CALDER Policymakers Council

Opinion Brief

APPROPRIATE STANDARDS OF EVIDENCE FOR EDUCATION POLICY DECISION-MAKING

Carrie Conaway
Massachusetts Department of Elementary and Secondary Education

Dan Goldhaber
University of Washington
American Institutes for Research/CALDER

Suggested citation:

Carrie, C. and Goldhaber, D. (2018). *Appropriate standards of evidence for education policy decision-making* (CALDER Opinion Brief). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. CALDER Policy Brief No. 4-0918-2.

The crafting and dissemination of this opinion brief was supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder.**WARNING: this brief contains the author's unmoderated opinions about controversial issues, which may cause dizziness, nausea, and/or seizures.** Note that the views expressed are those of the authors and do not necessarily reflect those of funders or the institutions to which the authors are affiliated. The authors are grateful to James Cowan, Bob Lee, Roddy Theobald, and two anonymous referees for helpful comments on earlier drafts.

1. Introduction

A key job of education policymakers 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 policymakers 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 policymakers 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 policymakers *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 the ways in which uncertainty affects decisions.

In this policy brief we review the way uncertainty factors into different types of decisions and illustrate how the standard of evidence for making decisions can be quite inconsistently applied. The type of information available about uncertainty, the nature of the decisions to be made, and the broader context for the decision are all critical factors—and all ones that are commonly overlooked in the way research findings are reported, making it all the harder for policymakers to appropriately consider uncertainty in their decisions. We also argue that inconsistency in evaluating the probabilities of risks and rewards can lead to suboptimal decisions for students, in part because risks and rewards are often judged by how the adults in the system are affected more than how students are affected. Finally, we offer some suggestions for how policymakers might think about the level of confidence they need to make different types of decisions.

1. The Use and Non-Use of Statistical Significance in Policymaking

We are not alone in raising concerns about how policymakers 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 & 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 that it is important to center this challenge within a context recognizable to policymakers, to make more specific how these issues apply to the decisions they face. In this case we focus on education policy, showing how uncertainty is handled in different situations, sometimes in ways that are seemingly inconsistent.

Testing and school accountability provide useful illustrations of how uncertainty enters the policymaking process. States must administer annual academic achievement tests to students in most grades and use the resulting data to make determinations of school quality. These data and determinations are inherently uncertain, yet uncertainty is inconsistently considered throughout the process—and in fact is considered least in the areas where it may be most consequential.

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 in which statisticians are highly confident (more on this below) that a student would receive a similar score were he/she to retake the test (American Educational Research Association, 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. But criterion-referenced tests 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 impossible 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 school level to describe school performance. Every state 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 *No Child Left Behind* and its successor, the *Every Student Succeeds Act*, 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. 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. Yet 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 rank schools according to test scores, graduation rates, and other criteria and classify them based on the rankings, then 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.

[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.

Small differences between schools in test scores are almost surely not indicative of true underlying differences in school quality (Kane & 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., Rouse, Hannaway, Goldhaber, & Figlio, 2013; Holden, 2016; Rockoff & Turner, 2010). 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' accountability system under the No Child Left Behind waiver (in place from 2012 to 2016), schools were counted as reaching their performance target if they came within one standard deviation of it. 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.

Thus, uncertainty is an integral part of generating, interpreting, and using assessment data, yet 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 policymaking. The inconsistency is troubling considering that 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, i.e., precisely the cases where uncertainty is not considered. Researchers and policymakers alike need to think more systematically about how they should consider uncertainty in the decision and policymaking processes.

2. Toward Appropriate Research Framing and Standards of Evidence

Decades of academic research speaks to how managers make decisions under uncertainty (e.g. Arrow & Lind, 1978; Bradley & Drechsler, 2013; Goodwin & Wright, 2014). But little of this work addresses how the ways research findings are reported affects managers' understanding of

or accounting for uncertainty, or how the context for a decision might influence how much certainty a decision-maker should seek.

