APPROPRIATE STANDARDS OF EVIDENCE FOR EDUCATION POLICY DECISION MAKING

Carrie Conaway
Graduate School of Education
Harvard University
Cambridge, MA 02138
carrie_conaway@gse.harvard.edu

Dan Goldhaber
(corresponding author)
Center for Education and Data & Research
University of Washington
Seattle, WA 98103
dgoldhab@uw.edu

Abstract

Education policy makers must make decisions under uncertainty. Thus, how they think about risks has important implications for resource allocation, interventions, innovation, and the information that is provided to the public. In this policy brief we illustrate how the standard of evidence for making decisions can be quite inconsistently applied, in part because of how research findings are reported and contextualized. We argue that inconsistencies in evaluating the probabilities of risks and rewards can lead to suboptimal decisions for students. We offer suggestions for how policy makers might think about the level of confidence they need to make different types of decisions and how researchers can provide more useful information so that research might appropriately affect decision making.

https://doi.org/10.1162/edfp_a_00301

INTRODUCTION

A key job of education policy makers is to make decisions under uncertainty. They must weigh the risks, rewards, and costs of different interventions, policies, and mixes of resources, and make decisions even when the likely outcome is uncertain. Sometimes decisions are informed by an abundance of empirical evidence; in those cases policy makers might be quite certain about the consequences of the decisions they make. But often decisions must be made in instances where the evidence is unavailable or inconclusive, or the evidence may even suggest that an informed decision is likely to yield uncertain outcomes.

How policy makers think about and deal with uncertainty has important implications for resource allocation, interventions, innovation, and the information that is provided to the public. We do not presume to judge how much risk policy makers *should* feel comfortable with in the face of uncertain educational decisions. Rather, we worry that the way uncertainty is described—particularly adherence to the statistician's standard for statistical significance—may lead to misunderstandings and inconsistencies in how uncertainty affects decisions.

In this policy brief we illustrate how the standard of evidence for making decisions can be quite inconsistently applied, in part because of how research findings are reported and contextualized. Academic papers and (especially) summaries, abstracts, and policy reports frequently exclude information about uncertainty, let alone the broader context for the findings. Without this information, it is hard for policy makers to appropriately consider uncertainty in their decisions. We also argue that inconsistencies in evaluating the probabilities of risks and rewards can lead to suboptimal decisions for students because risks and rewards are often judged by how the adults, rather than the students, in the system are affected. Finally, we offer some suggestions for how policy makers might think about the level of confidence they need to make different types of decisions and how researchers can provide more useful information so that research might appropriately affect decision making.

THE USE AND NONUSE OF STATISTICAL SIGNIFICANCE IN POLICY MAKING

We are not alone in raising concerns about how policy makers consider uncertainty. Most prominently, the widely cited 2016 statement by the American Statistical Association (ASA) raises a number of issues with interpretation of, and overreliance on, *p*-values, the measure commonly used to assess the level of statistical significance. It warns that "Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold" (Wasserstein and Lazar 2016, p. 131), but also notes that "Nothing in the ASA statement is new. Statisticians and others have been sounding the alarm about these matters for decades, to little avail" (p. 130).

The fact that sounding the alarm has seemingly yielded little change may itself be justification for continuing to try to get the message out about how to think about uncertainty. But we also believe it is helpful to center this challenge within a specific policy context, to make more tangible how these issues apply to the decisions policy makers face. In this case we focus on education policy.

Testing and school accountability provide useful illustrations of how uncertainty enters the policy-making process. Under the No Child Left Behind Act (NCLB) and

Table 1. Examples of Uncertainty in Test Score Use

Test Score Use	Examples of Sources of Uncertainty	Is This Source of Uncertainty Quantifiable?	Is it Commonly Reported?
Reporting individual scores	Random error (e.g., guessing, misreading)	Yes	Yes
	Only some domains of knowledge are included on the test	Not easily	No
Determining proficiency levels	*Above plus . . .*	Not easily	No
	Human judgment on performance levels		
Reporting aggregate school performance	*Above plus . . .*	Yes	No
	Sampling error		
Making school accountability determinations	*Above plus . . .*	Yes	No
	Choices on which measures to include and how much to weight them		

now Every Student Succeeds Act (ESSA), all states are required to administer annual academic achievement tests to students to measure their proficiency. The data from these tests are used for multiple purposes, among them to measure individual student achievement and assign students to interventions, and to measure and report on teacher and school performance (see table 1). As we argue in this brief, uncertainty is inherent in the testing process—yet it is not consistently considered or reported across all these uses.

