4. School Staffing and Teacher Quality

Thomas S. Dee

Executive Summary

Good teaching is deeply important for its immediate impact on both student learning and multiple, longer-run dimensions of educational and economic success. However, the effectiveness of individual teachers is highly variable and unevenly distributed across students. Fortunately, research over the last decade shows that strategies for improving the performance of in-service teachers have considerable promise. For example, focused training can significantly amplify the impact of teachers on student learning. Such professional development appears to be particularly effective when it emphasizes specific challenges of classroom practice. Similarly, performance-based teacher assessment systems can guide effective professional development and introduce high-powered incentives for teacher excellence, as well as establish informed procedures for directing chronically underperforming teachers out of the classroom.

However, the practical challenges to realizing such meaningful improvements in the effectiveness of the teaching workforce—and doing so consistently at scale—are considerable. For example, the exact ways to design and deliver consistently effective professional development for teachers are uncertain. This strongly indicates the need to embed new and ongoing professional development efforts within purposive cycles of design and evaluation. Similarly, there are several substantive logistical and political barriers to introducing effective and enduring systems of teacher performance assessment. These include the challenges of designing aligned and accurate data systems and assessment measures that reliably capture the variation in teacher performance and coupling them with the clear communication of reliably implemented incentives. The perception of political durability may also be key to the success of such teacher assessment reforms. However, recent research studies have identified several initiatives that serve as encouraging proof points for the promise of these reforms.

- *A Nation at Risk* placed a focus on teacher quality and anticipated some of the most dramatic education policy innovations of the past forty years. Creating large-scale, lasting changes related to teacher effectiveness has proved challenging, however.

- To move forward, we need a deeper understanding of how best to design reforms such as new teacher evaluation models or improved teacher professional development. Political barriers have also stood in the way of taking these reforms to scale.

- Policymakers should explore how to build broader coalitions around teacher effectiveness and perhaps use more incremental approaches to help build the evidence base for more lasting and scalable reform.

• • •

The publication of *A Nation at Risk* (*ANAR*) in 1983 was the defining moment of the "first wave" of education reform (NCEE 1983). It articulated improbably long-lived insights that continue to define education policy and discourse to this day. In particular, *ANAR* underscored, with uncommon rhetorical flourishes, the contrast between the ambitious ideals of a "Learning Society" and existing educational standards defined by modest minimum requirements, such as the low expectations embedded in high schools' minimum competency tests and "cafeteria-style" curricula. Clearly, *ANAR*'s most prominent recommendation was the adoption of high school graduation requirements grounded in a "New Basics" curriculum that would feature four years of English; three years of science, math, and social studies; a half year of computer science; and, for college-bound students, two years of foreign-language instruction.

However, *ANAR* also commented on several other dimensions of the education system in the United States, including the state of the teaching profession. In particular, *ANAR* concluded that "too many teachers are being drawn from the bottom quarter of graduating high school and college students" (22). The report also underscored the inadequate subject-matter focus of teacher training, low pay, teachers' limited influence on key professional decisions (e.g., textbooks), and the targeted character of teacher shortages. These findings—and the seven specific recommendations *ANAR* made regarding teaching—have been the focus of education research, commentary, and policymaking to this day.

In this chapter, I provide a compact overview of key insights from the research and policymaking that occurred in the wake of these recommendations. I focus specifically on the developments relevant to *in-service* teachers, while the important issues related to recruitment, induction, and mentoring in the teaching profession are addressed separately by Michael Hansen in chapter 3. *ANAR* made four specific recommendations relevant to in-service teachers. One is that teacher salaries should be "professionally competitive, market-sensitive, and performance-based" and linked to "an effective evaluation system" that rewards effective teachers and guides underperforming teachers toward improvement or termination. A related second recommendation advocates for collectively developed "career ladder" designations that distinguish beginning, experienced, and master teachers. *ANAR*'s remaining two recommendations for in-service teachers focus on supporting teacher improvement through funded time for professional development (30–31).

THEORIES OF ACTION

ANAR's recommendations for in-service teachers tacitly reflect two broad and complementary theories of action for improving teacher effectiveness and student outcomes. One involves improving the effectiveness of existing teachers. The intent is for this to occur through professional development activities and through the implementation of well-designed financial and professional incentives. Both of these intend to promote an understanding of high-quality classroom practices as well as their consistent use. The second theory of action focuses on selection—that is, performance assessment systems designed to retain and elevate the most effective teachers while ensuring that persistently ineffective teachers exit the classroom. Notably, these policy recommendations stand in sharp contrast to conventional efforts to promote teacher effectiveness through generic salary increases unrelated to performance or need and through reducing class sizes by hiring more teachers.

The motivations for *ANAR*'s theories of action rest upon several important stylized facts about teachers that have become increasingly well established since its publication. Arguably, the most foundational evidence concerns the variation in effectiveness across teachers. An older debate had questioned whether there are aspects specific to teaching that make it prohibitively difficult to measure teacher effectiveness in a valid and reliable manner (Murnane and Cohen 1986; Ballou 2001). However, richer data and methodological advances have led to a consensus about the general validity of teacher effectiveness measures while also acknowledging important evidence on the degree of noisiness common to such measures (Staiger and Rockoff 2010).

