Validity and Reliability Considerations for Educational Measurement

Jonathan Schweig, Susha Roy, Lauren Covelli
Last updated: March, 2026

Introduction

Federal and state laws often call for "valid" and "reliable" measures in school policy, research, and evaluation. These terms signal a shared goal: ensuring educational decisions, assessments, and evaluations are based on trustworthy information. In broad terms, validity refers to the degree to which evidence and theory support the interpretations of scores for their intended uses, indicating whether a measure captures what it is supposed to capture. In this way, validity is associated with the concept of accuracy. Reliability refers to the extent to which a measure produces scores that are consistent and free of error. In this way, reliability is associated with the concept of precision. Together, validity and reliability form the foundation for sound educational measurement.

The federal and state laws calling for valid and reliable measures refer to a wide range of school system features, including student assessment and teacher evaluation. At the federal level, the Every Student Succeeds Act (ESSA) of 2015 requires states to use standardized tests in ways appropriate for their verified accuracy and consistency—standards echoed in the Individuals with Disabilities Education Act (IDEA). At state level, many states specify that teacher evaluation systems must be valid and reliable before being used to inform personnel decisions. For example, Michigan law (sections 380.1249 and 380.1249b) requires that any changes to teacher evaluation systems preserve the system's accuracy and fairness. Similar terminology often appears in the criteria for competitive research grants, including those offered by the Institute of Education Sciences and The Spencer Foundation. These recurring references to validity and reliability reflect a widespread belief: that good policy depends on good data.

However, these terms are often misunderstood outside of research settings. Policy documents sometimes treat reliability and validity as fixed features that guarantee quality, rather than as ongoing interpretive processes that depend on how and where a measure is used. In reality, evaluating validity and reliability requires continual examination of how well a measure captures what it intends to capture and how consistently it performs across different populations and contexts. Weighing evidence of the strengths and weaknesses of a measure helps us ask critical questions about whether the evidence guiding decisions is fair and sound. While some skeptics have noted that there is a large divide between validity theory and the practical needs of

applied researchers,[1,2] and others have argued that many of the measures used in school settings serve only as proxies for the outcomes of interest,[3] validity is one of the most fundamental concepts in educational measurement.[4] As Popham wrote, "Valid inferences about students contribute to sensible decisions by educators about how to help those students; invalid inferences about students do the opposite."[5]

In this chapter, we highlight the important role that high-quality measurement plays in development and evaluation of education policy. We first provide a brief overview of validity and reliability, grounding the terms in a tradition that treats these concepts as interrelated processes whereby evidence is accumulated to evaluate whether interpretations of indicators produced from a measurement system are warranted for a specific purpose. We then present five illustrative examples drawn from real-world evaluation and policy contexts.

Key findings

Key finding #1. Evidence of validity and reliability is essential for understanding cause-and-effect relationships between policies and outcomes. Policymakers must have confidence that evidence of an intervention's success (or failure) reflects genuine change rather than artifacts of measurement error or shifting interpretations of what is being measured. When innovations are found to have statistically significant effects, evidence of validity helps ensure that these results are accurate and do not merely reflect changes in how students interact with instruments. When analyses yield nonsignificant effects, reliable measurement helps rule out random error as the explanation.

Key finding #2. Evidence of validity and reliability is critical for student proficiency trends on state standardized tests. State standardized tests are intended to support inferences about student proficiency, yet aspects of their development and administration can introduce systematic and random measurement error that undermines validity and compromises reliability. Systematic influences, such as shifts in test difficulty, construct-irrelevant content, or inequitable testing conditions or test mode can distort what scores represent and thereby weaken evidence on validity. Random variation, including idiosyncratic differences in testing environments and the timing of administration, influences reliability by adding noise and inconsistency to scores. Some fluctuation across these conditions is expected, but large fluctuations can interfere with the intended interpretations and uses of test scores, particularly the drawing of year-to-year comparisons of proficiency rates.

Key finding #3. Evidence of validity and reliability in administrative data is critical for understanding equity and school climate. Administrative data, such as student discipline records, are increasingly used to monitor equity and school climate. However, aspects of data collection and reporting can introduce systematic and random error that undermines validity and compromises reliability. Systematic influences, such as inconsistent definitions of infractions or differences in how staff apply discipline codes, distort what the data represent and weaken evidence on validity. Random mistakes in entering dates, codes, or consequences add unpredictable variation that reduces reliability. Understanding and addressing these sources of error is vital for evaluators to draw sound conclusions about equity across groups and over time.

Key finding #4. Evidence of validity, reliability, and fairness shapes stakeholder confidence in high-stakes decisions. Measurement choices influence not only

accuracy and precision but also stakeholders' perceptions of fairness. Objective measures used for high-stakes decisions must be perceived as credible. If measures are viewed as unfair, this undermines the credibility of the institutions using the measures. Changes in the use of standardized tests (including tests such as the SAT and ACT) for college admissions since 2019 exemplify how perceptions of fairness can seriously alter the usefulness of measurements.

Key finding #5. Even factual survey items require evidence of validity and reliability to support sound policy conclusions. Surveys are fundamental tools for education policy because they can efficiently capture large-scale information about schools, educators, and students. They often contain both subjective items (e.g., attitudes and perceptions) and factual items (e.g., counts, percentages, and program characteristics). Factual questions are commonly assumed to be more objective and therefore less prone to error. In practice, however, factual items are not immune to the same systematic and random influences that can undermine validity and compromise reliability. Respondents' cognitive and motivational processes shape their responses, and factual questions are not purely objective.

Key finding #6. Standards of validity and reliability must be commensurate with the stakes and intended uses of the measurement. It is common for validity and reliability to be discussed as if they were fixed traits. However, validity is actually a feature of how people interpret scores and use them. As the Standards for Educational and Psychological Testing emphasize, evidence and theory must support the interpretation of scores for each proposed use. Thus, the process of constructing a validity argument entails articulating the intended interpretations and accumulating evidence to justify them. Similarly, reliability is dependent on the population and aspects of administration. While many stakeholders seek clear rules or thresholds for determining whether reliability is "acceptable," the sufficiency of evidence on reliability is relative to purpose, consequence, and error tolerance. In other words, the quality and specificity required for evidence of validity and reliability are inherently connected to the stakes and consequences attached to those interpretations.