Standard psychometric practice for reporting test data (or any other statistical estimate) is to provide both an individual test score and a range of scores within which statisticians are highly confident (more on this later) that a student would receive a similar score were she to retake the test (AERA et al. 2014). This range is meant to reflect the fact that a student's test score on any given day is just an estimate of her true ability, in part because of random errors such as guessing or misreading questions (Koretz 2009). This type of error is easily quantified and thus typically included in describing test findings. But error at the individual level can also arise because tests themselves only measure a sample of a domain of knowledge; depending on the domains sampled, students may perform better or worse on the test (Koretz 2009). This is harder to quantify so is typically not represented statistically.

Criterion-referenced tests—whose purpose is to measure whether students are proficient, and not just how they perform relative to one another—also involve uncertainty in another, less obvious way. To determine which students are and are not proficient in a given subject, states must make decisions about what level of test performance is sufficiently high to meet that standard. States typically establish these thresholds by convening teams of educators to review test items and results, and to set cut points—that is, the minimum level of performance needed to attain each performance level on the test (e.g., "needs improvement," "proficient," "advanced"). As a result, the percent of students identified as proficient varies in part as a function of the differing judgments of different groups of educators, as opposed to meaningful differences in the challenge level of the content for that grade or students' preparation for learning that content. This type of uncertainty goes unnoticed because, unlike the uncertainty in student-level scores, it is not easy to quantify—yet it is critically important to how the results are interpreted and used downstream.

Similar issues arise when assessment data are aggregated to the teacher or school level to describe performance (see table 1). Every state, for instance, publishes extensive data on school-level assessment results, often along with other information, such as high school graduation rates. The provision of this information, required under both NCLB and ESSA, is a form of public accountability that is intended to inform schooling choices (Shober 2016), yet it ignores uncertainty in at least two critical ways. First, once a minimum school size threshold is met,[1] the information is typically reported publicly without any indication of a confidence interval around the results. Stated differently, sampling error is often not discussed. And second, these reports typically display the percentage of students scoring in each performance level, so rely heavily on the inherently uncertain performance-level categorizations described above. These public reports rarely, if ever, explicitly describe how uncertainty might matter for the results.

These same data are also used to rate school performance through accountability determinations. Standard practice is to assign weights to test scores, graduation rates, and other quantitative measures of school performance, to rank schools on those measures, to classify them into groups based on the rankings, and to report those classifications publicly. The designations schools receive lead to substantial rewards, sanctions, and prioritization in resource allocation. Yet, once again, this process pays little attention to whether the reported differences between schools are meaningful or how much the weighting of factors—a subjective choice—matters for the determinations.

The aggregation of test results across multiple students helps to ensure that observed differences in achievement reflect real differences in students' learning, as opposed to sampling or random errors.[2] Still, small differences between schools in test scores are almost surely *not* indicative of true underlying differences in school quality (Kane and Staiger 2002). In fact, the issues with uncertainty around small differences are so well established among researchers that a common research design for causal inference is to compare outcomes for schools just above and below an arbitrarily set cut point, on the argument that they are essentially equivalent, except for the random chance of which side of the cut they ended up on (e.g., Rockoff and Turner 2010; Rouse et al. 2013; Holden 2016). Yet the categorizations of schools affect how literally billions of dollars of education funding are allocated.

Some states do attempt to address uncertainty in some parts of the accountability process. For example, in Massachusetts's accountability system under the NCLB waiver (in place from 2012 to 2016), schools received credit for reaching their performance target if they came within half of a standard deviation. But accounting for uncertainty is not required, and thus states vary in whether and how they choose to address this issue. This leaves parents and the public with access to arbitrarily different information about school performance.

1. Here we have simplified the actual requirement that schools report the aggregate test performance of various student subgroups that exceed minimum threshold sizes. Interestingly, states vary in the thresholds they set for the minimum number of students in each subgroup for reporting requirements. Thus, the level of confidence in the differences in student achievement across subgroups varies from state to state according to the differences in their reporting thresholds.
2. Note, however, this does not alleviate concerns about whether tests are measuring the right domains, or the potential that testing itself negatively affects schools or students (Koretz 2017; Goldhaber and Özek 2019).

