Acknowledgements

NEPC Staff

Kevin Welner
Project Director

William Mathis
Managing Director

Alex Molnar
Publications Director


Funding: This review was made possible in part by funding from the Great Lakes Center for Educational Research and Practice.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

This publication is provided free of cost to NEPC’s readers, who may make non-commercial use of it as long as NEPC and its author(s) are credited as the source. For inquiries about commercial use, please contact NEPC at nepc@colorado.edu.
Executive Summary

A new report from the National Council on Teacher Quality (NCTQ) highlights six teacher evaluation systems claimed to be “yielding substantial benefits.” This comes at the end of a decade when reformed teacher evaluation systems that link teacher performance to measures of student growth have been at the center of educational debate. Disagreements range from the theoretical (e.g., is teacher quality fundamentally related to inequalities in student outcomes?) to technical (e.g., which measures should be included and how should they be defined?) to practical (e.g., how should ratings be used for personnel decisions?). Overall, the research regarding teacher evaluation is mixed, at best. Most notably, a recent multi-year RAND report suggests that a $500 million investment in teacher evaluation that heavily weighted student growth measures, largely funded by the Gates Foundation, did not improve student outcomes and, in some cases, exacerbated unequal access to effective teachers for low-income students and students of color. The new NCTQ report, while clearly promoting such teacher evaluation, does not seriously counter the groundswell of academic literature critiquing these systems. It does not address the relevant literature, present a compelling justification for its site selection or the inclusion criteria for evidence, or adequately consider disconfirming or contradictory evidence. These methodological flaws limit the validity of the report’s findings and conclusions, which ultimately diminishes the usefulness of the report for policy and practice.
I. Introduction

Teacher evaluation has been the topic of attention and debate over the past decade. Former evaluation systems were criticized for their inability to differentiate among teachers and for doing little to inform instructional improvement or personnel decisions. Inspired by federal incentives like Race to the Top grant funding and waivers from No Child Left Behind requirements, evaluation reform has become nearly ubiquitous. Nearly all states and many districts have adopted new or revised teacher evaluation systems, with a focus on student growth measures tied to individual teachers. This rapid adoption has been met with a nearly equal growth in related scholarship.

Numerous converging influences have provoked a backlash to this quick escalation of teacher evaluation reform. First, the 2015 Every Student Succeeds Act relaxed federal control over teacher evaluation. At the same time, teachers voiced persistent reservations about revised evaluation systems in general, and student growth measures in particular. These concerns, as well as the surge of lawsuits challenging evaluation systems across the country, have driven districts to rethink or revise their adopted models. Finally, it has become increasingly difficult to justify the continued use of measures that attribute student outcomes to individual educators amid warnings from researchers and professional organizations regarding their defensibility and reliability.

Another considerable blow to teacher evaluation reform may have been the June 2018 RAND evaluation report, which concluded that Gates’ investment in teacher evaluation systems had proven almost entirely unsuccessful. The RAND report concluded that student outcomes did not improve and low-income and minority students had no greater access to
In October 2018, a new report from the National Council on Teacher Quality (NCTQ) was released. *Making a difference: Six places where teacher evaluation systems are getting results*, written by Hannah Putman, Kate Walsh, and Elizabeth Ross, highlights six places that have “surmounted challenges to implement successful teacher evaluation systems that are yielding substantial benefits.”¹¹ The report concludes that, unlike the sites studied in the RAND report, these six effective evaluation systems have generated improvements in the overall quality of the teacher workforce.

In this review, we summarize the findings and conclusions of the NCTQ report, with special attention to its use of literature and the justifications given for the conclusions it draws. We argue that the report does not sufficiently address the evidence surrounding teacher evaluation and does not adequately consider disconfirming or contradictory evidence. These methodological flaws limit the validity of the report’s findings and conclusions, reducing the usefulness of the report for policy and practice.

## II. Findings and Conclusions of the Report

The NCTQ report analyzes six teacher evaluation systems reporting positive results, including four districts (Dallas, Denver, District of Columbia, and Newark) and two states (Tennessee and New Mexico).¹²

The report stresses the importance of evaluation systems that differentiate between varying levels of teacher effectiveness and use that information for personnel decisions. It lauds the six systems for continued progress in several key areas, including an increase in the frequency of evaluation and the number of rating levels used to evaluate teachers; the use of multiple measures, including “objective measures of student growth,”¹³ to determine teacher ratings; and the use of evaluation results in setting compensation. The report also highlights what these six systems have in common, organized by four principles of strong practice: (a) adherence to core principles, (b) a thoughtful approach to weighting individual components, (c) linking evaluation to personnel decisions, and (d) linking evaluation to compensation.