3a: Standard Statistical Practice Often Doesn't Reflect Policymakers' Needs

In academia, individual research 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 what it intended to measure, and generalizes to other contexts. Increases in these measures mean decreases in uncertainty about what a study implies for decision-making. But this overlooks the fact that in the abstracts, briefs, and media reporting that are most accessible to policymakers, findings are generally described not in terms of their reliability and validity, but their statistical significance—and the statistician's standard for what constitutes a "significant finding"—can often steer policymakers 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, if the treatment were repeated multiple times, 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.[2] Yet, some policymaker decisions suggest that 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.

[2] 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 (aka Type II error). This is particularly true in research 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.

Teacher preparation policy provides a helpful example of how policymakers might weight the risks of false positives and negatives differently than standard statistical practice. The quality of newly prepared teachers and the role of teacher preparation programs (TPPs) in developing teachers are issues receiving increased attention of late (Goldhaber, forthcoming). One natural question is whether TPPs vary meaningfully in the effectiveness of their graduates. In fact, a number of states have begun to hold TPPs accountable for teacher value added,[3] one measure of teacher effectiveness (von Hippel & Bellows, 2018).

Not surprisingly, ranking TPPs 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 TPPs (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, individual TPPs graduate teachers that differ from the average in effectiveness? (That is, which can we be reasonably certain credential especially strong or weak teachers?) The answer depend 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 TPPs are different from the average TPP in each state and conclude that "It is not meaningful to rank all the TPPs in a state. The true differences between most TPPs 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 below, which is based on analysis of TPPs in Washington state (Goldhaber, Liddle, & Theobald, 2013), illustrates this point.

[3] For more on value-added and other measures of teacher performance, see http://www.carnegieknowledgenetwork.org/briefs/value-added/value-added-other-measures/.

Figure 1 shows the estimated math value added of teachers from the 20 programs in Washington State.[4] The 95 percent confidence intervals (the thin line) 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, here defined as the average effectiveness of teachers who transfer in from out of state; when this happens, the program produces graduates that are not statistically distinguishable from teachers imported from outside Washington. In fact, no programs are, by this metric, significantly different from one other and only one is different from zero, i.e., the average out-of-state prepared teacher receiving a credential.[5]

But what if the standard were 80 percent confidence instead? Then 12 programs are different from one another, and six are different from the impact of the average out-of-state transfer teacher. What level of confidence is the right one for policymakers 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 TEPs 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 that the decision-maker is comparing discrete potential strategies and can make a decision by comparing the probability of the outcomes from each. But in reality, it is often the case that this type of information is not actually available in a way that meets decision-makers' needs.

For example, several well-executed studies now show that National Board for Professional Teaching Standards (NBPTS) teachers are more effective than those who are not.[6] This headline

[4] See Goldhaber et al. (2013) for details about the estimates (the estimates reported in Figure 1 are derived from the coefficients in Column 1 of Table 4).
[5] 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.
[6] See Cowan and Goldhaber (2016) for evidence from Washington state and a review.

emphasizes the statistical significance of these findings. But if a policymaker 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 that the statistically significant finding for NBPTS teachers means that policymakers have about a 55 to 60 percent chance of rewarding more effective teachers (Goldhaber, 2006). Whether that rate is high or low is a value judgment (which is what we have policymakers for!), but the framing around probabilities seems more in line with how this question might be debated in policy terms than whether a finding is statistical significant.

3b: The Policymaking Context Is Critical, Yet Frequently Overlooked

So, what standard of evidence *should* policymakers use when making policy decisions? Looking at individual studies, of course, policymakers should evaluate evidence by the same criteria that researchers use, with consideration to reliability, validity, and the appropriate standard of evidence. But policymakers 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 policymakers to start with 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, Tom Kane (2013) notes that a person on the way to one of two hospitals for treatment for a heart attack may well care (we sure would!) 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 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 TPPs solely on the basis of value-added measures (a policy, to be clear, that we are not recommending), then policymakers 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 relevance of prior evidence for the situation at hand. Part of what adds uncertainty to a policy decision is how confident policymakers 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, policymakers must weight 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 & Tang, 2011; Center for Research on Education Outcomes, 2013). In Massachusetts, however, the impact of charters in urban areas appears to be significantly larger, ranging from 0.2 to 0.4 standard deviations per year depending on subject and grade level (Abdulkadiroğlu et al, 2011). Thus, if policymakers 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 and accountability policy, particular approaches to pedagogy and school climate, and so forth.[7]

[7] 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 policymakers to underestimate the true riskiness of an intervention.