Thus, uncertainty is an integral part of generating, interpreting, and using assessment data, but its role and implications are inconsistently considered throughout that process. Where the uncertainty is easily quantified, it is more commonly reported—but this is only a small subset of the places where uncertainty matters for policy making. The inconsistency is troubling, considering the implications of making incorrect decisions based on test scores are arguably more profound when they are used to set proficiency levels or to drive resources and trigger interventions (precisely the cases where uncertainty is not considered).

TOWARD APPROPRIATE RESEARCH FRAMING AND STANDARDS OF EVIDENCE

Decades of academic research speaks to how managers make decisions under uncertainty (e.g., Arrow and Lind 1978; Bradley and Drechsler 2013; Goodwin and Wright 2014). But this work does not address how the typical reporting of research findings helps to frame the ways in which policy makers seek to account for uncertainty, or how the context for a decision might influence how much certainty a decision maker should seek. In particular, as we describe below, the statistician's standard for significance may serve to obscure effects to which policy makers should attend. And, the context for decision making—the policy goal, the weight of the evidence on an issue, the cost and reversibility of policy choices, and what is known about relevant policy alternatives—clearly matters for policy makers when interpreting and applying new evidence.

Standard Statistical Practice Often Doesn't Reflect Policy Makers' Needs

In academia, studies are often judged by their reliability and their internal and external validity, that is, the degree to which the study produces consistent measures, measures that it intended to measure, and generalizes to other contexts. Where a study's reliability and validity are strong, its implications for decision making are more certain. But this overlooks the fact that in the abstracts, briefs, and media reporting most accessible to policy makers,[3] findings are generally described not in terms of their reliability and validity but rather their statistical significance—and the statistician's standard for what constitutes a "significant finding" can steer policy makers in the wrong direction.

Specficially, in testing for differences between samples, the norm is to set a high standard for what constitutes a "real" difference, typically a probability (known as a *p*-value) of 5 percent or less, of stating that a difference exists when it does not. This then translates to a 95 percent confidence interval that defines the range in which the true population difference would lie with 95 percent certainty. This high standard limits the chance of finding a false positive (Type I error).

The 95 percent certainty standard is often uncritically adopted in the context of making education policy decisions. In fact, it is likely that many decision makers are unaware of the specific standards at all; they simply hear whether an initiative has a statistically significant effect or not.[4] Yet, some policy makers' decisions suggest that

3. Evidence suggests policy makers access research primarily through their professional networks (Penuel et al. 2017).

4. The standard practice when designing an experiment is to seek at least 80 percent confidence in avoiding falsely claiming that a difference doesn't exist when it really does. This in effect suggests that false positives are four times as problematic as false negatives, a standard that is certainly debatable. But we would also argue that researchers often fail to pay attention to false negatives (i.e., Type II error). This is particularly true in research

they also value avoiding false negatives. For instance, states devote substantial resources to collecting and publishing data about schools, even when the differences between them may not be meaningful.

Teacher preparation policy provides a helpful example of how policy makers might weight the risks of false positives and negatives differently from standard statistical practice. The quality of newly prepared teachers and the role of teacher preparation programs in developing teachers are issues receiving increased attention of late (Goldhaber 2019). One natural question is whether programs vary meaningfully in the effectiveness of their graduates. In fact, a number of states have begun to hold them accountable for teacher value added,[5] one measure of teacher effectiveness (von Hippel and Bellows 2018).

