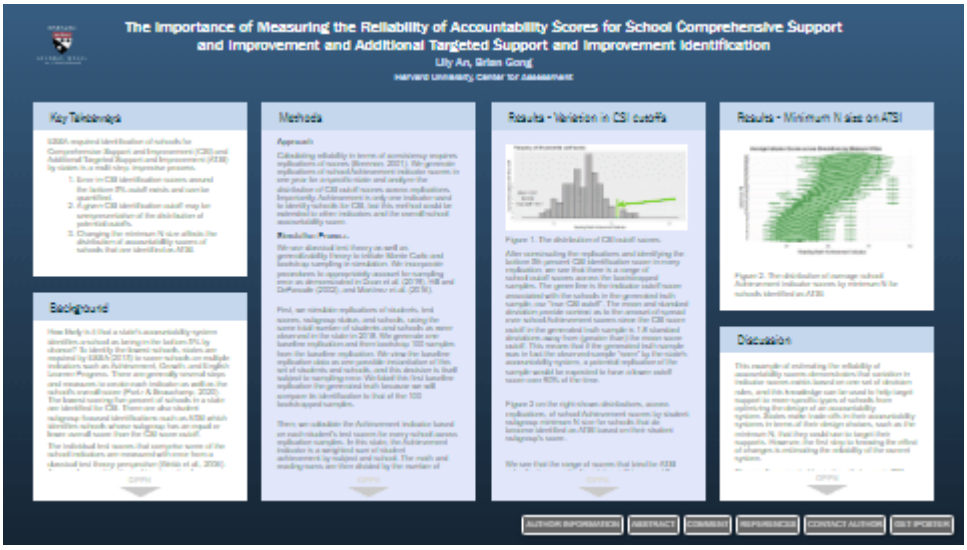
The Importance of Measuring the Reliability of Accountability Scores for School Comprehensive Support and Improvement and Additional Targeted Support and Improvement Identification



Lily An, Brian Gong

Harvard University, Center for Assessment

PRESENTED AT:

KEY TAKEAWAYS

ESSA-required identification of schools for Comprehensive Support and Improvement (CSI) and Additional Targeted Support and Improvement (ATSI) by states is a multi-step, imprecise process.

1. Error in CSI identification scores around the bottom 5% cutoff exists and can be quantified.
2. A given CSI identification cutoff may be unrepresentative of the distribution of potential cutoffs.
3. Changing the minimum N size affects the distribution of accountabillity scores of schools that are identified as ATSI.

BACKGROUND

How likely is it that a state's accountability system identifies a school as being in the bottom 5% by chance? To identify the lowest schools, states are required by ESSA (2015) to score schools on multiple indicators such as Achievement, Growth, and English Learner Progress. There are generally several steps and measures to create each indicator as well as the school's overall score (Portz & Beauchamp, 2020). The lowest-scoring five percent of schools in a state are identified for CSI. There are also student subgroup-focused identifications such as ATSI which identifies schools whose subgroup has an equal or lower overall score than the CSI score cutoff.

The individual test scores that comprise some of the school indicators are measured with error from a classical test theory perspective (Webb et al., 2006). As complex combinations of inputs, school accountability scores are also measured with error that undermines the ability of the accountability system to accurately identify schools that need support. Understanding the technical properties of accountability scores can improve alignment between intended purposes of accountability and the actualized system if policymakers are aware of them and can make changes to the identification determination processes.

The choice of minimum N count of students in any subgroup required for inclusion is an example of a lever for states to actualize their priorities: a larger minimum N runs the risk of failing to estimate subgroup performance for smaller subgroups but generates more trustworthy ATSI identifications for schools whose subgroups are large enough to meet minimum N.

METHODS

Approach

Calculating reliability in terms of consistency requires replications of scores (Brennan, 2001). We generate replications of school Achievement indicator scores in one year for a specific state and analyze the distribution of CSI cutoff scores across replications. Importantly, Achievement is only one indicator used to identify schools for CSI, but this method could be extended to other indicators and the overall school accountability score.

Simulation Process

We use classical test theory as well as generalizability theory to initiate Monte Carlo and bootstrap sampling in simulation. We incorporate procedures to appropriately account for sampling error as demonstrated in Doan et al. (2019), Hill and DePascale (2002), and Martínez et al. (2016).