### Adherence to core principles

NCTQ praises all six systems for following strong practices. Most notably, they all maintain the use of multiple measures, including student surveys and student growth measures. The
systems also used at least three teacher-rating categories, utilized annual evaluations and observations, and linked professional development to evaluation results.

**Thoughtful weighting of individual components**

Components of the various evaluation systems included observations, student achievement and/or growth, student surveys, professionalism, commitment to the school community, and school attendance. All systems at a minimum included observations and measures of student achievement/growth. The weight given to student achievement/growth ranged from 15-50%. The weight for student surveys varied from 0-15%. Two of the profiled systems also included a measure of professionalism, and one considered commitment to the school community and one included teacher attendance as part of the overall evaluation rating.

**Linking evaluation to personnel decisions**

All six systems tailored professional development in response to evaluation results. Some sites also used evaluations to select mentors for student teachers and teachers for leadership opportunities. Others based teacher dismissal decisions on evaluations or provided financial incentives for highly rated teachers who work in high-need schools.

**Linking evaluation to compensation decisions**

A high evaluation rating leads to a raise, bonus, or award in all six evaluation systems. Five of the six systems also compensate teachers for teaching a hard-to-staff subject or teaching in a high-need school.

**Report conclusions**

Although the report itself does not clearly delineate a set of conclusions, the NCTQ press release is more forthcoming. It asserts that the six evaluation systems have produced a meaningful and realistic measure of teacher talent. Thus, districts have been able to retain strong teachers and increase the rate at which weaker teachers choose to leave. The press release highlights increases in student proficiency that coincide with implementation of the new evaluation systems, suggesting that those reformed systems are responsible for the improved outcomes. It also suggests that the most significant component of each of the six systems is the link between teacher evaluation results and compensation, a suggestion that does not appear plainly in the full report.

**III. The Report’s Rationale for Its Findings and Conclusions**

The report does not offer an explicit rationale for its conclusions, which are presented pri-
III. The Report’s Use of Research Literature

The majority of references are drawn from the six teacher evaluation systems, and include policy documents, personal communication with stakeholders, and internal evaluations. Beyond those, the report cites news outlets, including *Education Week* and the *Washington Post*; publications from advocacy-oriented organizations, like TNTP and the Association for Supervision and Curriculum Development (ASCD); and research centers like the Center for Economic and Policy Research (CEPR) at Harvard University and the RAND Corporation. Seven endnotes refer to other NCTQ documents or projects.

The report does not sufficiently utilize academic or peer-reviewed work. Despite a robust literature base on teacher evaluation, only nine of the 127 endnotes in the report include unique citations from peer-reviewed journals. 19 Although the report seems to draw on an unstated conceptual framework that prioritizes evaluation, high-stakes testing, accountability, and pay-for-performance, it almost completely ignores the extensive literature regarding each reform, which presents mixed evidence at best. The report inadequately discusses the empirical evidence—or lack thereof—regarding the use of these reforms, the associated effect on teachers or the workforce, or their ultimate impact on student outcomes. In particular, it fails to address the body of work related to the challenges of attributing student growth to teachers, particularly more recent research that presents a less positive picture of evaluation reforms. 20 It also does not present any meaningful discussion of the role of evaluation on teacher development and improvement.

V. Review of the Report’s Methods

The NCTQ report does not discuss its methodology. This results in two major uncertainties. First, the report does not detail the selection criteria used for the inclusion of districts and states, simply noting that it was able to “identify six pioneers ... [whose] results are setting them apart.” 21 There is no discussion of how these systems were selected, what other locales were considered, or how the evidence base contributed to those decisions. In the absence of that information, one is left to wonder if NCTQ hand-selected locations based on instincts, anecdotes, or external pressures. Several of these locales have personal connections to NCTQ 22 and five of the six have received significant financial support from the Bill and Melinda Gates Foundation, 23 as has NCTQ. 24 The lack of information about the selection of these locales significantly hampers the ability to interpret how these systems’ results “set them apart” from anyplace else.

The second major methodological shortcoming is the approach used to synthesize data about the teacher evaluation systems in each district and state. NCTQ seems to imply that its
findings come from its own empirical investigation. A closer review suggests that the data presented do not in fact stem from original analyses, but are instead synthesized from “official policies, reviews of the evaluation system that these states and districts have conducted, independent studies, and interviews with district and state staff.” However, high-quality research syntheses should seek to “integrate empirical research for the purpose of generalizations.” When done well, research syntheses are complex approaches requiring numerous methodological decisions (e.g., how to select relevant studies, code their findings, and analyze their results to synthesize the most salient themes).