These studies indicate that the variation in teacher effectiveness is large, particularly relative to the effects of other promising education interventions. Specifically, a one-standard-deviation improvement in teacher effectiveness corresponds to a gain in student performance on standardized tests of roughly 0.1 to 0.2 standard deviations (e.g., Rivkin et al. 2005; Rockoff 2004; Aaronson et al. 2007; Staiger and Rockoff 2010). Critically, the manner in which teachers are currently assessed—that is, informal, "drive-by" evaluations—captures virtually none of this documented variation, rates the vast majority of teachers as satisfactory, and results in little performance-based attrition of low-performing teachers from the classroom.

Another important stylized fact is that, at the hiring stage, school leaders have little capacity to identify the teachers who will become more effective (Staiger and Rockoff 2010). This combination of facts—that teachers vary considerably in impact, but this impact can be observed much more easily after several years in the classroom than at the hiring stage—suggests the need for broader access to the teaching profession coupled with discerning assessment systems that guide subsequent personnel decisions. In particular, decisions to tenure rather than dismiss the lowest-performing teachers can have dramatic consequences given the length of teaching careers (Staiger and Rockoff 2010).

Over the past fifteen years, this evidence has motivated a number of ambitious public and philanthropic efforts to systematically improve the effectiveness of the teacher workforce through performance-based assessment systems. Recent research has also provided more credible evidence of direct initiatives designed to improve the performance of all in-service teachers through professional development. I discuss these policy innovations and the related research below.

IMPROVING TEACHER EFFECTIVENESS

ANAR recommended that teachers receive eleven-month contracts so that they could spend more time in professional development and provide additional instruction for students with special needs. While the eleven-month contract has not been widely adopted, broader efforts to improve the performance of in-service teachers through direct training and support involve a substantial expenditure of time and money. However, accurately identifying the magnitude of these outlays is not straightforward given the accounting challenges of categorizing such activities and their demands on time for both teachers and nonteaching staff. For example, a study by Alexander and Jang (2019) examined expenditure reports for Minnesota school districts and found that 1 percent of 2013–14 operational expenditures was spent on activities defined by the state as staff development. In contrast, a study by the New Teacher Project (2015) found that 2013–14 expenses related to teacher improvement constituted, on average, 8 percent of district budgets. This figure consisted of both direct expenditures on teacher improvement, such as professional development, coaching, and new-teacher support, as well as related indirect expenditures, such as the management, strategic, and operational expenses for these improvement efforts.

Focusing specifically on professional development, a study commissioned by the Gates Foundation (2014) found that the typical teacher spends sixty-eight hours per year on professional learning directed by districts, or eighty-nine hours when courses and self-guided professional learning are included. Most of the time spent by teachers in professional development occurs in workshops and professional learning communities conducted by district staff. The cost of this professional development was estimated at $18 billion per year in 2014. Teacher perceptions of the quality of these investments have generally not been encouraging, nor do they appear to have clear links to teacher performance or improvement (TNTP 2015; Gates Foundation 2014). The Gates report also stresses the overwhelming use of district staff instead of market-tested external providers to provide professional development, as well as limited teacher voice in choosing their training.

Despite the considerable expense and prominence of teacher professional development, credible research on the impact of these investments has also been quite limited over much of the period since *ANAR*'s publication. For example, Yoon et al. (2007) reviewed more than 1,300 studies potentially addressing the impact of teacher professional development on student learning and found only nine studies that met the evidence standards in the federal What Works Clearinghouse: six randomized controlled trials and three quasi-experimental studies

conducted between 1986 and 2003. However, what these studies revealed suggests a striking proof of concept: teachers who received substantial professional development could boost the achievement of the average control-group student by 21 percentile points. Notably, these nine professional development initiatives focused on elementary grades but differed in their theories of action (Yoon et al. 2007).

However, other quasi-experimental studies serve as a reminder that implementing effective professional development consistently at scale is a serious challenge. Jacob and Lefgren (2004) examined the effect of teacher training in Chicago Public Schools using a credible natural experiment in which schools with low baseline test scores received additional resources for staff development. They found that this initiative had "no statistically or academically significant effect" on math or reading achievement of elementary students. Similarly, Harris and Sass (2011) examined student-level longitudinal data linked to teacher data for the state of Florida and did not find an overall impact of professional development on teacher productivity. However, they did find positive effects of content-focused math professional development on student outcomes at the elementary and middle-school levels.

Over the past decade, experimental studies of teacher professional development have proliferated. In general, they have provided mixed evidence of the learning impact of investments in professional development. For example, experimental studies by Garet et al. (2008, 2010) found that reading- and math-focused training changed teacher knowledge and practice but without clearly improving student achievement. However, meta-analytic summaries of such experimental professional development evaluations suggest that positive effects exist but vary considerably by program design. For example, Basma and Savage (2018) examined seventeen literacy-focused professional development studies and found an overall effect size for reading achievement of 0.225. Similarly, in a meta-analysis of ninety-five STEM-focused professional development studies with experimental and quasi-experimental designs, Lynch et al. (2019) report an average effect size of 0.21.