Background

Decisions that affect students, classrooms, and schools depend on data that are trustworthy. For evaluators to make sound inferences, measurement approaches must be consistent, accurate, and fair. In measurement terms, this means minimizing both random and systematic sources of error that can obscure or bias results.[6] Users need to be confident that the indicators they rely on truly capture the qualities they intend to measure in ways that support their intended interpretations and uses. Measures that meet these expectations are said to have evidence of validity and reliability.

Validity refers to how well evidence and theory support the specific interpretations and uses of a measure. It is concerned with avoiding systematic error that can induce construct-irrelevant variance and otherwise cause scores to represent something other than the intended construct.[7] Each distinct use (e.g., measuring growth or change, appraising differences across student subgroups, classifying schools, assessing school or teacher accountability) requires its own evidence base. All measures have advantages and limitations, and establishing validity involves gathering data and exercising professional judgment to weigh these factors.[8] A measure is said to have evidence of validity when its limitations are not large enough

to undermine the intended interpretation or use.[9] Weaknesses are problematic if they are deemed large enough to interfere with intended interpretations and uses.

The Standards for Educational and Psychological Testing specify four broad sources of evidence of validity, each bringing a different perspective on the question of whether a measure works as intended:

> **1. Evidence based on content.** Whether evidence is content based is determined by whether the assessment tasks or items adequately represent the domain or construct being measured in terms of both their relevance to the domain and the adequacy with which those tasks or items represent the domain in its entirety.[10] For example, on an end-of-year summative math test intended to determine proficiency, the included items should reflect the full range of grade-level content standards and not simply focus on a small subset of standards. Evidence based on content is often collected by convention of a panel of experts to examine the topics and items included on an assessment.
>
> **2. Evidence based on response processes.** This kind of evidence refers to how test-takers or raters engage with the assessment to confirm that the mental and behavioral processes of test takers are consistent with the constructs or domains the test is intended to measure. Suppose a mathematics task is meant to gauge how well students use graphs to solve problems. If observations show that students achieve correct answers by calculating values in a table or using trial-and-error methods instead, the task may not be measuring the construct as intended. Such information can be collected, for example, through cognitive interviews with respondents as they complete items.
>
> **3. Evidence based on internal structure.** Here, analysts examine evidence that items, scales, or domains within the assessment relate to each other in theoretically anticipated ways. For example, in an assessment that claims to measure self-efficacy, student–teacher relationships, and growth mindset, we would expect to see higher correlations among items measuring the same domain (e.g., two items measuring self-efficacy) than among items measuring different domains (e.g., one item measuring growth mindset and one item measuring student–teacher relationships). Evidence based on internal structure is often collected by means of factor analysis or item response theory (IRT), statistical tools for exploring relationships among indicators of complex concepts.
>
> **4. Evidence based on relations to other variables.** This kind of evidence relates to how assessment results connect with other measures or outcomes. For example, if a survey measures a student's engagement in school, survey-based variables would be expected to show positive correlations with other indicators of engagement, such as academic self-efficacy, attendance, and intrinsic motivation. Evidence based on relations with other variables is often collected by means of multivariate regression models.

Woven throughout these four sources of evidence is an attention to concerns of fairness, which the Standards frame as an essential component of validity.[11] Fairness involves ensuring that measure interpretations are equitable and comparable across relevant populations. For student assessment, these populations include students who differ by gender, race, ethnicity, socioeconomic background, language status, or disability. For teacher evaluation, they include teachers who differ by school context and grades and subjects taught. Evaluating fairness therefore requires gathering

evidence that the measure assesses the intended construct consistently for all groups and is relatively free from systematic error that biases the scores of certain populations. Evidence of fairness can come from statistical analyses of differential item functioning (a procedure that tests whether individuals from different groups with the same overall ability have differing probabilities of answering specific items correctly), for example. When unfairness exists, validity is compromised. Thus, attention to fairness is integral to the broader validity argument for any measure used in education policy and practice.

Reliability refers to the precision or stability of scores produced by a measure—that is, the extent to which a measure is consistent and largely free from random measurement error. Random error reflects unpredictable fluctuations in scores caused by chance influences—such as temporary conditions, sampling of items, or small variations among raters—that affect the precision of measurement.[12] In simple terms, reliability reflects a measure's ability to yield similar results across items, occasions, settings, or evaluators. Every measure contains potential sources of error that induce random score fluctuations. For example, an assessment designed to assess a student's ability to add and subtract fractions might include ten items, though many other items could plausibly have been chosen to represent the same construct. Because specific items are sampled from a larger universe of possible items, responses may vary randomly depending on which items are used. Similar variability can arise from the occasion of measurement: Scores may differ from one day to another simply because of temporal factors. In assessments involving human raters, such as scoring open-ended responses, the raters themselves are viewed as a sample from a broader population of possible evaluators. Some raters may be slightly more lenient and others more stringent, which could introduce chance variation unrelated to the construct of interest. Evidence about the magnitude of these errors of measurement provides an opportunity for researchers to judge whether the errors are sufficiently large to interfere with intended uses.[13] Excessively large errors must be mitigated in some way, either by means of statistical adjustments or revisions to the measurement process.

The concepts of validity and reliability are interrelated and reinforcing. Evidence of reliability helps users judge whether differences in scores are likely to represent real variation rather than random error or inconsistency. At the same time, accumulating evidence of validity involves demonstrating the extent to which systematic error—bias or construct-irrelevant variance—has been minimized. Low reliability can seriously limit confidence in any interpretation or policy decision based on the results.

Evidence

Key finding #1. Evidence of validity and reliability is essential for understanding cause-and-effect relationships between policies and outcomes. Policymakers must have confidence that evidence of an intervention's success (or failure) reflects genuine change rather than artifacts of measurement error or shifting interpretations of what is being measured. When innovations are found to have statistically significant effects, evidence of validity helps ensure that these results are accurate and do not merely result from changes in how students interact with instruments. When analyses yield nonsignificant effects, reliable measurement helps rule out random error as the explanation.

We illustrate how aspects of validity and reliability influence causal interpretation through an example situated in a teacher preparation program.