Policy choices are also 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; Bosworth, 2014; Rivkin, Hanushek, & Kain, 2005; Cho, Glewwe, & Whitler, 2012; Schwartz, Zabel, & Leardo 2017).[8] 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 hire 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 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, policymakers 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 policy consideration about uncertainty, and surely the most overlooked, is *the relevant policy alternative*. Again consider the issue of rating TPPs. The AERA (2015) statement about using value added to evaluate or rate TPPs 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. These forms of uncertainty are just harder to quantify and thus more easily overlooked. Policymakers should therefore recognize that the issue of uncertainty in TPP ratings (or ratings of any sort) is not limited to the uncertainty inherent in value-added models.

Ironically, the policy alternative that may be most frequently overlooked is sticking with the status quo. And when the status quo is the alternative, policymakers should be particularly cautious about making changes to a successful status quo policy or program, and they should

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

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. 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 while research finds that teachers rapidly improve as they gain experience early in their careers, there is strikingly little evidence supporting 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). Policymakers wishing to compensate for teacher performance 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.

Why is the master's pay premium sticky despite the empirical evidence that it does not seem to be well aligned with teacher effectiveness? One obvious 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. From the student perspective, it is highly certain that paying teachers more for master's degrees will not enhance their 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. Policymakers should remember that the consequences of policy changes are not relative to a preexisting nirvana, but relative to current conditions—which may or not be ideal for achieving a particular policy goal.

This highlights a final point: Because the purpose of education is to improve outcomes for students, policymakers 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 and pay insufficient attention to 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.

3. Concluding Thoughts

All policy decisions require policymakers 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 policymaking communities. It renders many research findings irrelevant for policy because too little information was provided about their context. And it may cause policymakers to err on the side of inaction, and/or to make relatively uniformed bets.

The bottom line: It is too easy to fall into the trap of always using the statistician's standard 95 percent confidence threshold . But nothing about this standard is special, and neither policymakers nor researchers should blindly adhere to it. Rather, they 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 associated policy-relevant confidence intervals is central to good policymaking.

References

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *The Quarterly Journal of Economics*, *126*(2), 699-748.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing.* Washington, DC. American Educational Research Association.

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1-27.

Arrow, K. J., & Lind, R. C. (1978). Uncertainty and the evaluation of public investment decisions. In *Uncertainty in Economics* (pp. 403-421).

Betts, J. R., & Tang, Y. E. (2011). The Effect of Charter Schools on Student Achievement: A Meta-Analysis of the Literature. *Center on Reinventing Public Education*.

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

Bradley, R., & Drechsler, M. (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, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593-1660.

Cowan, J. and Goldhaber, D. (2016). National Board Certification and Teacher Effectiveness: Evidence from Washington State. *Journal of Research on Educational Effectiveness*. 9(3): 233-258.

Gilraine, M. (2017). Multiple treatments from a single discontinuity: An application to class size. University of Toronto.

Goldhaber, D. (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, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29-44.

Goldhaber, D. (forthcoming). Evidence-Based Teacher Preparation: Policy Context and What We Know. *Journal of Teacher Education.*

Goodwin, P., & Wright, G. (2014). *Decision Analysis for Management Judgment 5th ed* (No. 5th). John Wiley and sons: London, England.

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

Kane, T. (2016) Never judge a book by its cover—use student achievement. Brookings. Series: Evidence Speaks; Report. Retrieved from https://www.brookings.edu/research/never-judge-a-book-by-its-cover-use-student-achievement-instead/.

Holden, K.L. (2016). Buy the Book? Evidence on the Effect of Textbook Funding on School-Level Achievement. *American Economic Journal: Applied Economics*, *8*(4), pp.100-127.

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.

Kane, T.J. (2013). Presumed Averageness: The Mis-Application of Classical Hypothesis Testing in Eduction. The Brown Center Chalkboard Series Archieve (Wednesday, December 4). Brookings Institution.