However, other multisubject meta-analyses suggest smaller but still positive effects on student learning. For example, Fletcher-Wood and Zuccollo (2020) identified fifty-three experimental evaluations of teacher professional development and found an overall effect size of 0.09. Similarly, Sims et al. (2021) reviewed 104 experimental evaluations and found an overall effect size of 0.05. Given the considerable financial expense of most training investments, effects of this size, though positive, raise serious questions about cost-effectiveness.

These reviews also note and seek to examine the considerable variation across professional development programs in terms of impact. Kennedy (2016) argues that the widely discussed design features of teacher professional development—namely program duration, emphasis on content knowledge, and use of professional learning communities—are far less relevant than whether the training addresses any of the four persistent challenges of teaching: portraying content, managing student behavior, enlisting student participation, and knowing what students understand. In a similar vein, Sims et al. (2021) characterize professional development programs by the more general ways they change teacher skills and behaviors. Specifically,

they characterize teacher professional development by four "IGTP" traits that indicate whether teachers are provided with new insights (I), goal-oriented behaviors (G), and techniques (T) that are embedded in practice (P). And they conclude that professional development programs with all four traits have an effect size on student learning of 0.17. However, these assessments may obscure the relevance of professional development initiatives that focus on the most effective elements of content and practice, such as an emphasis on "science of reading" approaches in literacy-focused training.

Overall, this evidence indicates that *ANAR* was prescient in emphasizing the need for ongoing training of in-service teachers. The available evidence suggests that such training can have substantial effects on student learning. However, realizing the increasingly well-established potential of this training is not straightforward. It involves the perennial challenge of translating research findings—that is, the critical design features of effective professional development—into genuine changes in high-impact practice at scale.

TEACHER EVALUATION AND PERFORMANCE-BASED INCENTIVES

ANAR also made prominent recommendations to dramatically change how we pay and evaluate public school teachers. In general, the status quo to this day compensates teachers according to single-salary schedules that rigidly structure pay according to years of experience and observed qualifications (e.g., a graduate degree) that do not consistently predict teacher effectiveness. This approach has historical origins in well-intentioned efforts to eliminate overt discrimination and capriciousness in teacher pay. Today, critics allege that this inflexible approach has led to low and undifferentiated salaries that do little to attract, motivate, and retain the most-effective teachers and to direct the least-effective teachers out of the classroom, particularly in hard-to-staff schools and high-need subjects. Furthermore, this approach to pay is coupled with low-stakes, "drive-by" teacher evaluations that capture little of the variation in teacher performance and do not provide reliable guidance for professional learning (Weisberg et al. 2009).

ANAR envisioned an alternative in which teacher compensation was substantially higher but also based on performance in a manner that would direct persistently underperforming teachers either to improve or to leave the profession. In the aftermath of *ANAR*'s publication, several states and districts experimented with providing teachers with extra pay and career-ladder recognitions for demonstrated merit (though, not generally, dismissing chronically underperforming teachers). These reforms tended to be short-lived despite encouraging results (Cornett and Gaines 1994). While the rollback of these reforms was clearly a policy choice, the underlying causes are debated. Ballou (2001) argued that it largely reflected the opposition of teachers' unions. Murnane and Cohen (1986) contended that it reflected the distinctive character of teachers' professional practice—that is, multidimensional and difficult to observe. However, random-assignment evidence from a comparatively well-implemented

career ladder program in Tennessee indicates that it was effective in identifying teachers who raised student achievement (Dee and Keys 2004).

The past two decades have witnessed a diverse variety of ambitious efforts, often encouraged by prominent philanthropic and federal initiatives, to measure teacher performance and to link it to improvement supports and incentives such as financial benefits, career-ladder designations, and dismissal threats. The research on these different reforms suggests their promise but also underscores the nontrivial challenges (e.g., design features, implementation, and political credibility) that make the consistent realization of this promise difficult. For example, the Obama administration's Race to the Top (RttT) initiative disbursed more than $5 billion to states in a competition based in part on their commitment to developing systems for promoting teacher effectiveness. While RttT was effective in promoting state policy adoption (Howell and Magazinnik 2017), its effects on key design features and implementation are far less clear. In particular, while states were more likely to have multiple measures of teacher performance in the wake of RttT, the use of this data to inform salary and retention decisions remained uncommon (Hallgren, James-Burdumy, and Perez-Johnson 2014). The state reforms over this period were "rarely sustained over time," offered low bonuses, and rated fewer than 1 percent of teachers as unsatisfactory (Bleiberg et al. 2021).