Researchers often rely on self-report surveys to assess outcomes such as social and emotional learning.[14,15] Between 2021 and 2023, RAND researchers conducted an evaluation of Teach For Nigeria (TFN), the local partner of the global Teach For All network, which recruits and trains promising graduates to teach for two years in underresourced schools. TFN aims to improve classroom instruction and develop long-term leadership capacity in response to Nigeria's education crisis, which is characterized by low literacy and numeracy levels among students and limited support for teachers.[16] Central to the organization's theory of change is the premise that students with TFN teachers will show greater academic, social, and emotional growth than comparable peers.

Using a quasi-experimental design that compared students taught by TFN teachers to similar students taught by teachers who had entered the profession through traditional pathways, the study aimed to examine the impacts of TFN teachers on students' academic, social, and emotional growth, consistent with this theory of change. All students completed a survey on growth mindset, self-efficacy, and social awareness at the beginning and end of the school year. At baseline, the outcomes in the two groups were equivalent, but by year's end, the treatment group (i.e., students with TFN teachers) reported lower self-efficacy scores than the comparison group.

At first glance, such results appear to provide evidence contradicting the program's theory of change. But how should this evidence be interpreted? Three explanations are plausible.

1. The first interpretation is that the TFN program genuinely reduced self-efficacy, meaning the theory linking TFN's teaching approach to students' social-emotional growth is flawed. This could represent theory failure, in which the intervention's underlying assumptions are invalid.[17]
2. A second explanation is weak implementation. Even if the theory is sound, evidence of program impact depends on the fidelity with which it is delivered. Novice TFN teachers may have struggled to integrate ambitious whole-child instructional practices, resulting in limited or inconsistent classroom application—an implementation failure that compromises internal validity.
3. A third possibility centers on the measurement itself. Because the evaluation relied on student self-reports, the outcomes may have been influenced by reference bias or response shift—changes in students' internal standards or understanding of the constructs being measured due to participation in the program. Exposure to teachers who explicitly emphasized self-efficacy may have caused students to redefine what it means to be "self-efficacious." In this case, lower end-of-year scores may not have signaled diminished self-efficacy but a more critical self-assessment. The measurement would then no longer capture the same construct at both points in time, compromising validity and clouding causal interpretation.

Such response shifts have been documented in education research for nearly fifty years.[18] In studies of social-emotional development, students in more academically and behaviorally demanding schools often report lower scores on surveys, not necessarily because those traits decline but because students reconfigure their definition of what these constructs mean and subsequently rate themselves more critically.[19] Evaluations of antibullying initiatives often find increased reports of bullying behaviors after implementation not because bullying worsened but because heightened awareness led students to recategorize incidents as bullying.[20]

In each of these cases, reliable and valid inference depends on distinguishing real changes in the underlying construct from changes in measurement. Without such distinction, even well-designed quasi-experimental or randomized studies can produce misleading conclusions. Evidence assumed to test a theory of change may merely reflect measurement artifacts.[21] Additionally, noisy or unreliable measures could obscure genuine effects, leading to underestimation of a policy's benefits.

The TFN example illustrates why evaluations of the validity and reliability of measures are not mere technical concerns but central determinants of the credibility of policy evidence. Misinterpretation of self-reported data can not only limit a study's contribution to knowledge but also compromise strategic decision-making. [22] Evidence of validity and reliability, by contrast, provides the confidence necessary to make well-founded causal claims.

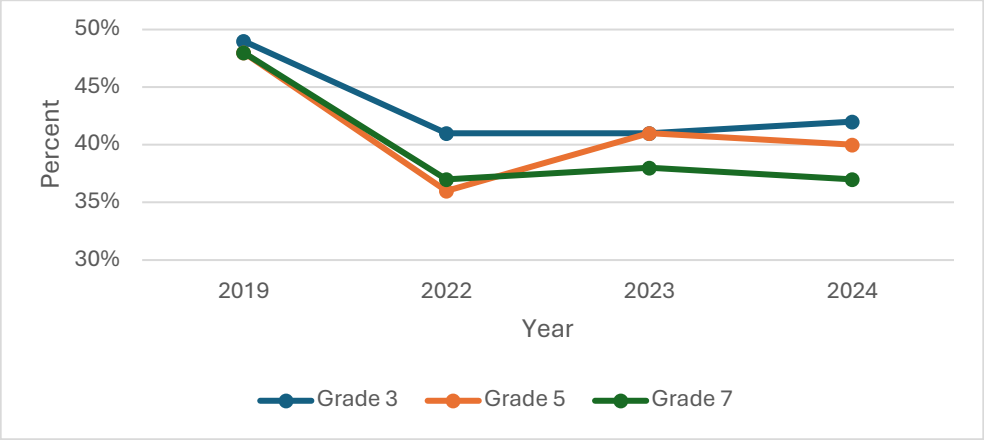
Key finding #2: Evidence of validity and reliability is critical for student proficiency trends on state standardized tests. State standardized tests are intended to support inferences about student proficiency, yet aspects of their development and administration can introduce measurement error that undermines evidence of validity and compromises reliability. Systematic influences, such as shifts in test difficulty, construct-irrelevant content, or inequitable testing conditions or test mode can distort what scores represent and thereby weaken evidence on validity. Random variation, including idiosyncratic differences in testing environments and the timing of administration, influences reliability by adding noise and inconsistency to scores. Some fluctuation across these conditions is expected, but large fluctuations can interfere with the intended interpretations and uses of test scores, particularly the drawing of year-to-year comparisons of proficiency rates. For example, examining whether this year's seventh-grade students appear more or less proficient than last year's seventh-grade students can offer insight into broad trends only when other sources of score variation have been examined and accounted for.

Proficiency rates themselves are limited indicators of achievement.[23] As Ho (2008) notes, the percentage of students deemed proficient depends on policy decisions about cut scores rather than any natural learning threshold. Because proficiency rates are sensitive to both test difficulty and arbitrary thresholds, they can give a distorted picture of achievement and raise concerns about the validity of these interpretations.