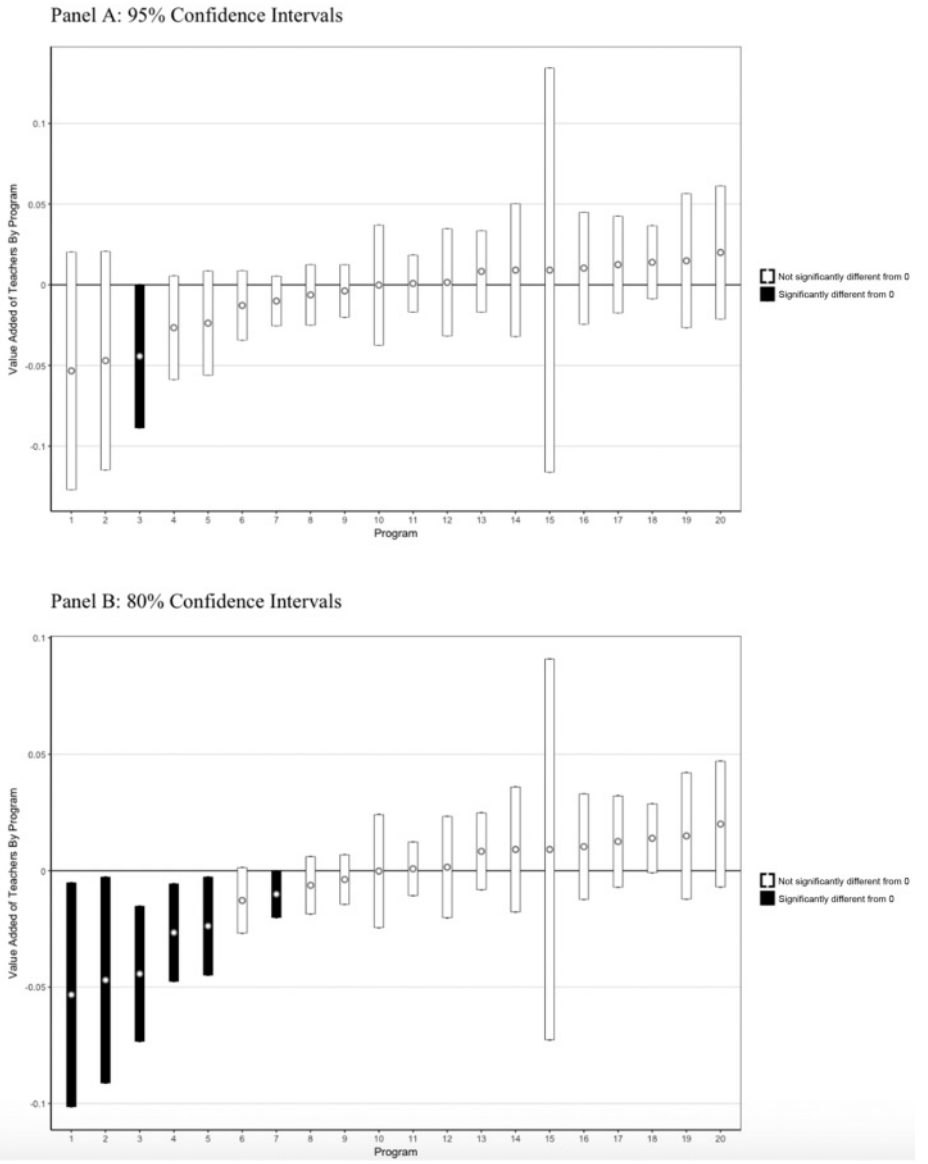
Not surprisingly, ranking teacher preparation programs is controversial, especially when it comes to rankings based on value-added measures and using these rankings for program accountability. The American Educational Research Association (AERA), for instance, released a statement raising substantial cautions about the use of value-added models to evaluate programs (AERA 2015). One of the concerns raised is that value added should "always be accompanied by estimates of uncertainty to guard against overinterpretation of differences [between programs]" (p. 50).

So, how many, and which, programs produce especially strong or weak teachers? The answer depends in large part on the statistical standards used to determine whether the differences are meaningful. In an analysis of studies from six states, von Hipple and Bellows (2018) note that few programs are different from the average in a state and conclude that "It is not meaningful to rank all the [teacher preparation programs] in a state. The true differences between most [teacher preparation programs] are too small to matter, and the estimated differences consist mostly of noise" (p. 13). But the von Hippel and Bellows conclusion is based largely on the typical statistician's standard of evidence, and a standard other than the 95 percent confidence level might yield a different conclusion. Figure 1, which is based on analysis of teacher preparation programs in Washington State (Goldhaber, Liddle, and Theobald 2013), illustrates this point.

Figure 1 shows the estimated math value added of teachers from the twenty programs in Washington State.[6] The 95 percent confidence intervals, which are shown in panel A, often overlap across programs, suggesting those programs are not readily distinguishable from one another, at least with 95 percent confidence. The 95 percent confidence intervals often also include zero (cases where it does have white bars and where it does not have black bars), here defined as the average effectiveness of teachers who transfer in from out-of-state; when this happens, the program produces graduates who are not statistically distinguishable from teachers imported from outside Washington. By this metric, no programs are significantly different from one another, and only

on nonexperimental data where the sample is fixed, creating a tradeoff between Type I and Type II errors. Studies testing against a null hypothesis using the 95 percent confidence standard often lack sufficient power to detect what might be considered to be reasonably sized treatment effects.