A more granular focus on the available evidence from specific initiatives provides richer insights into these issues of design, implementation, and political durability. For example, several studies focused narrowly on simply providing teachers with incentives for improved performance. These studies often found null (or weak) effects that are likely to reflect the unique character of these programs. "Cash for test scores" experiments with individual incentives for teachers in Nashville (Springer et al. 2012) and group incentives for teachers in Round Rock, Texas (Springer et al. 2013), found little to no evidence of effects on teacher practices, attitudes, and the learning gains of their students. Similarly, studies of a group-based teacher-incentive experiment in New York City (Goodman and Turner 2013; Fryer 2013) found that they had no overall effects on key teacher or student outcomes.

Critics of teacher incentives suggest that these null findings reflect a misunderstanding of teacher motivations and the manner in which such incentives might debase intrinsic motivation (e.g., Murnane and Cohen 1986). However, three design features of these studies could also contribute to these null findings and have important implications for performance-based assessment and compensation. First, the fact that participants know that these experimental incentives have a short term (e.g., two years) can sharply attenuate the resulting motivation to undertake changes in professional practices. This same concern can also apply to the incentives embedded in at-scale policy reforms that are viewed as faddish and unlikely to endure politically. Second, these initiatives generally focused on student achievement as the incentivized outcome. This may weaken the impact of incentives if teachers do not see or understand how they should change everyday practice to realize these rewards. A related third point is that these incentive studies generally did little to support and guide teachers in how they could change their professional practices to earn these rewards.

Three other studies suggest the potential importance of other design features. A teacher-incentive study in Chicago Heights, Illinois, found positive effects on student achievement (but only in the first wave of the experiment) when the incentives were framed as the loss of an award rather than a gain (Fryer et al. 2022). Second, the Talent Transfer Initiative (TTI) found positive effects when offering high-performing teachers a high-powered incentive ($20,000) linked to a distinctly clear, easily observed, and important behavior: working in a hard-to-staff school for two years (Glazerman et al. 2013). However, it is notable that these incentive-based gains were difficult to realize. More than 1,500 teachers had to be approached in order to fill only eighty-one vacancies. Third, the Accelerating Campus Excellence (ACE) program in Dallas similarly provided large incentives to highly effective teachers willing to work in hard-to-staff schools. Morgan et al. (2023) presented evidence that ACE produced dramatic gains in student performance: a 0.3 effect size in reading and 0.4 in math. This study also found that this success replicated as the program went to scale and that these gains were reversed when the program was eliminated.

Notably, these focused incentive programs all fall short of the more comprehensive system of assessments, supports, and incentives recommended by *ANAR*. TAP: The System for Teacher and Student Advancement (formerly known as the Teacher Advancement Program), which was introduced in 1999 and is currently active in "nearly twenty states and hundreds of school districts across the US" (Cohodes, Eren, and Ozturk 2023), is closer to *ANAR*'s vision. Specifically, the defining features of TAP include career ladder designations for teachers and job-embedded, professional learning led by master teachers. In support of this professional learning, TAP also provides teachers with comprehensive evaluations of their professional practice. However, it is not clear that this "instructionally focused accountability" articulates clear mechanisms for directing consistently low-performing teachers out of the classroom (the selection mechanism in *ANAR*'s theory of change). Finally, TAP includes performance pay typically linked to observations of teachers' professional practice, such as classroom observation, portfolios, and interviews, as well as test scores.

The available evidence suggests that TAP is effective in improving teacher performance and student outcomes. Specifically, in a quasi-experimental study based on 1,200 schools from two states, Springer, Ballou, and Peng (2014) found that TAP increased student performance, particularly at the elementary school level, with effect sizes varying from 0.12 to 0.34 by grade. Similarly, Cohodes, Eren, and Ozturk (2023), leveraging the rollout of TAP across schools in South Carolina, found that it generated improvements in several long-run outcomes, including educational attainment, criminal activity, and the take-up of government assistance. However, a random-assignment evaluation of TAP in Chicago schools by Glazerman and Seifullah (2012) found that it did not improve student achievement and that it was also vexed by the challenges of implementing this reform with fidelity, such as teacher payouts being smaller than originally stated and no rewards based on value added because of inadequate data systems.

Two other high-profile studies provided further evidence of the serious challenges of implementing comprehensive reforms of teacher assessments and compensation as well as of credibly assessing their effects. The first example is the federal Teacher Incentive Fund (TIF). Congress established TIF in 2006 to provide grants to high-need schools implementing

performance-based compensation systems. The four required components of TIF reforms also resembled those suggested by *ANAR*: (1) measures of teacher performance, including observations of classroom practice; (2) large, differentiated, difficult-to-earn performance bonuses; (3) additional pay for career-ladder opportunities, such as becoming a master teacher and coach; and (4) professional development linked to the teacher assessments. A congressionally mandated study of TIF focused on the 2010 grant recipients in more than 130 school districts and found it led to student achievement of 1 to 2 percentile points higher in reading and math (Chiang et al. 2017).