We illustrate how one aspect of sampling, test difficulty, can influence inferences about student progress through an example situated in the Massachusetts Comprehensive Assessment System (MCAS). Since the 2001 No Child Left Behind Act, statewide standardized testing has been a prominent feature of U.S. public education, providing policymakers with evidence on whether students meet state standards in subjects such as mathematics and reading.[24] In the wake of pandemic-related school closures, such assessments have also supported monitoring of educational recovery. Beginning in the 2020–2021 school year, the U.S. Department of Education (U.S. ED) encouraged states to use annual tests to appraise improvements in student progress, identify areas of persistent struggle, and inform resource allocation.[25]

In 2019, 49% of Grade 3 students met or exceeded expectations on the MCAS math assessment (Figure 1). By 2024, this figure was 42%, a seven-point decline. Grade 7 students showed a similar drop, from 48% to 39%, leading researchers to conclude that students in Massachusetts are still lagging behind prepandemic levels.[26]

Figure 1: Percentage of students meeting or exceeding expectations, MCAS math



Source: https://profiles.doe.mass.edu/statereport/mcas.aspx

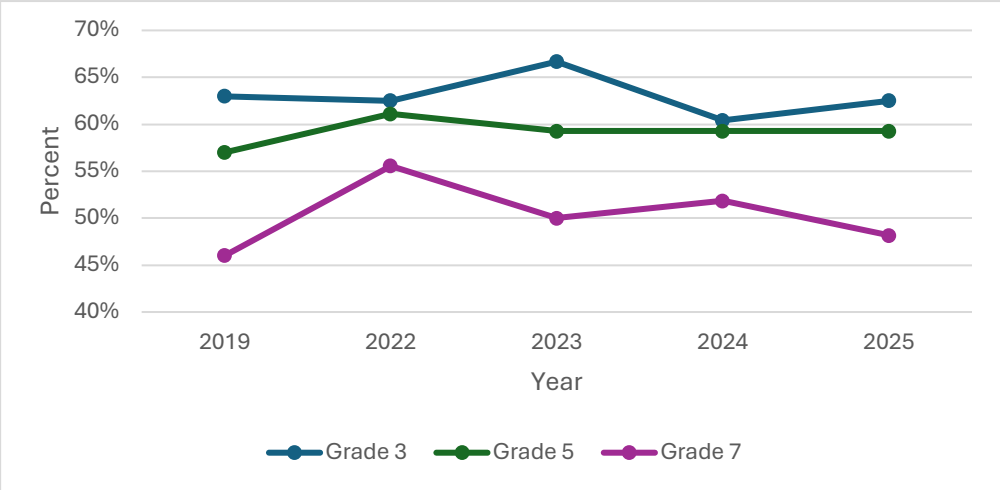
Such conclusions depend on evidence that year-to-year differences accurately reflect changes in student learning and not changes introduced by extraneous factors. Environmental conditions, time of administration, and incidental distractions are typically random influences that affect reliability. By contrast, systematic differences in test difficulty, scoring, or content sampling can undermine evidence of validity by introducing construct-irrelevant variance. Policymakers often read annual proficiency trends such as those in Figure 1 as indicators of progress. But if an assessment in one year is substantively more difficult than in another, fewer students will appear proficient even if true learning has not changed. Conversely, easier tests could make recovery appear stronger than it really is. Because test forms are assembled by sampling questions from large item pools, variation in difficulty is inevitable: Some forms will, by chance, be harder or easier than others.

Test developers therefore aim to ensure that such error is small enough not to undermine intended interpretations.[27,28] To ensure that judgments about changes in proficiency are accurate and do not reflect systematic differences in test design or administration, testing programs use psychometric procedures to equate tests across years, mapping raw scores to a common scale. In practice, this means that the raw score (the number of "correct" items) needed to reach proficiency will be slightly higher in some years and slightly lower in others. On MCAS Grade 7 math, for example, the minimum raw score needed to meet expectations for Grade 7 jumped from 46% correct in 2019 to 56% in 2022 and then dropped to 50% in 2023. Focusing only on that decline might suggest a "lower bar" for proficiency, but that interpretation misunderstands the purpose of the equating. Such adjustments do not inflate results by enabling more students to "pass"; rather, they increase confidence that year-to-year changes reflect true changes in student proficiency. Nevertheless, misinterpretation of equating adjustments is common. News stories have portrayed changes to raw scores as evidence of artificial progress. Reporting on similar recalibrations in New York, Albany's Times Union and the New York Post suggested that adjusted scoring raised "questions about boasts of substantial academic progress."[29]

The MCAS example illustrates how test difficulty might influence interpretations of annual changes in proficiency rates. Because statewide tests carry high stakes for accountability decisions, weak evidence of validity or low reliability can have

disproportionate consequences. Ensuring that random and systematic errors are sufficiently small to prevent their interfering with the scores' intended interpretations strengthens trust in proficiency data. Evidence of reliability confirms that sampling and random fluctuations do not unduly influence results, and evidence of validity demonstrates that construct representation is consistent across time and context. Together, they provide a strong foundation for sound judgments about student progress.

Figure 2: Raw score needed to meet expectations, MCAS math



Source: https://profiles.doe.mass.edu/statereport/mcas.aspx

Key finding 3: Evidence of validity and reliability in administrative data is critical for understanding equity and school climate. Administrative data, such as student discipline records, are increasingly used to monitor equity and school climate. However, aspects of data collection and reporting can introduce systematic and random error that undermines validity and compromises reliability. Systematic influences, such as inconsistent definitions of infractions or differences in how staff apply discipline codes, distort what the data represent and weaken validity evidence. Random mistakes in entering dates, codes, or consequences add unpredictable variation that reduces reliability. Understanding and addressing these sources of error is vital for evaluators to draw sound conclusions about equity across groups and over time.

Under ESSA, all states are required to publicly report student discipline information, including exclusionary actions such as suspensions and expulsions, and incidents of school violence. These data are among the few administrative indicators reported with near universal coverage, and many states, including California, Rhode Island, and West Virginia, incorporate these rates into accountability systems to inform local improvement efforts.[30,31] Moreover, discipline outcomes are widely viewed as central indicators of school climate and fairness, and reducing disproportionality in discipline is a core focus of equity improvement initiatives nationwide. At the same time, discipline data are particularly susceptible to inconsistent documentation and systematic differences in how incidents are recorded or categorized. When these data are recorded accurately, they can illuminate disparities in student experiences and

help ensure that policies promote fairness; when reporting practices vary substantially, the resulting inferences about equity can be unreliable or misleading.