5. For more on value added and other measures of teacher performance, see www.carnegieknowledgenetwork.org/briefs/value-added/value-added-other-measures/.
6. The estimates reported in figure 1 are derived from the coefficients in column 1 of table 4 in Goldhaber, Liddle, and Theobald (2013).



Note: The estimates reported in figure 1 are derived from the coefficients in column 1 of table 4 in Goldhaber, Liddle, and Theobald (2013).

Figure 1. Estimated Mathematics Value Added by Teacher Preparation Programs, Washington State

one is different from zero, that is, the average out-of-state prepared teacher receiving a credential.[7]

But what if the standard were 80 percent confidence instead, as is shown in panel B? Then twelve programs are different from one another (i.e., have nonoverlapping

7. The reality of this type of comparison is more complex than we present here (for the sake of parsimony) as it involves multiple comparisons (von Hippel and Bellows 2018), but the general idea holds.

confidence intervals), and six are different from the impact of the average teacher who comes into the Washington workforce from out-of-state. What level of confidence is the right one for policy makers to use in this context? Although 95 percent confidence is the default figure, this is by no means a magic number. The right value depends critically on contextual factors, such as the anticipated behavioral responses to the identification of individual programs or the alternative policy options for judging the quality of programs.We return to these points in the next subsection.

Exacerbating these issues, reporting only magnitudes and statistical significance of findings neglects to provide other crucial information for decision making. Much of the literature on how managers make decisions, for instance, presumes the decision maker is comparing discrete potential strategies and can make a decision by comparing the probability of the outcomes from each. In reality, this type of information is often not available in a way that meets decision makers' needs.

For example, several well-executed studies now show that teachers certified by the National Board for Professional Teaching Standards (NBPTS) are more effective on average than those who are not.[8] This headline emphasizes the statistical significance of these findings. But if a policy maker were considering highlighting specific NBPTS teachers as exemplars of excellence in their community or providing them with greater compensation, a more relevant question might be: What is the probability that recognizing NBPTS teachers in my district or state would be rewarding teachers who are more effective than average? The answer to this question, at least in one context, is about 55 to 60 percent (Goldhaber 2006). Whether that rate is high or low is a value judgment, but the framing around probabilities seems more in line with how this question might be debated in policy terms than whether a finding is statistically significant.

The Policy-Making Context is Critical, Yet Frequently Overlooked

So, what standard of evidence *should* policy makers use when making policy decisions? Looking at individual studies, of course, policy makers should evaluate evidence by the same criteria that researchers use, with consideration to reliability, validity, and the appropriate standard of evidence. But policy makers also need to consider contextual factors, such as the degree of uncertainty in findings across multiple studies and the relevant policy alternatives.

A good place for policy makers to start is a careful consideration of the policy goal and what it implies for the standard of evidence they should adopt. For instance, if the goal is to inform individuals about the decisions they face, the standard of evidence may not need to be terribly high. In an apt analogy, Kane (2013) notes that a person on the way to one of two hospitals for treatment for a heart attack may well care whether the mortality rate for heart attack patients is 75 percent at one hospital versus 20 percent at the other—even if the differences between the two hospitals are not statistically significant. Similarly, information about student test results is meant to describe and contextualize a student's performance. It could contribute one piece of data among many that might inform parents' decisions around, say, placing their child in tutoring services. This type of use doesn't require much certainty in the test scores. One would

8. See Cowan and Goldhaber (2016) for evidence from Washington State and a review.

want to be much more certain, however, if those test results are the only factor being used to make those decisions.

This same principle applies to decisions about institutions. Returning again to the teacher preparation example: If the policy objective were to close low-performing programs solely on the basis of value-added measures (a policy, to be clear, that we are not recommending), then policy makers might wish to seek very high levels of certainty that a program is underperforming before taking such a drastic action. By contrast, if the goal were to identify high-performing programs to study more closely for potential effective practices to share with others, or to identify lower-performing programs that might deserve a bit more scrutiny or review, then a lower bar for identifying outliers might be more than sufficient.