However, there are two important caveats to this evidence of modest impact. First, the implementation of these reforms in the study districts was incomplete. Only about half of the participating districts reported implementing all four components of the reforms required by TIF. In particular, professional development was frequently not provided, and most teachers received bonuses, "a finding inconsistent with making bonuses challenging to earn" (Chiang et al. 2017). Second, the treatment–control contrast assessed in this random-assignment study did not examine the effect of TIF versus "business as usual." Instead, the treatment schools in the study were intended to receive pay-for-performance bonuses while the control group received automatic bonuses. And all study participants, both treatment and control, were assigned access to the three other TIF components: career ladder responsibilities and rewards, evaluative feedback, and professional development. In this critical but often overlooked detail, the federal study of TIF more closely resembles the studies of teacher incentives noted above than a true evaluation of teacher assessment systems.

The Gates-funded Intensive Partnerships for Effective Teaching initiative is a second widely discussed example of implementing and evaluating teacher assessment systems. This initiative sought to introduce assessment reforms within three school districts and four charter management organizations. Similar to both TAP and TIF, this effort featured focused professional development and career ladder incentives along with performance pay and retention decisions based on direct, structured observation of teacher practice and value-added scores. A quasi-experimental study found that these reforms did not clearly improve the focal student outcomes of high school graduation and college attendance (Stecher et al. 2018). However, the implementation of the reforms appears to have been weak. The teacher evaluations flagged few teachers as poor performers, and in sites with available data, only 1 percent were dismissed for poor performance. As with the federal TIF evaluation, the treatment contrast that was studied was muted because the comparison schools in this study often adopted similar policies.

IMPACT, the highly controversial teacher assessment reforms introduced in the District of Columbia Public Schools (DCPS), is distinctive as a seminal and enduring effort to implement *ANAR*'s recommendations with fidelity. IMPACT evaluated DCPS teachers on multiple measures with a heavy emphasis on structured classroom observations, including some conducted by district staff, and linked professional development. These evaluations resulted in measures of teacher performance that exhibited variation rather than being largely uniform. IMPACT linked these measures to high-stakes consequences: substantial pay increases for "highly effective" teachers, particularly those in high-poverty schools; dismissal for a small

number of "ineffective" teachers; and a dismissal threat for "minimally effective" teachers who did not become effective within a year.

A quasi-experimental study of the incentive contrasts embedded in IMPACT found it had positive effects on teacher performance (Dee and Wyckoff 2015). This study's design leveraged a feature of IMPACT in which teachers with performance scores just below a threshold value were deemed "minimally effective" and subject to a dismissal threat while those with scores at or above the threshold were not. A comparison of teachers just below and above this threshold found that the threat of dismissal caused minimally effective teachers either to leave the district or to improve their measured performance substantially. A powerful financial incentive for highly effective teachers to repeat their prior performance also appeared to have positive effects.

Three other aspects of IMPACT merit emphasis. First, the political credibility and resiliency of IMPACT appeared to be highly salient. In 2010, when the city (and district) leadership who championed IMPACT were forced out of office, the first "minimally effective" designations did not appear to change teacher behavior. However, the ratings reported in the summer of 2011, when it appeared that IMPACT would endure, did drive changes in teacher behavior.

Second, evidence indicates that IMPACT not only improved the performance of existing teachers but also replaced underperforming teachers who exited with substantially more effective instructors. Specifically, a quasi-experimental study by Adnot et al. (2017) finds that, when a low-performing teacher exited, their replacement raised student performance by 0.14 standard deviations in reading and 0.24 standard deviations in math. Third, the performance benefits of IMPACT's incentives endured through subsequent revisions to the teacher supports and ratings structure (Dee, James, and Wyckoff 2021).

A second district reform of note (and one with strong parallels to IMPACT) began in the Dallas Independent School District in 2015. Specifically, like IMPACT, the Teacher Excellence Initiative (TEI) replaced a single-salary schedule with compensation based on multiple measures of teacher performance. Furthermore, like IMPACT, it also did so in the context of accountability for school principals. TEI also implemented a unique design feature to discourage inflated or arbitrary ratings of teachers. It fixed the overall distribution of ratings and penalized principals for subjective ratings that were highly misaligned with test-based ratings. A synthetic-control study by Hanushek et al. (2023) found that these reforms led to statistically significant increases in student achievement that grew over time to a roughly 0.2 standard deviation in math and a 0.1 standard deviation in reading.

CONCLUDING THOUGHTS

ANAR's recommendations that focused on improving the effectiveness of in-service teachers were a harbinger of some of the most dramatic education policy innovations of the past forty years. And these innovations have provided us with several proofs of concept and new

insights that establish the potential to improve student learning through dramatic changes in teacher evaluation, in-service training, and compensation.

However, it must also be acknowledged that there has clearly not been large-scale, lasting change regarding *ANAR*'s teacher-focused recommendations. Uninformative, low-stakes assessments of professional practice and rigid single-salary schedules are still the norm for the vast majority of teachers in US public schools. And while in-service teachers do engage in extensive professional development, the impact of these expensive and highly variable investments is uncertain at best.