We highlight the operational and procedural challenges of capturing accurate discipline data through an example situated in a large, diverse school district, Midstate Unified School District, in a western state. Approximately 10% of its students are African American, nearly one-third are Hispanic, and about 45% are socioeconomically disadvantaged. Suspension rates vary markedly across groups: Roughly 10% of African American students are suspended each year, while only 3% of White students are. African American students also constitute over 40% of recorded incidents involving physical fights, even though they comprise only 10% of the population. At face value, such figures might suggest behavioral differences among subgroups. However, differences in recorded data may also arise from both random and systematic error in the measurement process itself.

The process of recording the disciplinary infraction includes multiple opportunities to introduce two distinct types of error: random measurement error, which stems from unpredictable mistakes or inconsistencies that impact reliability, and systematic error, which reflects patterned or construct-irrelevant influences that threaten validity. Typically, staff record each incident along with details such as the date, location, student identifiers, infraction type, and assigned consequence. Random measurement error could be introduced when administrators or others responsible for inputting the data into the record system accidentally input the wrong information. For example, they could input the incorrect date of an incident, or they could mistype the code for the type of infraction. More problematic are sources of systematic error that result from human discretion and subjective interpretation. For instance, staff may interpret behaviors or apply rules across schools or student groups differently. These sources of error could lead to misleading conclusions about whether observed disparities reflect true differences or inconsistencies in documentation.

In discipline systems, discretion arises at two points: when educators decide which infraction code best describes an incident and when they determine the appropriate consequence. Districts typically provide a menu of codes for physical aggression, disruption, or insubordination, but the application of definitions is rarely enforced rigorously. Two educators might code similar behaviors differently—one as “minor disruption,” another as “physical aggression.” Similarly, selecting the consequence (e.g., detention, in-school suspension, or restorative practice) often depends on local norms and individual judgment. Research suggests that implicit bias can influence both classification and sanctioning decisions, leading educators to view the same behaviors as more severe when they are exhibited by Black students and to recommend harsher penalties.[32,33] Because these decisions determine the inputs that ultimately populate administrative databases, bias at the point of data entry becomes bias in the data itself.

The random and systematic errors that result from such processes have important consequences for the validity of inferences drawn from discipline data. Random error influences reliability and makes subgroup comparisons less precise. Systematic error induced by educator bias undermines the validity of discipline data, making it harder—or even inappropriate—to use these data to inform policy. In efforts to assess racial disparities in student discipline over time, the potential for bias in the recording of both the behavior and the associated consequence limits evaluators' ability to pinpoint the source of the disparities and develop appropriate policy responses.

For example, if disparities stem primarily from educator bias, schools aspiring to address racial inequality in exclusionary discipline may consider undertaking teacher

bias training to reduce bias. If differences reflect genuine behavioral variation, supports might instead rely on interventions that proactively prevent encounters with the disciplinary system or alternatives to exclusionary discipline such as restorative justice programs. When measurement error prevents policymakers from distinguishing between these explanations, they cannot accurately diagnose the problem or select the most appropriate remedy.

Improving the quality of discipline data therefore requires strategies aimed at minimizing systematic error. Districts can establish standardized definitions for infractions, specifying clear decision rules for classifying incidents, and auditing data for consistency across schools. Organizationally, investments in professional learning that increase awareness of implicit bias and promote reflective discipline practices can reduce the likelihood of subjective judgments distorting the data. Together, these actions improve the fairness and interpretability of disciplinary indicators and minimize the opportunity for systematic error.
The Midstate Unified example underscores that high-quality measurement is indispensable for policymaking. While discipline data is comprehensively collected and readily available, making it a critical source of information to monitor the overuse of exclusionary discipline, it must be accurately recorded in administrative datasets for it to be useful in informing policy decisions.

Key finding 4: Evidence of validity, reliability, and fairness shapes stakeholder confidence in high-stakes decisions. Measurement choices influence not only accuracy and precision but also stakeholders' perceptions of fairness. Objective measures used for high-stakes decisions must be perceived as credible. If measures are viewed as unfair, this undermines the credibility of the institutions using the measures. Changes in the use of standardized tests (including tests such as the SAT and ACT) for college admissions since 2019 exemplify how perceptions of fairness can seriously alter the usefulness of measurements.

In the United States, colleges have long used standardized tests as a factor in admissions decisions. These tests include the SAT, developed by the College Board and administered by ETS, and ACT, developed and administered by ACT Incorporated.[34] Scholars have amassed a considerable body of research examining the validity of such tests in this context.[35,36] This includes evidence based on content[37] and evidence based on relations to other variables, as they predict student success in college and later outcomes. [38 , 39] However, persistent disparities in test performance by student background characteristics—particularly race and socioeconomic status—have led to criticism regarding fairness. [4041] To address these concerns, the College Board introduced an "adversity score" in 2019 intended to contextualize SAT results.[42] The score combined environmental indicators such as median family income, local crime rates, and school resources into a single numerical estimate of each student's exposure to adversity. The score was provided to colleges alongside each applicant's SAT score. Understanding where students came from could allow colleges to situate specific scores in a relevant context—an attempt to improve transparency around the factors that contribute to systemic disparities in test-based performance. However, in practice, the fairness of the adversity score itself was called into question.

Although its intent was to increase fairness and transparency, the scoring system raised new concerns. Students could not access or verify their adversity scores, and their calculation method was not publicly disclosed. The lack of transparency undermined the measure's perceived objectivity and prompted widespread skepticism. The College Board quickly recognized and responded to the criticism, discontinuing the adversity score within three months and noting that "the idea of a single score was confusing because it seemed that all of the sudden the College Board

was trying to score adversity. That's not the College Board's mission. The College Board scores achievement, not adversity."[43] The College Board replaced the adversity score with a more transparent contextual data system that avoided summarizing multiple factors into a single index. This episode illustrates how even well-intentioned efforts to enhance fairness can lose legitimacy when users perceive the measurement process as unfair.