First, we simulate replications of students, test scores, subgroup status, and schools, using the same total number of students and schools as were observed in the state in 2018. We generate one baseline replication and then bootstrap 100 samples from the baseline replication. We view the baseline replication data as one possible instantiation of this set of students and schools, and this decision is itself subject to sampling error. We label this first baseline replication the generated truth because we will compare its identification to that of the 100 bootstrapped samples.

Then, we calculate the Achievement indicator based on each student's test scores for every school across replication samples. In this state, the Achievement indicator is a weighted sum of student achievement by subject and school. The math and reading sums are then divided by the number of student scores in each subject. If the number is below the state's minimum N, 30 students per school, then the subject's sum is excluded from further use. The resulting values are averaged to make an overall Achievement score. We also calculate subgroup-level Achievement indicators for subgroup members. Within replication sample, the lowest five percent of school-level scores are identified as CSI. If the school subgroups' Achievement indicator is below the CSI school cutoff, then the school will be identified as needing ATSI.

Finally, to understand the effects on reliability of changing decision rules in accountability systems, we re-calculated the Achievement indicator for minimum N's from 1 to 30 at both the school and student subgroup levels, and new fifth percentile CSI cutoff scores were generated across samples.

RESULTS - VARIATION IN CSI CUTOFFS

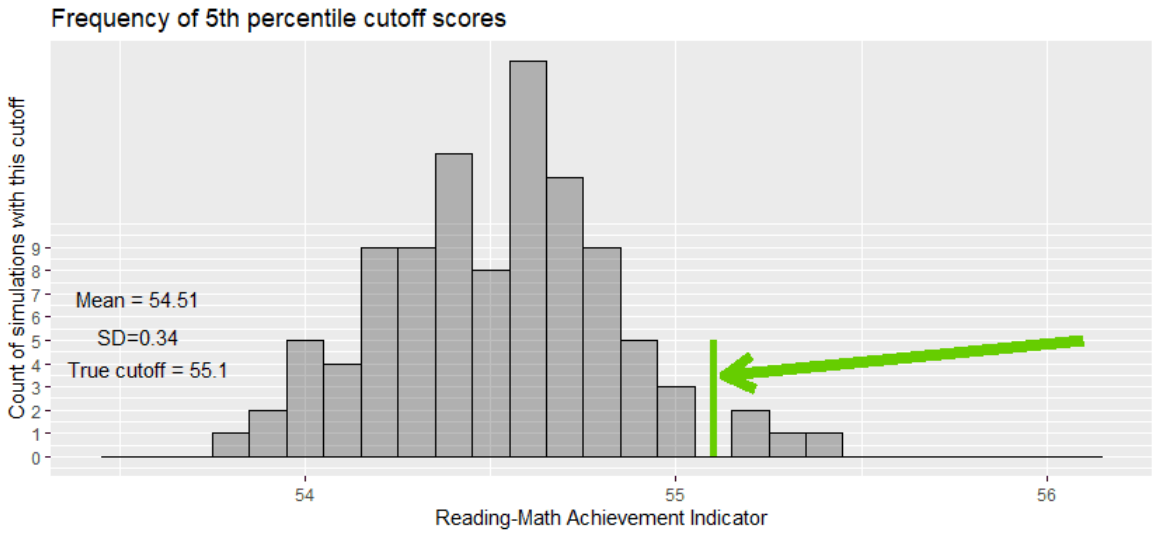


Figure 1. The distribution of CSI cutoff scores.

After constructing the replications and identifying the bottom 5th percent CSI identification score in every replication, we see that there is a range of school cutoff scores across the bootstrapped samples. The green line is the indicator cutoff score associated with the schools in the generated truth sample, our “true CSI cutoff”. The mean and standard deviation provide context as to the amount of spread over school Achievement scores since the CSI score cutoff in the generated truth sample is 1.8 standard deviations away from (greater than) the mean score cutoff. This means that if the generated truth sample was in fact the observed sample “seen” by the state's accountability system, a potential replication of the sample would be expected to have a lower cutoff score over 90% of the time.