Nowhere does the report describe the approach it took for each of these methodological decisions, although we were able to recreate an inventory of the studies and data sources utilized, presented in Table I.

Table 1: Data Sources Used in NCTQ Analysis

<table>
<thead>
<tr>
<th>Locale</th>
<th>Interview/ personal communication</th>
<th>Internal research or other internal documents</th>
<th>External source (not peer-reviewed)</th>
<th>External source (peer-reviewed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dallas Independent School District (DISD)</td>
<td>Suzy Smith, Manager of Performance Management; John Vega, Deputy Chief, Human Capital Management</td>
<td>2015-16 survey results report; 2016-17 TEI evaluation and guidebook; 2015-16 ACE evaluation</td>
<td><strong>Commit!</strong> State achievement data; <strong>Dallas News</strong> article re: state school accountability ratings</td>
<td>None</td>
</tr>
<tr>
<td>Denver Public Schools (DPS)</td>
<td>Nicole Wolden, Director of Growth and Performance</td>
<td>* ** press release re: student growth in 2018</td>
<td>** CADRE Center evaluation of Denver ProComp; * ERS presentation to DPS leadership</td>
<td>None</td>
</tr>
<tr>
<td>District of Columbia Public Schools (DCPS)</td>
<td>Betsy Press, Former Deputy Chief of IMPACT; Christopher Lewis, Deputy Chief of IMPACT</td>
<td>* 2010-11 and 2017-18 IMPACT system overview; * 2017-18 LIFT guidebook; * ** Press release re: NAEP results; * ** FY 2018 salary schedule</td>
<td>FutureEd analysis of DCPS teacher reforms; Bellwether Education’s analysis of DCPS exit survey; Mathematica analysis of DCPS effectiveness under IMPACT; * Washington Post article re: IMPACT scores; ** Nation’s Report Card 2009 vs 2017 NAEP results</td>
<td>Adnot et al. (2017) article re: teacher turnover under IMPACT; Dee et al. (2015) article re: teacher retention/performance under IMPACT</td>
</tr>
<tr>
<td>Newark Public Schools (NPS)</td>
<td>Larisa Shambaugh, Former Chief Talent Officer</td>
<td>2016 state-operated school district annual report to the NJ State Board of Education</td>
<td>AIR 2016 external evaluation: NPS teacher contract; CEPR 2017 study: NPS reform effect on achievement; MarGrady 2017 analysis of NPS achievement trends</td>
<td>None</td>
</tr>
</tbody>
</table>
This summary makes clear that the majority of district and state profiles relied heavily on interviews with one or two staff members within each agency. In at least one location (Newark Public Schools), NCTQ was not able to collect any data directly from current staff, instead relying exclusively on data from a former staff member, in addition to data drawn from one PowerPoint presentation to the New Jersey Board of Education and three external reports. This summary also elucidates how little evidence was gathered from “independent studies” or external research: With the exception of DC Public Schools and New Mexico, each case study was created without the benefit of any data or evidence drawn from peer-reviewed empirical research.

While our summary of evidence may provide a better understanding of the data included in the report, it does not answer other important questions about NCTQ’s data-gathering approach, such as how thorough or selective this data-gathering was. For example, the reader cannot ascertain whether NCTQ conducted an exhaustive search of all publicly available evidence or if they confined themselves to seeking out evidence related to the report’s key principles. If it is the latter, the evidence and associated conclusions are largely invalid.

### VI. Review of the Validity of the Findings and Conclusions

Below, we briefly evaluate the NCTQ report using Cooper and Hedges’ suggestions for sound research syntheses.

http://nepc.colorado.edu/thinktank/review-teacher-eval
The report pays little attention to relevant theory

Good research practice would entail drawing on the relevant prior literature, including scholarship regarding teacher evaluation, professional improvement, and performance pay. However, as noted above, NCTQ does not adequately address or reference the prior literature.

The report relies on an underdeveloped and incomplete analysis of the relevant research, and fails to address disconfirming evidence

In addition to using a narrow set of sources, there are several examples where the report seems to misinterpret or mischaracterize findings, or leave out data that was less than positive. For example, survey results cited within the report suggest that only 41% of teachers in Dallas Independent School District were satisfied with the system.\(^8\) Although the report praises Dallas because this number represents an increase over prior years’ data, a 41% satisfaction rate simply fails to represent the perceptions of teachers in a well-functioning system. The report does not mention that nearly half the teachers cite concerns about the system’s fairness.\(^8\) To its credit, the report does highlight a concern among Dallas teachers that the system may disadvantage teachers in lower-performing schools, citing this as an opportunity for improvement in the district. However the report mischaracterizes this finding, claiming that “roughly a quarter of teachers felt the evaluation system was unfair to teachers facing greater challenges in low-performing schools.”\(^8\) In fact, 51% of teachers surveyed about this issue selected it as a concern, although the question was asked of only a subset of teachers. In total, these omissions and mischaracterizations suggest that the report does not take a critical or objective lens regarding the evidence base.