The COVID-19 pandemic accelerated a second shift in college admissions. Testing disruptions in 2020 led many institutions to suspend or eliminate standardized testing requirements, building on preexisting debates over fairness.[44] The University of California (UC) system, for instance, has not considered test scores in its admissions processes since 2020. The UC's permanent move to "test-blind" admissions occurred after a lawsuit argued that standardized test scores disadvantaged certain students.[45] While the plaintiffs in the lawsuit made the case that the use of the SAT and ACT had a disparate negative impact on marginalized students, it did not consider the alternative measures that UC may use in the absence of test scores and whether these alternatives could be worse for applicants.

While the decision not to consider test scores in admissions aimed to promote equity, there may be unintended consequences. In the absence of test scores, admissions committees have less information about students, which some argue could harm applicants of disadvantaged backgrounds.[46] Additionally, admissions committees may rely more heavily on alternative indicators such as high-school GPA or course titles, which vary substantially across contexts and may themselves be unreliable measures of college readiness. In the presence of imperfect information on applicants, colleges may use observable characteristics of individuals as a proxy for unobservable characteristics that correlate to the outcome of interest, such as academic capability.[47] A recent report from UC San Diego found reduced academic preparedness of incoming college students under test-blind admissions.[48] One hypothesis is that without test scores, the admissions committee had more difficulty identifying students likely to succeed at UCSD. If this is the case, the perception that test-based measures were not fair may have had significant unintended consequences—for both individuals and policy.

Key finding 5: Even factual survey items require evidence of validity and reliability to support sound policy conclusions. Surveys are fundamental tools for education policy because they can efficiently capture large-scale information about schools, educators, and students.[49] They often contain both subjective items (e.g., attitudes and perceptions) and factual items (e.g., counts, percentages, and program characteristics). Factual questions are commonly assumed to be more objective and therefore less prone to error. In practice, however, factual items are not immune to the same systematic and random influences that can undermine validity and compromise reliability. Respondents' cognitive and motivational processes shape their responses, and factual questions are not purely objective.

Federal and state agencies frequently rely on survey data to guide policy decisions by describing key aspects of instructional conditions, workforce characteristics, and school organization.[50] The Schools and Staffing Survey (SASS, now the National Teacher and Principal Survey), and RAND's American Educator Panels, are two prominent examples of surveys designed to provide timely insights into how practitioners respond to policy reforms, offering evidence that can recalibrate standards, supports, and implementation strategies.[51]

These surveys provide both subjective and objective information: They capture beliefs and perceptions not easily assessed by other methods[52] and factual information about

enrollment, staffing, budget, and programs. However, surveys are also susceptible to biases that threaten the validity of inferences. Response biases, including acquiescence bias (a tendency to agree with all items), halo effects (positive perceptions about one topic influencing perceptions of other topics), and social desirability bias are well documented in survey research[53,54] and may particularly impact items that focus on sensitive topics.[55,56]

We illustrate how measurement error can influence objective, fact-based survey items using RAND's American Mathematics Educator Study (AMES). AMES annually administers surveys to nationally representative samples of teachers and principals through the American Educator Panels specifically related to their mathematics policies and instructional practices.[57]

In the 2025 AMES administration, principals were asked to report the percentage of students in their school that were proficient in math last year (Figure 3). Information from this survey item can be used to explore relationships among school mathematics policies and practices and student math proficiency.

Figure 3: Survey item

What approximate proportion of students at your school achieved proficiency in mathematics on your last standardized mathematics assessment?

○ 0–10% ○ 11–20% ○ 21–30% ○ 31–40% ○ 41–50%	○ 51–60% ○ 61–70% ○ 71–80% ○ 81–90% ○ 91–100%	○ I don't know ○ Not applicable because students at my school do not take a state standardized mathematics assessment

By proficiency, we are referring to the standard proficiency cut scores used by your state for their standardized mathematics assessment. If students in your school have taken a standardized mathematics assessment that does not provide a standard proficiency cut score, do your best to estimate the percentage of students who achieved proficiency, defined as solid academic performance and competency over challenging subject matter.

In contrast to subjective items, factual items can often be validated against external data sources. To validate this item, the principals' responses were compared with publicly available proficiency data from state departments of education in the four largest states represented in the sample—California, New York, Florida, and Texas. For each principal, the responses were classified as (a) accurately aligned with the school's true proficiency range, (b) within five or ten percentage points of the true rate, or (c) over- or underestimates of the true rate.

Across all states, less than half of the principals reported math proficiency rates that fell within the 10-percentage-point range they had selected (Table 1). In other words, most of the reported ranges did not include the school's actual proficiency rate. In Texas, California, and Florida, the principals were likelier to overestimate their school's proficiency; in New York, they were likelier to underestimate their school's proficiency. Even though this question involved objectively verifiable information (i.e., if the principal was not sure of her school's math proficiency rate, she could easily look it up to verify before responding to the survey item), the responses exhibited

substantial measurement error. Had we assumed the principals' reports were valid, our conclusions about the relationship between mathematics proficiency and school practices would have been misleading.

Table 1. Accuracy of principal reports of math proficiency by state

Within Range	Within +/- 5 Points	Within +/- 10 Points	Overestimate	Underestimate
9%	11%	19%	87%	4%
37%	41%	63%	24%	36%
48%	53%	75%	32%	16%
24%	26%	49%	41%	33%

Why does such error occur in factual survey items? Several mechanisms may be at work, each affecting the reliability and interpretability of results in different ways. Measurement error may be random or systematic. Recall problems and incomplete record-keeping can generate random error, reducing reliability.[lviii] Principals may not remember exact scores or may approximate on the basis of past reports.

Other mechanisms result in errors that are more systematic. Mondak (2001) finds that respondents avoid answering "I don't know" and make educated guesses on fact-based questions.[lix] What's more, these guesses are not random; they are often rooted in personal beliefs or heuristics. For example, principals who believe their instructional programs are strong may overestimate their school's performance, while those who perceive local challenges may underestimate it. Such systematic misreporting can bias correlations between perceived performance and other variables.