Any serious effort to reimagine the assessment, training, and compensation of in-service teachers should begin by confronting the factors that have contributed to the long durability of the status quo. There appear to be three broad and interrelated impediments to substantive change. The first is the need to improve the knowledge base of how best to design the key features of these reforms. For example, efforts to improve teacher evaluation and introduce performance-based teacher pay rely critically on valid and reliable measures of teacher performance. Promising gains in measuring teacher effectiveness are likely to come from continued improvements to structured rubrics for classroom practices. Incentives can better guide the professional improvement of teachers when they are linked to the high-impact, everyday classroom practices teachers directly control and can enhance through complementary training.

Another important area where improved knowledge is critical to driving at-scale change concerns the design of teacher professional development. The typical professional development experience, workshops directed by internal district staff, is often criticized (e.g., the New Teacher Project 2015). At the same time, a recent and growing body of experimental studies indicates that purposively designed professional development can have substantial impact. This literature generally emphasizes the particular benefits of in-service training that focuses on meeting more general challenges of teacher practice (e.g., Kennedy 2016; Sims et al. 2021). While more can be learned about the design of professional development, the question of how to design its delivery is even more uncertain. A study from the Gates Foundation (2014) suggests that relying more on external providers of professional development will make it easier to move nimbly to market-tested and effective approaches. However, several of the teacher assessment reforms discussed here instead emphasize redesigning internally provided professional development to rely on master teachers who may be better positioned to serve as coaches providing embedded and relevant training. These issues underscore the need to build a complementary learning agenda around any new reforms (e.g., inquiry cycles, networked improvement communities).

A second impediment to realizing *ANAR*'s vision concerns the multifaceted operational challenges of implementing meaningful reforms effectively at scale. The null findings from credibly identified studies of professional development in at-scale field settings suggest this issue (Jacob and Lefgren 2004; Harris and Sass 2011). However, more-direct and sobering evidence comes from several well-funded, high-profile efforts to introduce teacher assessment and

compensation reforms at some scale. These include (1) the failure to deliver value-added bonuses because of data-system inadequacies in TAP (Glazerman and Seifullah 2012); (2) the limited variation in teacher ratings and their infrequent use in personnel decisions in the Gates Foundation's Intensive Partnership for Effective Teaching (Stecher et al. 2018); (3) the inconsistent delivery of professional development and the broad distribution of bonuses under the federal Teaching Incentive Fund (Chiang et al. 2017); and (4) the limited use of teacher evaluations to guide salary and retention decisions under the RttT initiative (Hallgren, James-Burdumy, and Perez-Johnson 2014).

A third and closely related impediment is political opposition. With regard to introducing performance-based pay, this most obviously refers to the opposition of teachers' unions. However, it can also involve unresponsive public-sector bureaucracies. Furthermore, reform efforts can also fail when their success and durability rely on politically determined funding commitments. The political opposition to reform in the broader public also turns on misinformation about what the existing evidence discussed here actually indicates. Specifically, opponents of the types of reforms recommended by *ANAR* often argue that investments in professional development are effective while performance-based pay has failed.

Given these interlocking issues, a compelling way to achieve change at scale may involve forming political coalitions around compelling reforms that adopt some but not all of *ANAR*'s proposals. For example, it may be possible to move school districts toward more effective professional development delivered by a carefully curated set of outside vendors if their provision involved cost-sharing that saved district resources. Alternatively, it may be possible to achieve durable political support for a teacher evaluation system if that system focuses narrowly on identifying master teachers and providing them with training and extra pay to coach their peers but takes a more incremental approach toward dismissing underperforming teachers. Intentionally combining such efforts with careful evaluation could, over the longer term, seed further evidence-based change in this important domain.

HESI PRACTITIONER COUNCIL RESPONSES

Essays in this series were reviewed by members of the Hoover Education Success Initiative (HESI) Practitioner Council. For more information about the Practitioner Council and HESI, visit us online at hoover.org/hesi.

We are worse off than we were forty years ago relative to the state of the teaching profession. I wonder the extent to which our reforms have contributed to the decay. In the years following the publication of *ANAR*, my grandmother, who had graduated at the top of her class, retired from teaching earlier than she had planned because her state instituted a teacher test she was not interested in taking. Twenty years later, I started teaching in my home district. Had I not been certified at the time, I would not have been considered for an interview. Due to the teacher shortage today, states are creating a plethora of on-ramps into the profession, and

districts are forced to hire anyone they can find. While the teacher reform initiatives have been well intended, the application has come at high cost and low benefit. I applaud and wholeheartedly support Dee's assertion that we should build political coalitions and pivot to focus teacher evaluations on identifying excellence. We need to also prioritize improvement-focused feedback in our support and evaluation systems.

—Holly Boffy, Louisiana State Board of Elementary and Secondary Education

Author Thomas Dee concludes: "There has clearly not been large-scale, lasting change regarding *ANAR*'s teacher-focused recommendations. Uninformative, low-stakes assessments of professional practice and rigid single-salary schedules are still the norm for the vast majority of teachers in US public schools. And while in-service teachers do engage in extensive professional development, the impact of these expensive and highly variable investments is uncertain at best." His research review supports that conclusion, and I agree. However, I feel that we should not stop these efforts; rather, we should significantly improve them. While teacher evaluation systems may never be implemented into a performance pay system, the information (ratings) is critical in ensuring appropriate staffing, developing meaningful career ladders, and identifying targeted personal professional development.