Another contextual consideration is the weight of the evidence on an issue. Part of what adds uncertainty to a policy decision is how confident policy makers can be in the likely impact of a policy, based on prior research. But the research literature often does not consistently point in the same direction regarding the likely impact of a policy, and all evidence is contextually specific—generated from a particular group of students, assigned to teachers with particular qualifications, in a particular type of school and district, in a particular time period and policy environment. To decrease uncertainty in a policy outcome, policy makers must weigh these factors to determine which findings have greatest relevance for their needs. For example, much of the national research on charter schools suggests that charters, on average, have impacts on student outcomes that are fairly similar to those of traditional public schools (Betts and Tang 2011; CREDO 2013). In Massachusetts, however, the impact of charters in urban areas appears to be substantially larger, ranging from 0.2 to 0.4 standard deviations per year depending on subject and grade level (Abdulkadiroğlu et al. 2011). Thus, if policy makers wish to be more certain of a positive impact from introducing charters, they might consider how well their context matches what makes urban charters successful in Massachusetts: a strong state authorizing an accountability policy, particular approaches to pedagogy and school climate, and so forth.[9]

The cost and reversibility of policy choices also matters. Choices are inherently riskier when they are harder to reverse, whether because of the level of investment, political considerations, or both. Class size reduction, for example, is a risky investment from the point of view of likely impact on student achievement, as most recent studies show little to no effect (Hoxby 2000; Rivkin, Hanushek, and Kain 2005; Cho, Glewwe, and Whitler 2012; Bosworth 2014; Schwartz, Zabel, and Leardo 2017).[10] Further, it is expensive relative to the likely gain, and it can create unanticipated negative impacts on average teacher quality as districts must dig deeper into their hiring pools to employ sufficient teachers (Schrag 2006; Gilraine 2017). But it is also a policy that, once implemented, is extremely hard to reverse, as it creates difficult conversations in

9. A related but subtler point is that an intervention may have positive effects across all contexts but be more successful relative to some baselines than others. Confidence intervals are rarely reported in a way that quantifies the variation across treatment effects, which may cause policy makers to underestimate the true riskiness of an intervention.

10. Note, however, that although class size reduction appears to have limited effects on student test scores, some evidence suggests smaller classes may positively affect later life outcomes, such as college attendance (Chetty et al. 2011).

schools when parents see the number of chairs in their child's classroom increasing and worry about whether their child is receiving sufficient individual attention. For all these reasons, policy makers should be more cautious when considering a class size reduction policy than another option that represents a smaller investment or is otherwise easier to reverse.

Arguably the most important contextual feature is the relevant policy alternative. Again, consider the issue of rating teacher preparation programs. The AERA (2015) statement about using value added to evaluate or rate programs notes, "There are promising alternatives currently in use in the United States that merit attention . . . [such as] teacher observation data, peer assistance and review models" (p. 451). These methods may well have promise for characterizing the quality of teacher preparation programs, but they also inherently involve uncertainty. The uncertainty inherent in other forms of program evaluation may not be quantifiable but that does not mean it does not exist.

The policy alternative that may be most frequently overlooked is sticking with the status quo. When the status quo is the alternative, policy makers should be particularly cautious about making changes to a successful status quo policy or program, and they should tolerate a bit more risk when the status quo is likely to be yielding poor results. Teacher compensation is an instance where the status quo has powerful inertia but perhaps should not. The overwhelming majority of teachers are paid according to a single salary schedule that rewards years of experience and, generally, having a master's degree (USDOE 2012). Presumably, a goal of this policy is to pay more effective teachers more than less effective teachers, since they contribute more to student improvement. But although research finds that teachers rapidly improve as they gain experience early in their careers (e.g., Rockoff 2004), strikingly little evidence supports the notion that attaining a master's degree has an impact on teacher effectiveness—or even that teachers with master's degrees tend to be more effective.[11] Policy makers wishing to compensate for teacher effectiveness—at least, as measured by impact on student test scores—should therefore be more cautious about tinkering with changes to rewards associated with teacher experience than they are about changing the master's premium. Despite this, however, most school systems in the country still pay teachers with master's degrees more than those with bachelor's degrees.[12]

This highlights a final point: Because the purpose of education is to improve outcomes for students, policy makers should make judgments about benefits, costs, and uncertainty from a student perspective. But too often the focus is on the risks of a change to the adults in the system rather than the risk of the status quo on students. This can cause inertia and ultimately may harm the students the education system is intended to serve.