Whatever the source of the measurement error, policymakers and other data users must consider whether the amount of error is tolerable for the intended uses of the data.[lx] In the case of AMES, it was determined that with less than half of principals accurately reporting their school's proficiency, the data were unusable to support valid and reliable inferences about the relationships between math practices and student proficiency.

Key finding 6: Standards of validity and reliability must be commensurate with the stakes and intended uses of the measurement. It is common for validity and reliability to be discussed as if they were fixed traits. However, validity is actually a feature of how people interpret scores and use them. As the Standards for Educational and Psychological Testing [lxi] emphasize, evidence and theory must support the interpretation of scores for each proposed use. Thus, the process of constructing a validity argument entails articulating the intended interpretations and accumulating evidence to justify them.[lxii] Similarly, reliability is dependent on the population and aspects of administration.[lxiii] While many stakeholders seek clear rules or thresholds

for determining whether reliability is "acceptable,"[lxiv] the sufficiency of evidence on reliability is relative to purpose, consequence, and error tolerance.[lxv] In other words, the quality and specificity of validity and reliability evidence required are inherently connected to the stakes and consequences attached to those interpretations.

In low-stakes contexts, including formative classroom assessments or research surveys, there may be minimal consequences for individuals or institutions, and accordingly, there may be greater tolerance for error. In such contexts, the burden of evidence of validity and standards for reliability may be less stringent. On the other hand, in high-stakes systems, including those used to inform accountability ratings or personnel decisions, there are tangible consequences for individuals and institutions. In such contexts, the burden of evidence is greater. Error that undermines validity or compromises reliability can lead to spurious conclusions about schools, teachers, or students, thereby distorting decision-making. Each interpretation must be supported by commensurate levels of empirical and theoretical evidence demonstrating that errors are small enough that the intended uses remain defensible.[lxvi]

This principle also highlights a broader implication for policy: Evidence of validity and reliability gathered in low-stakes research settings may not generalize to high-stakes operational contexts. Changes in population, administration procedures, or decision consequences can affect measurement error and the credibility of inferences. [lxvii] Studies documenting evidence of validity and reliability should therefore specify the conditions and stakes under which the evidence was obtained and evaluate whether those conditions align with current operational use.

References

[1] Borsboom, Denny, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. The Concept of Validity. Psychological Review 111(4): 1061.
[2] Lissitz, Robert W., and Karen Samuelsen. 2007. A Suggested Change in Terminology and Emphasis Regarding Validity and Education. Educational Researcher 36(8): 437–448.
[3] Jacob, Brian A., and Lars Lefgren. 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. Journal of labor Economics 26, no. 1: 101–136.
[4] Popham, W. James. 2010. Everything School Leaders Need to Know About Assessment. Corwin Press.
[5] Popham, 2010. p 17.
[6] American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). 2010. Standards for Educational and Psychological Testing.
[7] AERA, APA, and NCME, 2014.
[8] Kane, Michael. Validation. 2006. Educational Measurement 4(2): 17–64.
[9] Kane, Michael. 2011. The Errors of Our Ways. Journal of Educational Measurement 48(1): 12–30.
[10] Sireci, Stephen, and Molly Faulkner-Bond. 2014. Validity Evidence Based on Test Content. Psicothema: 100–107.
[11] AERA, APA, and NCME, 2014.
[12] AERA, APA, and NCME, 2014.
[13] Kane, 2011.
[14] Augustine, Catherine H., John Engberg, Geoffrey E. Grimm, Emma Lee, Elaine Lin Wang, Karen Christianson, and Andrea A. Joseph. 2018. Can Restorative Practices Improve School Climate and Curb Suspensions? An Evaluation of the Impact of Restorative Practices in a Mid-sized Urban School District. RAND Corporation.
[15] Cipriano, Christina, Michael J. Strambler, Lauren H. Naples, Cheyeon Ha, Megan Kirk, Miranda Wood, Kaveri Sehgal, et al. 2023. The State of Evidence for Social and Emotional Learning: A Contemporary Meta-analysis of Universal School-Based SEL Interventions. Child Development 94(5): 1181–1204.
[16] Mihaly, Kata, Jonathan D. Schweig, Elaine L. Wang, and Sophie Lee. 2024. Teach For Nigeria Evaluation: Quantitative and Qualitative Study Findings. RAND Corporation. RRA1870-1.
[17] Raudenbush, Stephen W., and Sally Sadoff. 2008. Statistical Inference When Classroom Quality Is Measured with Error. Journal of Research on Educational Effectiveness 1(2): 138–154.
[18] Howard, George S., and Patrick R. Dailey. 1979. Response-Shift Bias: A Source of Contamination of Self-Report Measures. Journal of Applied Psychology 64(2): 144.
[19] West, Martin R., Matthew A. Kraft, Amy S. Finn, Rebecca E. Martin, Angela L. Duckworth, Christopher FO Gabrieli, and John DE Gabrieli. 2016. Promise and Paradox: Measuring Students' Non-cognitive Skills and the Impact of Schooling. Educational Evaluation and Policy Analysis 38(1): 148–170.
[20] Smith, J. D., B. H. Schneider, P. K. Smith, and K. Ananiadou. 2004. The Effectiveness of Whole-School Antibullying Programs: A Synthesis of Evaluation Research. School Psychology Review 33(4): 547–60.
[21] Bereiter, Carl, and C. W. Harris. 1963. Problems in Measuring Change. University of Wisconsin at Madison.
[22] Fokkema, Marjolein, Niels Smits, Henk Kelderman, and Pim Cuijpers. 2013. Response Shifts in Mental Health Interventions: An Illustration of Longitudinal Measurement Invariance. Psychological Assessment 25(2): 520.
[23] Ho, Andrew Dean. 2008. The Problem with "Proficiency": Limitations of Statistics and Policy Under No Child Left Behind. Educational Researcher 37(6): 351–360.
[24] Perie, Marianne, Scott Marion, and Brian Gong. 2009. Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments. Educational Measurement: Issues and Practice 28(3): 5–13.