The report makes only oblique references to future research

Although the report acknowledges that “the knowledge base for building a strong evaluation system is still young and is continuously being refined,”\(^8\) it does not return to this in its conclusions and fails to identify areas for future research or acknowledge limitations or uncertainties that warrant additional study.

VII. Usefulness of the Report for Guidance of Policy and Practice

Had the report more fully considered the body of evidence surrounding teacher evaluation in each of these locales and presented a more holistic narrative of its successes and challenges, it might have provided a meaningful contribution. Instead, its narrow focus and questionable inclusion criteria make it more misleading than helpful.

Teacher evaluation reform has not yet been sufficiently vetted in the empirical literature
and it has certainly not resulted in a demonstrable effect on student achievement.\textsuperscript{84} The NCTQ report does little to extend that literature. Given the timing of its release immediately following the RAND evaluation, it is worth considering whether the NCTQ publication is a thinly veiled attempt to shift the national discourse away from negative press that characterized teacher evaluation reform as an “expensive experiment,”\textsuperscript{85} “a bust,”\textsuperscript{86} and “one of the biggest failures in K-12 philanthropy.”\textsuperscript{87}

While not the focus of this review, it is noteworthy that the RAND evaluation provides more thoughtful guidance for policy and practice than is noted here. It acknowledges that while it is difficult to determine the exact cause of the lackluster results, they may be attributable to the reform’s narrow focus: in order for transformative change to be possible, RAND points out, “many other factors might need to be addressed, ranging from early childhood education, to students’ social and emotional competencies, to the school learning environment, to family support.”\textsuperscript{88} We agree but also encourage policymakers and practitioners to think beyond short-term interventions aimed at improving schools or teachers like those described in the NCTQ report. Instead, real educational change may require greater understanding of how structural inequalities and oppressive systems create and perpetuate opportunity gaps with real consequences for students.\textsuperscript{89}

The report does not address teacher development or professional learning in a meaningful way, and instead celebrates the six systems without offering comprehensive empirical evidence of their putative successes. It also ignores the influx of additional funding that enabled these agencies to design and implement the profiled systems, including at least $75 million dollars in direct grants from the Bill and Melinda Gates Foundation.\textsuperscript{90} This omission is noteworthy because it suggests that even if these reforms are effective, those results may not be possible elsewhere without a similar external investment. Although NCTQ tries to suggest otherwise, perhaps that money would have been better spent investing in teacher and student learning and programs to address persistent opportunity gaps. As it stands, this report does little to enrich an already tired conversation about linking teacher evaluation to student achievement.
Notes and References


Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS Education Value-Added Assessment System (EVAAS). *Education Policy Analysis Archives, 22*(98), 1-41;


6 This literature is sizable; the following references provide an introduction to the key concerns: Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge;


For example, NCTQ's board of directors currently includes F. Mike Miles, the former superintendent of Dallas Independent School District who was largely credited with the development and implementation of the teacher evaluation system there (https://www.dallasnews.com/news/education/2015/06/23/dallas-isd-chief-mike-miles-announces-resignation-after-contract-changes-rejected). Former chancellor of education Michelle Rhee—who instituted much of the reform in DC Public Schools—also served on their technical advisory board (https://www.washingtonpost.com/blogs/answer-sheet/post/ravitch-what-is-netq-and-why-you-should-know/2012/05/23/gJQA7CrIU_bloghtml/utm_term=.e9c9aa604c1).

According to their publicly available database of funded projects, Gates provided over $8 million in direct grants to Dallas Independent School District, $28 million to Denver Public Schools and the Denver Public Schools Foundation, $7 million to the DC Public Education Fund, and $7 million to the New Schools Venture Fund to support the development of teacher evaluation reforms and data systems in Newark and Washington D.C. In Tennessee, Gates' investment has been even larger, with a combined total of over $140 million in education grants awarded to various districts, agencies, and foundations in the state. Of that $140 million, $14 million was awarded directly to the Tennessee Department of Education, and another $15 million to the Tennessee State Collaborative on Reforming Education. Bill and Melinda Gates Foundation. (n.d.). *Grantmaking awarded grants.* Seattle, WA: Author. Retrieved October 15, 2018, from https://www.gatesfoundation.org/How-We-Work/Quick-Links/Grants-Database

NCTQ has also received a significant investment from Gates, with seven grants totaling more than $15 million.