11. Note that this is the finding for master's degrees in general (Ladd and Sorenson 2015; Goldhaber 2016). Evidence does suggest that holding master's degree in a subject, particularly math and science, predicts teacher effectiveness in that subject area (Goldhaber and Brewer 1997; Coenen et al. 2017; Bastian 2019).
12. Why is the master's pay premium sticky despite the empirical evidence that it is not well aligned with teacher effectiveness? We believe one reason is that the risks involved are typically framed around the adults in the system rather than the students that the school system is supposed to serve. It is pretty certain that paying teachers more for master's degrees will not enhance student learning. But from an adult perspective, what might replace the master's premium, and therefore how one might earn future salary raises, is highly uncertain.

CONCLUSION

All policy decisions require policy makers to make a bet on the future with the information available today. Ignoring the nuances inherent in how information from research will be used, and thereby holding all purposes to an equivalent, arbitrary standard of statistical significance, does a disservice to both the research and policy-making communities. It renders many research findings irrelevant for policy because too little information was provided about their context. And it may cause policy makers to err on the side of inaction or to make uninformed bets.

How can researchers help? For starters, by not perpetuating the problem. For example, they can report confidence intervals rather than up or down interpretations of *p*-values (Amrhein, Greenland, and McShane 2019). They can show how their findings might differ under different standards of evidence, as we demonstrated here with the teacher preparation program example. They can also provide more information about the reliability and validity of their findings and the context in which they were produced. It is particularly important they do so in the more accessible versions of their work likely to be seen directly by policy makers: abstracts, summaries, and policy briefs.

Likewise, policy makers should recognize the statistician's standard for statistical significance often results in a message of "this intervention works" or "it doesn't work"—but this message is oversimplified. Nothing about the 95 percent confidence standard is special, and neither they nor researchers should blindly adhere to it. Rather, both should carefully consider the context in which decisions are made and the policy alternative for the decision, as well as how both factors influence the level of confidence they need for making policy choices. Sometimes context will call for making decisions that research suggests will lead to (precisely estimated) marginal improvements, but other times it will be appropriate to go with the (underpowered) moonshot. Thinking clearly about the full range of options and the standard of evidence each requires is central to good policy making.

ACKNOWLEDGEMENTS

We are grateful to James Cowan, Bob Lee, Roddy Theobald, Katharine Strunk, and two anonymous referees for helpful comments on earlier drafts. Note that the views expressed are those of the authors and do not necessarily reflect the views of the institutions with which the authors are affiliated.

REFERENCES

Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak. 2011. Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *Quarterly Journal of Economics* 126(2): 699–748.

American Educational Research Association (AERA). 2015. *AERA issues statement on the use of value-added models in evaluation of educators and educator preparation programs*. Available https://www.aera.net/Newsroom/News-Releases-and-Statements/AERA-Issues-Statement-on-the-Use-of-Value-Added-Models-in-Evaluation-of-Educators-and-Educator-Preparation-Programs. Accessed 10 September 2019.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for educational and psychological testing*. Washington, DC: AERA.

Amrhein, Valentin, Sander Greenland, and Blake Mcshane. 2019. Scientists rise up against statistical significance. *Nature* 567(7748): 305–307. doi: 10.1038/d41586-019-00857-9.

Arrow, Kenneth J., and Robert C. Lind. 1978. Uncertainty and the evaluation of public investment decisions. In *Uncertainty in economics*, edited by Peter Diamond and Michael Rothschild, pp. 403–421. New York: Academic Press.

Bastian, Kevin C. 2019. A degree above? The value-added estimates and evaluation ratings of teachers with a graduate degree. *Education Finance and Policy* 14(4): 652–678. doi: 10.1162/edfp_a_00261.

Betts, Julian R., and Y. Emily Tang. 2011. The effect of charter schools on student achievement: A meta-analysis of the literature. Seattle, WA: National Charter School Research Project, Center on Reinventing Public Education.

Bosworth, Ryan. 2014. Class size, class composition, and the distribution of student achievement. *Education Economics* 22(2): 141–165.

Bradley, Richard, and Mareile Drechsler. 2013. Types of uncertainty. *Erkenntnis* 79(6): 1225–1248.

Center for Research on Education Outcomes (CREDO). 2013. *National charter school study*. Palo Alto: CREDO, Stanford University.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmannuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics* 126(4): 1593–1660.