In addition to Dee's closing recommendations, which I think are helpful, a significant piece still missing from the overall analysis is the critical need for strong school leadership to address the effectiveness of in-service teachers. This has not been seriously addressed in the research designs nor in the implementation of teacher evaluation systems.

—Angelika Schroeder, Colorado State Board of Education

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25, no. 1 (January): 95–135.

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2017. "Teacher Turnover, Teacher Quality, and Student Achievement in DCPS." *Educational Evaluation and Policy Analysis* 39, no. 1 (March): 54–76.

Alexander, Nicola A., and Sung Tae Jang. 2019. "Expenditures on the Professional Development of Teachers: The Case of Minnesota." *Journal of Education Finance* 44, no. 4 (March): 385–404.

Ballou, Dale. 2001. "Pay for Performance in Public and Private Schools." *Economics of Education Review* 20, no. 1 (February): 51–61.

Basma, Badriah, and Robert Savage. 2018. "Teacher Professional Development and Student Literacy Growth: A Systematic Review and Meta-Analysis." *Educational Psychology Review* 30 (June): 457–81. https://doi.org/10.1007/s10648-017-9416-4.

Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew Springer. 2021. "The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms." EdWorkingPaper 21-496. Retrieved from Annenberg Institute at Brown University. https://doi.org/10.26300/b1ak-r251.

Chiang, Hanley, Cecilia Speroni, Mariesa Herrmann, Kristin Hallgren, Paul Burkander, and Alison Wellington. 2017. *Evaluation of the Teacher Incentive Fund: Final Report on Implementation*

and Impacts of Pay-for-Performance across Four Years. NCEE 2018-4004. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Cohodes, Sarah, Ozkan Eren, and Orgul Ozturk. 2023. "Teacher Performance Pay, Coaching, and Long-Run Student Outcomes." National Bureau of Economic Research Working Paper Series no. w31056.

Cornett, Lynn M., and Gale F. Gaines. 1994. "Reflecting on Ten Years of Incentive Programs: The 1993 SREB Career Ladder Clearinghouse Survey." Atlanta: Southern Regional Education Board.

Dee, Thomas, and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34, no. 2 (October): 267–97.

Dee, Thomas S., Jessalynn James, and James Wyckoff. 2021. "Is Effective Teacher Evaluation Sustainable? Evidence from District of Columbia Public Schools." *Education Finance and Policy* 16, no. 2 (Spring): 313–46.

Dee, Thomas S., and Benjamin J. Keys. 2004. "Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment." *Journal of Policy Analysis and Management* 23, no. 3 (Summer): 471–88.

Fletcher-Wood, Harry, and James Zuccollo. 2020. *The Effects of High-Quality Professional Development on Teachers and Students: A Rapid Review and Meta-Analysis*. London: Education Policy Institute.

Fryer, Roland G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics* 31, no. 2 (April): 373–407.

Fryer, Roland G., Jr. Steven D. Levitt, John List, and Sally Sadoff. 2022. "Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment." *American Economic Journal: Economic Policy* 14, no. 4 (November): 269–99.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa et al. 2008. *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement*. NCEE 2008-4031. Washington, DC: National Center for Education Evaluation and Regional Assistance

Garet, Michael S., Andrew J. Wayne, Fran Stancavage, James Taylor, Kirk Walters, Mengli Song, Seth Brown et al. 2010. *Middle School Mathematics Professional Development Impact Study: Findings after the First Year of Implementation*. NCEE 2010-4009. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Gates Foundation. 2014. *Teachers Know Best: Teachers' Views on Professional Development*. Seattle: Bill & Melinda Gates Foundation.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max. 2013. "Executive Summary." In *Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment.* NCEE 2014-4004. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Glazerman, Steven, and Allison Seifullah. 2012. *An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years*. Princeton, NJ: Mathematica Policy Research.

Goodman, Sarena F., and Lesley J. Turner. 2013. "The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." *Journal of Labor Economics* 31, no. 2 (April): 409–20.

Hallgren, Kristin, Susanne James-Burdumy, and Irma Perez-Johnson. 2014. *State Requirements for Teacher Evaluation Policies Promoted by Race to the Top*. NCEE Evaluation Brief. NCEE 2014-4016. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Hanushek, Eric A., Jin Luo, Andrew J. Morgan, Minh Nguyen, Ben Ost, Steven G. Rivkin, and Ayman Shakeel. 2023. "The Effects of Comprehensive Educator Evaluation and Pay Reform on Achievement." National Bureau of Economic Research Working Paper Series no. w31073.

Harris, Douglas N., and Tim R. Sass. 2011. "Teacher Training, Teacher Quality and Student Achievement." *Journal of Public Economics* 95, nos. 7–8 (August): 798–812.