The report’s website explicitly refers to the report as a study (e.g., “In this study, NCTQ examines evidence of the impact of teacher evaluation in six places....”) and the formal press release refers to it as an original analysis (i.e., “Our analysis suggests that moving forward with the teacher evaluation systems presents students and teachers with a huge opportunity”). National Center for Teacher Quality (n.d.). *Making a difference: Six places where teacher evaluation systems are getting results* (web landing page). Washington, D.C.: Author. Retrieved October 30, 2018, from https://www.nctq.org/publications/Making-a-Difference;

Cooper, H., & Hedges, L.V. (2009). Research synthesis as a scientific process. In H. Cooper, L.V. Hedges, & J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis (2nd ed.)*, pp. 3-16. New York,
It is worth noting that sources cited within each profile but which were not used as data sources or pieces of evidence about the teacher evaluation system in question are not included in this table. For example, the Dallas profile cites an ASCD article on the provision of feedback to improve teaching practice; this source is used to buttress the claim that the district’s short “spot evaluations” have the power to “save principals time and give them more real-time opportunities to engage with teachers and learn about their strengths and areas for growth” (p. 5). However, the ASCD article does not actually present evidence about Dallas—it instead addresses effective feedback practices in general—so it does not appear in the table.


The vast majority of the data presented in the profile of Denver Public Schools (DPS) appears to have been drawn from a single interview—and subsequent personal communication—with Nicole Wolden, the Director of Growth and Performance at DPS. Her interview and other personal communications are cited 10 times in the report, while other references regarding the DPS system are only cited three times. The report also acknowledges Sarah Almy, DPS Executive Director of Talent Management, but the district profile
never references any data or evidence based on an interview with her. It is therefore unclear whether she participated in an interview or simply provided feedback on the final report.


41 The NCTQ report’s endnotes only reference personal communication—not an interview—with Christopher Lewis, the then current Deputy Chief of IMPACT in DCPS.


54 Like Denver, much of the data presented in the profile of Newark Public Schools (NPS) appears to have been drawn from a single interview with Larisa Shambaugh, Former Chief Talent Officer at NPS. This interview is cited nine times in the 15 included endnotes. Putman, H., Walsh, K., & Ross, E. (2018). Making a difference: Six places where teacher evaluation systems are getting results. Washington, DC: National Council on Teacher Quality.

55 Unlike Denver, NCTQ was not able to engage any current NPS staff in their data collection efforts, noting in endnote 72 that “current district leadership did not respond to requests to review the information before publication” (p. 23). Putman, H., Walsh, K., & Ross, E. (2018). Making a difference: Six places where teacher evaluation systems are getting results. Washington, DC: National Council on Teacher Quality.


60 It appears Christopher Eide, the Director of Educator Quality at the New Mexico Public Education Department, participated in an email exchange with NCTQ in lieu of a full interview; see endnote 97 of Putman, H., Walsh, K., & Ross, E. (2018). Making a difference: Six places where teacher evaluation systems are getting results. Washington, DC: National Council on Teacher Quality, pg. 24.


http://nepc.colorado.edu/thinktank/review-teacher-eval
The NCTQ report draws several pieces of evidence regarding system performance from the following NMPED source, which unfortunately appears to have been removed from the NMPED website: New Mexico Public Education Department (2018). *Elevating the teaching profession in New Mexico: 2018 NMTEACH results, statewide overview*. Santa Fe, NM: Author.


http://nepc.colorado.edu/thinktank/review-teacher-eval


80 It appears that this piece of data is only the tip of the proverbial iceberg: a closer examination of Dallas’ data suggests this dissatisfaction persists among the most highly rated teachers. Dallas teachers are rated into nine effectiveness levels, including unsatisfactory, progressing (I and II), proficient (I, II, and III), exemplary (I and II), and master. Among the 5,498 teachers in the system, only 54% of the 1,214 teachers rated in the top two proficient and top two exemplary categories reported being satisfied or very satisfied with the system. Barton, M., & Palladino, D. (2018). 2016-17 evaluation of Teacher Excellence Initiative (TEI). Dallas, TX: Dallas Independent School District, Department of Evaluation and Assessment. Retrieved October 30, 2018 from, https://www.dallasisd.org/cms/lib/TX01001475/Centricity/domain/98/evaluation/16-17/finalrpts/EA18-539-2%20Full%202017-18%20TEI%20Report.pdf