Figure 2 on the right shows distributions, across replications, of school Achievement scores by student subgroup minimum N size for schools that do become identified as ATSI based on their student subgroup's score.

We see that the range of scores that bind for ATSI schools at any particular minimum N is around 2 Achievement indicator points. However, the distributions are not consistent across minimum N size - there is a larger variance in Achievement scores for higher minimum N sizes. This means that if the state uses a higher minimum N size, the actual score corresponding to a school being identified as ATSI can be quite different across schools. Additionally, the score distribution for smaller minimum N’s have, on average, much lower means than minimum N’s above ten. This demonstrates that changing the minimum N identifies different kinds of schools in terms of student test performance.

RESULTS - MINIMUM N SIZE ON ATSI

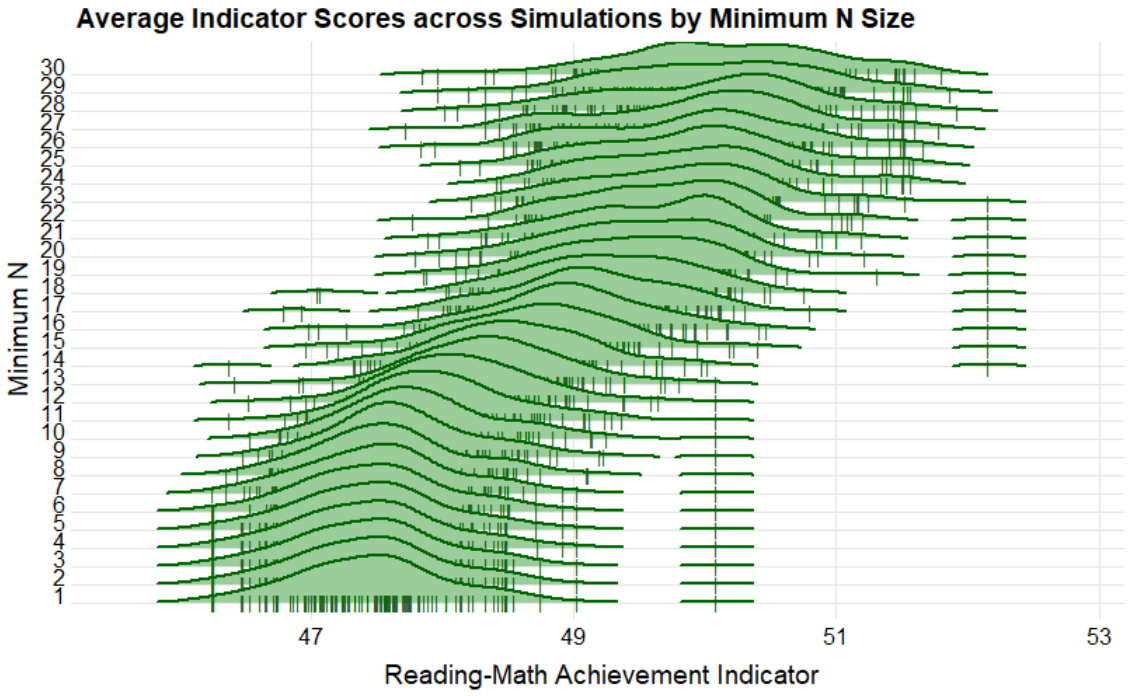


Figure 2. The distribution of average school Achievement indicator scores by minimum N for schools identified as ATSI.

DISCUSSION

This example of estimating the reliability of accountability scores demonstrates that variation in indicator scores exists based on one set of decision rules, and this knowledge can be used to help target support to more specific types of schools from optimizing the design of an accountability system. States make trade-offs in their accountability systems in terms of their design choices, such as the minimum N, that they could use to target their supports. However, the first step to knowing the effect of changes is estimating the reliability of the current system.

The results presented here show that error in CSI identification exists and can be quantified. We calculated the standard deviation around the mean cutoff score and found a disconnect between the generated truth cutoff, which represents one instantiation of possible students and student scores, and cutoffs from replications of the system. This first result is surprising and could be checked by others states using their student and school statistical properties as inputs. Importantly, decisions made using one year's observed scores should investigate the generalizability of results based on that sample of students. Policymakers should be informed about the representativeness of their CSI cutoff in the distribution of plausible cutoff scores.