Cho, Hyunkuk, Paul Glewwe, and Melissa Whitler. 2012. Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review* 31(3): 77–95.

Coenen, Johan, Ilja Cornelisz, Wim Groot, Henriette Maassen van denBrink, and Chris Van Klaveren. 2017. Teacher characteristics and their effects on student test scores: A systematic review. *Journal of Economic Surveys* 32(3): 848–877. doi: 10.1111/joes.12210.

Cowan, Joshua, and Dan Goldhaber. 2016. National Board certification and teacher effectiveness: Evidence from Washington State. *Journal of Research on Educational Effectiveness* 9(3): 233–258.

Gilraine, Michael. 2017. Multiple treatments from a single discontinuity: An application to class size. Unpublished paper, University of Toronto.

Goldhaber, Dan. 2006. National Board teachers are more effective, but are they in the classrooms where they're needed the most? *Education Finance and Policy* 1(3): 372–382.

Goldhaber, Dan. 2016. In schools, teacher quality matters most: Today's research reinforces Coleman's findings. *Education Next* 16(2): 56–62.

Goldhaber, Dan. 2019. Evidence-based teacher preparation: Policy context and what we know. *Journal of Teacher Education* 70(2): 90–101.

Goldhaber, Dan, and Dominic Brewer. 1997. Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources* 32(3): 505–523. doi: 10.2307/146181

Goldhaber, Dan, Stephanie Liddle, and Roddy Theobald. 2013. The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review* 34:29–44.

Goldhaber, Dan, and Umut Özek. 2019. How much should we rely on student test achievement as a measure of success? *Educational Researcher* 48(7): 479–483. doi: 10.3102/0013189X19874061

Goodwin, Paul, and George Wright. 2014. *Decision analysis for management judgment*, fifth edition. London: John Wiley & Sons.

Holden, Kristian L. 2016. Buy the book? Evidence on the effect of textbook funding on school-level achievement. *American Economic Journal: Applied Economics* 8(4): 100–127.

Hoxby, Caroline M. 2000. The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics* 115(4): 1239–1285.

Kane, Thomas J. 2013. *Presumed averageness: The mis-application of classical hypothesis testing in education*. Available https://www.brookings.edu/research/presumed-averageness-the-mis-application-of-classical-hypothesis-testing-in-education/. Accessed 10 September 2019.

Kane, Thomas J., and Douglas O. Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16(4): 91–114.

Koretz, Daniel. 2009. *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Koretz, Daniel. 2017. *The testing charade: Pretending to make schools better*. Chicago: University of Chicago Press.

Ladd, Helen F., and Lucy C. Sorensen. 2015. Do Master's degrees matter? Advanced degrees, career paths, and the effectiveness of teachers. CALDER Working Paper No. 136, American Institutes for Research.

Penuel, William R., Derek C. Briggs, Kristen L. Davidson, Corinne Herlihy, David Sherer, Heather C. Hill, Caitlin Farrell, and Anna Ruth Allen. 2017. How school and district leaders access, perceive, and use research. *AERA Open* 3(2): 1–17. doi: 10.1177/2332858417705370.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73(2): 417–458.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94(2): 247–252.

Rockoff, Jonah, and Lesley J. Turner. 2010. Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy* 2(4): 119–147.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy* 5(2): 251–281.

Schrag, Peter. 2006. Policy from the hip: Class-size reduction in California. *Brookings Papers on Education Policy* 9(2006/2007): 229–243.

Schwartz, Amy Ellen, Jeffrey Zabel, and Michele Leardo. 2017. Class size and resource allocation. ESE Policy Brief. Malden, MA: Massachusetts Department of Elementary and Secondary Education.

Shober, Arnold F. 2016. Individuality or community? Bringing assessment and accountability to K–16 education. In *The convergence of K–12 and higher education: Policies and programs in a changing era*, edited by Christopher P. Loss and Patrick J. McGuinn, pp. 67–86. Cambridge, MA: Harvard Education Press.

United States Department of Education (USDOE). 2012. *Schools and staffing survey.* Washington, DC: National Center for Education Statistics.

von Hippel, Paul T., and Laura Bellows. 2018. How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review* 64:298–312.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA's statement on p-values: Context, process, and purpose. *American Statistician* 70(2): 129–133.