[25] U.S. Department of Education, Office of Elementary and Secondary Education, State Requests for Waivers of ESEA Provisions for SSA-Administered Programs, last modified June 29, 2021.
[26] https://cepr.harvard.edu/news/2025/09/massachusetts-students-still-lag-behind-pre-pandemic-mcas-scores
[27] Kane, 2011.
[28] Perrie, 2009.
[29] https://nypost.com/2025/08/16/us-news/nys-lowered-the-bar-for-some-students-to-pass-2025-reading-math-exams/
[30] Kostyo, Stephen, Jessica Cardichon, and Linda Darling-Hammond. 2018. Building a Positive School Climate. Making ESSA's Equity Promise Real: State Strategies to Close the Opportunity Gap. Research Brief. Learning Policy Institute.
[31] Holahan, Cathy, and Brooklyn Batey. 2019. Measuring School Climate and Social and Emotional Learning and Development: A Navigation Guide for States and Districts. Council of Chief State School Officers.
[32] Okonofua, Jason A., and Jennifer L. Eberhardt. 2015. Two Strikes: Race and the Disciplining of Young Students. Psychological Science 26(5): 617–624.
[33] Owens, Jayanti, and Sara S. McLanahan. 2020. Unpacking the Drivers of Racial Disparities in School Suspension and Expulsion. Social Forces 98(4): 1548–1577.
[34] Amrein-Beardsley, Audrey, Zarrina T. Azizova, Norman P. Gibbs, Chukwu Ikegwuonu, Jeongeun Kim, Deborah Michele La Torre, Matthew R. Lavery, Margarita Pivovarova, and Yi Zheng. 2025. A Validation Review of the SAT and ACT for College and University Admissions Decisions. Education Policy Analysis Archives 33(28): n28.
[35] Amrein-Beardsley, et al., 2025.
[36] Sackett, Paul R., and Nathan R. Kuncel. 2018. Eight Myths About Standardized Admissions Testing. Measuring Success: Testing, Grades, and the Future of College Admissions, 13–39.
[37] Shaw, E. J. 2015. An SAT® Validity Primer. College Board.
[38] Friedman, John N., Bruce Sacerdote, Douglas O. Staiger, and Michele Tine. 2014., Standardized Test Scores and Academic Performance at Ivy Plus Colleges. In AEA Papers and Proceedings 115: 676–681.
[39] Chetty, Raj, David J. Deming, and John N. Friedman. 2025. Diversifying Society's Leaders? The Determinants and Causal Effects of Admission to Highly Selective Private Colleges. The Quarterly Journal of Economics: qjaf050.
[40] Rothstein, Jesse M. 2004. College Performance Predictions and the SAT. Journal of Econometrics 121(1-2): 297–317.
[41] Kamenetz, Anya. 2019. College Board to Give Students "Adversity Score" Based on Social and Economic Factors. NPR.
[42] Goldstein, Dana. 2019. Your Questions About the New Adversity Score on the SAT, Answered. The New York Times.
[43] Allyn, Bobby. 2019. College Board Drops Its "Adversity Score" for Each Student After Backlash. NPR.
[44] Bennett, Christopher T. 2022. Untested Admissions: Examining Changes in Application Behaviors and Student Demographics Under Test-Optional Policies. American Educational Research Journal 59(1): 180–216.
[45] Middleton, K.V., Omonkhodion, C.H., Amoateng, E.Y., Okam, L.O., Cardoza, D. and Oakley, A. 2024. From Mandated to Test-Optional College Admissions Testing: Where Do We Go from Here? Educational Measurement: Issues and Practice, 43(4): 33–37.
[46] Sacerdote, Bruce, Douglas O. Staiger, and Michele Tine. 2025. How Test Optional Policies in College Admissions Disproportionately Harm High Achieving Applicants from Disadvantaged Backgrounds. Working Paper No. w33389. National Bureau of Economic Research.
[47] Arrow, Kenneth J. 1973. Information and Economic Behavior.
[48] https://senate.ucsd.edu/media/740347/sawg-report-on-admissions-review-docs.pdf
[49] Desimone, Laura M. 2009. Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures. Educational Researcher 38(3): 181–199.
[50] Goldring, Rebecca, Soheyla Taie, and Minsun Riddles. 2014. Teacher Attrition and Mobility: Results from the 2012-13 Teacher Follow-up Survey. First Look. NCES 2014-077. National Center for Education Statistics.

[51] Hamilton, Laura S., Julia H. Kaufman, and Melissa Diliberti. 2020. Teaching and Leading Through a Pandemic: Key Findings from the American Educator Panels Spring 2020 COVID-19 Surveys. Data Note: Insights from the American Educator Panels. Research Report. RR-A168-2. RAND Corporation.
[52] Walston, Jill, Jeremy Redford, and Monica P. Bhatt. 2017. Workshop on Survey Methods in Education Research: Facilitator's Guide and Resources. REL 2017-214. Regional Educational Laboratory Midwest.
[53] Krosnick, Jon A. 1991. Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. Applied cognitive psychology 5(3): 213–236.
[54] Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. The Psychology of Survey Response.
[55] Tourangeau, Roger, and Ting Yan. 2007. Sensitive Questions in Surveys. Psychological Bulletin 133(5): 859.
[56] Krumpal, Ivar. 2013. Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. Quality & quantity 47(4): 2025–2047.
[57] Schweig, Jonathan, David Grant, Rakesh Pandey, Dorothy Seaman, Julia H. Kaufman, and Elizabeth D. Steiner. 2025. American Mathematics Educator Survey: 2025 Technical Documentation and Survey Results.
[lviii] Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. Measurement Error in Survey Data. In Handbook of Econometrics 5: 3705–3843.
[lix] Mondak, Jeffrey J. 2001. Developing Valid Knowledge Scales. American Journal of Political Science, 224–238.
[lx] Kane, 2011.
[lxi] AERA, APA, and NCME, 2014.
[lxii] Kane, 2006.
[lxiii] Michael T. Kane, 2013. Validating the Interpretations and Uses of Test Scores, Journal of Educational Measurement 50(1): 1–73.
[lxiv] Nunnally, Jum C., and Ira H. Bernstein. 1994. Psychometric Theory, 3rd ed.
[lxv] Kane, 2011.
[lxvi] Kane, 2011.
[lxvii] Ganimian, Alejandro J., Andrew D. Ho, and Alejandra Campos. 2025. The Reliability of Classroom Observations and Student Surveys in Non-Research Settings: Evidence from Argentina.