Beyond demonstrating the imprecision of CSI cutoff scores, we also see an interesting policy lever for a state with regards to their minimum N value - this decision affects the distribution of Achievement scores of schools that are identified as ATSI. Therefore, this one decision rule in an accountability system could be used to help target support to specific types of schools. There are other decision rules whose effects on identification could be analyzed and quantified.

Calculations like these can help provide policmakers with an understanding of the error in accountability scores, and therefore the extent to which school accountability scores fall below the CSI cutoff by chance. There are opportunities to optimally pick indicators and set decision rules that create indicators to maximize the likelihood of the accountability system reflecting a state's priorities.

We encourage policymakers and researchers to examine technical properties of these consequential scores in their own systems. We also note several caveats to this study, the first being that only one indicator within the accountability system was calculated and ranked to identify the CSI cutoff score here. The full system is compensatory, meaning that other indicators soften the effects of any one indicator. Based on the findings of Douglas and Mislevy (2010), who estimated classification accuracy for complex rules including compensatory systems, we would expect the inclusion of other indicators to decrease the rate of false positive identification while increasing false negative identification, though effects on reliability are unclear. Additionally, our student subgroup was randomly selected among students. If subgroup scores were modeled to be consistently different than school averages, effects of minimum N changes would likely change. Both limitations will be addressed in future work.

AUTHOR INFORMATION

Lily An, Harvard University, lily_an@g.harvard.edu

Brian Gong, Center for Assessment, bgong@nciea.org

TRANSCRIPT

ABSTRACT

Estimating the precision of school accountability scores is complex, as the scores are composites of school- or subgroup-level scores. Nevertheless, accountability scores have technical properties and error similarly to test scores. Calculating their reliability can explain the extent to which identification occurs by chance. Using Generalizability theory to initiate Monte Carlo and bootstrap sampling in simulation, this study estimates the variance of school accountability scores for a specific state, finding that the generated true CSI identification cutoff score is over one standard deviation away from the average cutoff score across simulated repetitions. Additionally, as the minimum count required for inclusion of a student group in an ATSI calculation increases, lower performing schools are less likely to be identified for ATSI.

REFERENCES

Brennan, R. L. (1992). Generalizability theory. Educational Measurement: Issues and Practice, 11(4), 27-34.

Brennan, R. L. (2001). An Essay on the History and Future of Reliability from the Perspective of Replications. Journal of Educational Measurement, 38(4), 295–317. http://www.jstor.org/stable/1435452

Doan, S., Schweig, J. D., & Mihaly, K. (2019). The consistency of composite ratings of teacher effectiveness: evidence from New Mexico. American Educational Research Journal, 56(6), 2116-2146.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. Journal of Educational and Behavioral Statistics, 35(3), 280-306.

Every Student Succeeds Act, 20 U.S.C. § 1111 (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177

Gordon, N. (2017). How state ESSA accountability plans can shine a statistically sound light on more students. Evidence Speaks Reports, 2(17), 1-6.

Hill, R. K., & DePascale, C. (2002). Determining the reliability of school scores. The National Center for the Improvement of Educational Assessment Inc.

Hoffman, R. G. & Wise, L. L. (2003). The Accuracy of School Classifications for the 2002 Accountability Cycle of the Kentucky Commonwealth Accountability Testing System (Report No. FR-03-06). HumRRO for Kentucky Department of Education.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. Journal of Economic perspectives, 16(4), 91-114.

Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. Educational Evaluation and Policy Analysis, 38(4), 738-756.

Portz, J., & Beauchamp, N. (2022). Educational Accountability and State ESSA Plans. Educational Policy, 36(3), 717–747. https://doi.org/10.1177/0895904820917364

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1992). Generalizability theory. American Psychological Association.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. Handbook of statistics, 26, 81-124.

Wisconsin ESSA Plan. (2018). Wisconsin's Consolidated ESSA State Plan § 4ii. https://dpi.wi.gov/sites/default/files/imce/esea/pdf/01%2011%2018%203rd%20Resubmission%20with%20Tracked%20Changes.pdf

SCREEN TIME