that classifies most answers to this central question. The quadrants are distinguished by whether the use of scores has higher or lower stakes and whether the scores describe individuals or groups (Ho, 2022b):



1) The Classroom Quadrant (for individuals at low stakes): Teachers and students using scored worksheets, quizzes, or tests to inform and improve teaching and learning.
2) The Selection Quadrant (for individuals at higher stakes): Admissions, awards, classification, and certification committees and organizations using test scores in whole or in part to select or track students into alternative programs or curricula or award or certify students as eligible for certain career or educational pathways.

3) The Monitoring Quadrant (for groups at low stakes): Policymakers and the public using proficiency rates and average scores to monitor academic proficiency and progress for classrooms, schools, districts, states, or countries to improve awareness of educational trends and inform educational leadership and policy, e.g., the National Assessment of Educational Progress and the Trends in International Mathematics and Science Study.
4) The Accountability Quadrant (for groups at higher stakes): Policymakers and education leaders using proficiency rates and average scores in whole or in part to determine sanctions and rewards for teachers and school or district leaders.

These contrasting actors, scores, and purposes suggest that no single test can serve all purposes well. This motivates simple guidance to developers and users of educational tests: "Don't Cross Quadrants." Yet it is common for developers and proponents of educational testing programs to blur the lines between quadrants to build constituency for their tests. Examples of crossing quadrants include the following questions: Why can't administrators use teachers' classroom tests for higher stakes admissions decisions? Why can't state agencies use high-stakes accountability tests to monitor educational progress? Why can't we use the National Assessment of Educational Progress to inform classroom instruction?

The belief of test users that a single test can serve multiple purposes rests on at least three fallacies that I call: 1) the Equating Fallacy, 2) the Aggregation Fallacy, and 3) the Test-Worth-Teaching-To Fallacy. The reasons why these fallacies are incorrect are also describable in a mnemonic, The 3 I's, that test scores are more Incomparable, Imprecise, or Inflated than users believe:

1) The Equating Fallacy is described well by Braun and Mislevy (2005) as the belief that, "any two tests that measure the same thing can be made interchangeable with a little 'equating' magic" (p. 493). Crossing from the Classroom Quadrant to the Selection Quadrant, as one example, assumes at the least that classroom test scores and the teachers or systems that rate them are comparable.
2) The Aggregation Fallacy is the belief that properties and inferences of individual scores are the same as those of aggregate scores. In fact, the two can differ dramatically depending on how they are derived (Haertel & Ho, 2016). In general, aggregate scores are more precise than individual scores, so crossing upper quadrants to lower quadrants sacrifices substantial precision.
3) The Test-Worth-Teaching-To Fallacy, as the name implies, is the belief that tests can be accurate high-stakes targets while maintaining sound measurement properties. Koretzian Inflation Theory (e.g., Koretz, 2008) shows clearly how even well-designed tests for groups will show inflated scores over time under high-stakes pressure. This can also manifest at the classroom level as students focus on unrepresentative subsets of disproportionally tested content (Ho, 2022b). In this way, tests designed for the righthand quadrants do not serve purposes in the lefthand quadrants well.

Paving a path to "systemic validation," where different actors, tests, and derived scores can achieve their separate aims in a coherent assessment system, requires not only clarifying purposes but improving the use of test scores for each purpose. For selection, systemic validation requires a clear link between the mission of an organization and the basis for selectivity. For accountability, systemic validation requires multiple measures, achievable targets, and the mantra, "no stakes without support." For classrooms, systemic validation requires focusing on student growth and dispelling student, teacher, and parent fallacies that test scores are more meaningful, precise, and permanent than they are (Ho, 2022c). For monitoring, systemic validation requires strong equating designs and investment in longitudinal linking to distinguish population changes from academic progress (e.g., An, Ho, & Davis, 2022).

References

An, L. S., Ho, A. D., & Davis, L. L. (2022). Disrupted data: Using longitudinal assessment systems to monitor test score quality. *Educational Measurement: Issues and Practice*, *41*(1), 28-32. [DOI] [Public]

Braun, H. I., & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan, 86(7),* 489-497. [DOI] [Public]

Haertel, E. H. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives, 11,* 1-18. [DOI]

Haertel, E. H., & Ho, A. D. (2016). Fairness using derived scores. In N. Dorans and L. Cook (Eds.), *Fairness in educational assessment and measurement* (217-237). New York, NY: Routledge. [DOI] [Public]

Ho, A. D. (2013). The epidemiology of modern test score use: Anticipating aggregation, adjustment, and equating. *Measurement: Interdisciplinary Research and Perspectives, 11*, 64-67. [DOI] [Public]

Ho, A. D. (2014). Variety and drift in the functions and purposes of assessment in K-12 education. *Teachers College Record, 116*(110307), 1-18. [DOI] [Public]

Ho, A. D. (2022a). Specifying the three Ws in educational measurement: Who uses which scores for what purpose? *Journal of Educational Measurement, 59*, 418-422. [DOI] [Public]

Ho, A. D. (2022b, June 6). Towards assessment literacy: Questions to ask about educational tests. Retrieved from https://scholar.harvard.edu/files/andrewho/files/testquestions.pdf

Ho, A. D. (2022c, June 25). The fragile compatibility of test scores and progressive pedagogy. *Progressive Philosophy and Pedagogy.* Retrieved from https://www.hanahauoli.org/pdc-blogposts/2021/29

Keng, L., & Marion, S. (2020). Comparability of aggregated group scores on the "same test." In A. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), pp. 49-73, Washington, DC: National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2020/06/Comparability-of-Large-Scale-Educational-Assessments.pdf

Klugman, E., & Ho, A. D. (2020) How can released state test items support interim assessment purposes in an educational crisis? *Educational Measurement: Issues and Practice, 39*(3), 65-69. [DOI] [Public]

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us.* Cambridge, MA: Harvard University Press. [DOI] [Public]

Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. *The Journal of Economic Education, 44,* 339-352. [DOI] [Public]

Reardon, S. F., Kalogrides, D., & Ho A. D. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics.* [DOI] [Public]