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2), 417–458.

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

Rouse, Celia Elena, Hannaway, Jane, Goldhaber, Dan and Figlio, David. (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, P. (2006). Policy from the Hip: Class-Size Reduction in California. Brookings Papers on Education Policy, (9), 229–243.

Schwartz, A., Zabel, J., & Leardo, M. (2017). "Class Size and Resource Allocation." Massachusetts Department of Elementary and Secondary Education. Retrieved from http://www.doe.mass.edu/research/reports/2017/12class-size.docx.

Shober, Arnold F. 2016. "Individuality or Community? Bringing Assessment and Accountability to K-16 Education." In Christopher P. Loss and Patrick J. McGuinn, eds., The Convergence of K-12 and Higher Education. Cambridge: Harvard Education Press, 67-86.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child development*, *85*(3), 842-860.

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review*.

Wasserstein, R. L. and Nicole A. L. (2016). "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician*, 70:2 (129–133).

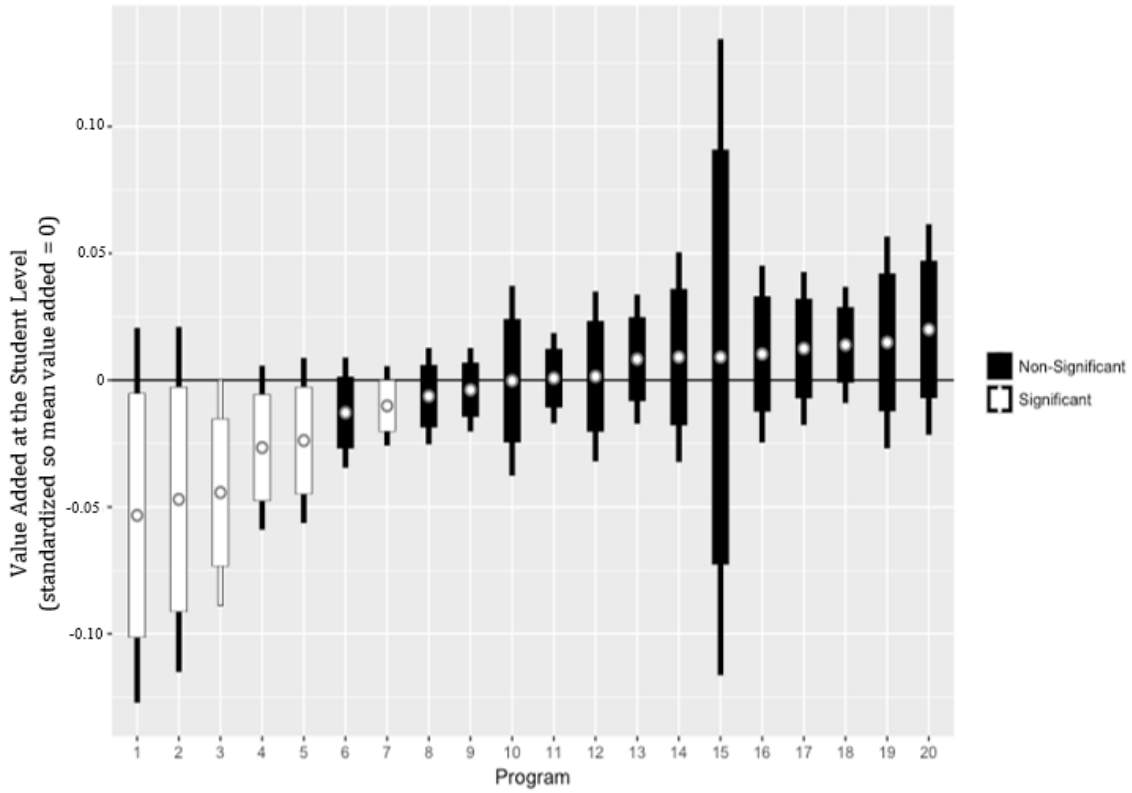

Figure 1: Estimated Mathematics Value Added by Teacher Preparation Programs, Washington State