RESEARCH BRIEF • DECEMBER 2025

Stop Using Test Scores to Measure Test Results

Based on BFI Working Paper No. 2025-143, "Do Test Scores Misrepresent Test Results? An Item-by-Item Analysis," by Jesse Bruhn, Brown University; Michael Gilraine, Simon Fraser University; Jens Ludwig, University of Chicago; and Sendhil Mullainathan, Massachusetts Institute of Technology

Collapsing students' responses to individual test questions to single scores discards useful information about teacher efficacy and student knowledge that predicts outcomes like graduation, discipline, and future earnings.

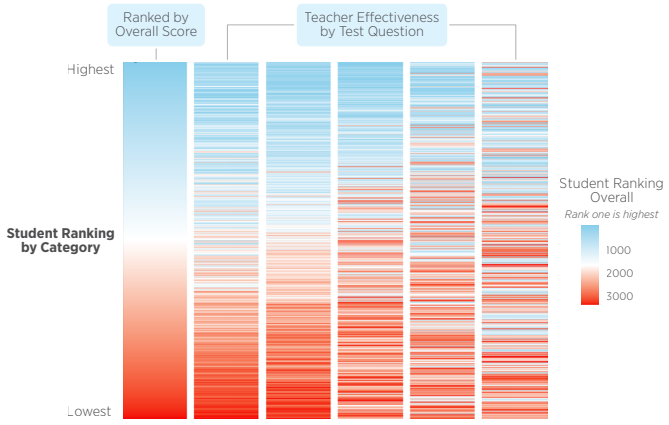
Imagine a teacher evaluating an in-class math exercise. The teacher is likely to analyze each student's answers: Who could use more help on long division? Who has mastered their multiplication tables? They are less likely to simply sort their students by their overall math ability based on their percentage of correct responses.

Despite this intuition, when it comes to standardized testing, we tend to follow the first approach. We assume aggregate test scores sufficiently capture students' ability, and use scores to inform high stakes decisions like which students to give extra help to or which teachers to hire or fire. Schools invest as much as 18% of student time spent on testing and preparation.

In this paper, the authors examine whether this approach throws away useful information. They ask: Do individual responses to test questions better predict student outcomes than aggregate scores do?

Using data from five million Texas students spanning eight years, the authors connect responses to 1.31 billion individual test questions with student outcomes ranging from course grades and disciplinary problems to graduation, college attendance, and earnings. They compare predictions made using individual questions to those made using aggregate scores—both simple averages and Item Response Theory scores (a slightly more sophisticated aggregate that weights harder questions more heavily than easier ones). For teachers, the authors measure effectiveness using

Figure 1 · Teachers Ranked 'Excellent' Overall Show Surprising Weaknesses; Teachers Ranked 'Poor' Overall Have Hidden Strengths



Note: Each vertical line represents one of 2,280 fourth-grade math teachers in Texas in 2016. Teachers are color-coded by their ranking on the overall average test score (left column), then re-ranked for each test content area while keeping their original color. If a single measure captured all teacher effectiveness, each column would show the same smooth gradient from blue (top) to red (bottom). Instead, the dramatic color mixing reveals that 'good' teachers have specific weaknesses and 'struggling' teachers have specific strengths—information that is completely lost when we look only at average scores. This hidden structure represents 66% of the predictable variation in teacher performance.

both approaches to determine whether aggregation hides meaningful performance patterns.

They find the following:

- Aggregation destroys substantial information about both students and teachers. When teachers are ranked by their effectiveness using aggregate scores, then re-ranked based on individual test questions, the rankings shift dramatically. The correlation between the two rankings is only 0.66 for teachers and 0.75 for students, far from the 1.0 that would indicate perfect agreement.
- The lost information matters for high-stakes teacher decisions. When identifying which teachers fall in the bottom 5% for their effectiveness at improving student outcomes, aggregate scores and item-level data point to different teachers. The two approaches disagree 51.6% of the time when measuring teachers' impact on student class failure rates, 42.4% for student disciplinary infractions, 44.9% for student graduation, and 39.6% for student college attendance.
- Question-level data improves decisions about which teachers to replace. Because test questions change year-to-year, the authors develop a method to categorize similar items across years by identifying which types of questions consistently differentiate teachers based on their patterns of comparative advantage. Using these item categories, rather than aggregate test scores, to identify the bottom 5% of teachers produces 21% more graduates for the same number of teacher replacements.
- Switching to item-level analysis is highly cost-effective. Schools already collect individual question responses, but discard them before analysis. The additional cost of storing and analyzing item-level rather than aggregate data is negligible, yielding an infinite **marginal value of public funds** at implementation costs below $4 million annually.

It's time to rethink how we use testing data in both research and practice. This study shows that aggregating test responses into single scores discards information that could improve critical decisions about students and teachers. Education is not alone in this problem. Aggregation occurs throughout economics—in health, housing, transportation, crime, public finance, consumer finance, and even inflation measurement. Evidence that aggregation destroys valuable information in education suggests we should question whether it's harmless in these other domains.

Marginal Value of Public Funds (MVPF): The MVPF is designed to measure long-run policy effectiveness. It is calculated as the ratio of two numbers: the benefits that the policy provides, divided by the government cost. The numerator (benefits) captures the extent to which the policy improves the lives of beneficiaries (described by economists as individuals' "willingness to pay"), and the denominator reflects net government cost.

READ THE WORKING PAPER

NO. 2025-143 · NOVEMBER 2025

Do Test Scores Misrepresent Test Results? An Item-by-Item Analysis

bfi.uchicago.edu/working-papers/do-test-scores-misrepresent-test-results-an-item-by-item-analysis

ABOUT OUR SCHOLAR

Jens Ludwig
Edwin A. and Betty L. Bergman Distinguished Service Professor, Harris School of Public Policy

THE UNIVERSITY OF CHICAGO
HARRIS SCHOOL
OF PUBLIC POLICY

Written by Abby Hiller • Designed by Maia Rabenold