Howell, William G. 2015. "Results of President Obama's Race to the Top." *Education Next* 15, no. 4 (Fall): 58–67.

Howell, William G., and Asya Magazinnik. 2017. "Presidential Prescriptions for State Policy: Obama's Race to the Top Initiative." *Journal of Policy Analysis and Management* 36, no. 3 (Summer): 502–31.

Jacob, Brian A., and Lars Lefgren. 2004. "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago." *Journal of Human Resources* 39, no. 1 (Winter): 50–79.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." MET Project Research Paper, Bill & Melinda Gates Foundation.

Kennedy, Mary M. 2016. "How Does Professional Development Improve Teaching?" *Review of Educational Research* 86, no. 4 (February): 945–80.

Lynch, Kathleen, Heather C. Hill, Kathryn E. Gonzalez, and Cynthia Pollard. 2019. "Strengthening the Research Base That Informs STEM Instructional Improvement Efforts: A Meta-Analysis." *Educational Evaluation and Policy Analysis* 41, no. 3 (September): 260–93. https://doi.org/10.3102/0162373719849044.

Morgan, Andrew J., Minh Nguyen, Eric A. Hanushek, Ben Ost, and Steven G. Rivkin. 2023. "Attracting and Retaining Highly Effective Educators in Hard-to-Staff Schools." National Bureau of Economic Research Working Paper Series no. w31051.

Murnane, Richard J., and David K. Cohen. 1986. "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive." *Harvard Educational Review* 56, no. 1 (April): 1–18.

NCEE (National Commission on Excellence in Education). 1983. *A Nation at Risk: The Imperative for Educational Reform*. Washington, DC: US Department of Education.

Podgursky, Michael J., and Matthew G. Springer. 2007. "Teacher Performance Pay: A Review." *Journal of Policy Analysis and Management* 26, no. 4 (Autumn): 909–50.

Rivkin, Steve G., Eric A. Hanushek, and John G. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73, no. 2 (March): 417–58.

Rokoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94, no. 2 (May): 247–52.

Sims, Sam, Harry Fletcher-Wood, Alison O'Mara-Eves, Sarah Cottingham, Claire Stansfield, Jo Van Herwegen, and Jake Anders. 2021. *What Are the Characteristics of Teacher Professional Development That Increase Pupil Achievement? A Systematic Review and Meta-Analysis*. London: Education Endowment Foundation.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2013. *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives.

Springer, Matthew G., Dale Ballou, and Art Peng. 2014. "Estimated Effect of the Teacher Advancement Program on Student Test Score Gains." *Education Finance and Policy* 9, no. 2 (Spring): 193–230.

Springer, Matthew G., John F. Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan Freeman Burns, Laura S. Hamilton, and Brian Stecher. 2012. "Team Pay for Performance: Experimental Evidence from the Round Rock Pilot Project on Team Incentives." *Educational Evaluation and Policy Analysis* 34, no. 4 (December): 367–90.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24, no. 3 (Summer): 97–118.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet et al. 2018. *Improving Teaching Effectiveness. Final Report*. Santa Monica, CA: RAND Corporation.

The New Teacher Project (TNTP). 2015. *The Mirage: Confronting the Hard Truth about Our Quest for Teacher Development*. New York: TNTP Inc. https://tntp.org/assets/documents/TNTP-Mirage_2015.pdf.

US Department of Education, Office of Elementary and Secondary Education, *State and District Use of Title II, Part A Funds in 2020–21*, Washington, DC, 2022.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. New York: The New Teacher Project.

Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley. 2007. *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement*. Issues & Answers Report, REL 2007-No. 033. Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

28 27 26 25 24 23 7 6 5 4 3 2 1

ABOUT THE AUTHOR

THOMAS S. DEE

Thomas S. Dee is the Barnett Family Professor at Stanford University's Graduate School of Education and a senior fellow (courtesy) at the Hoover Institution. Dee is also a senior fellow at the Stanford Institute for Economic Policy Research and a research associate with the programs on education, children, and health at the National Bureau of Economic Research.

A Nation at Risk + 40

The modern school-reform movement in the United States was set in motion by the release of the report *A Nation at Risk* in 1983. Countless education policy changes at the local, state, and national levels came as a result. *A Nation at Risk + 40* is a research initiative designed to better understand the impact of these efforts. Each author in this series has gone deep in a key area of school reform, exploring the following questions: *What kinds of reforms have been attempted and why? What is the evidence of their impact? What are the lessons for today's education policymakers?* As the nation's schools work to recover from the effects of the COVID-19 pandemic, this series not only describes the education-reform journey of the past forty years, it also provides timely and research-driven guidance for the future.

The Hoover Institution gratefully acknowledges Allan B. and Kathy Hubbard, the Daniels Fund, the William and Flora Hewlett Foundation, and the Koret Foundation for their generous support of the Hoover Education Success Initiative and this publication.

Hoover Institution, Stanford University
434 Galvez Mall
Stanford, CA 94305-6003
650-723-1754

Hoover Institution in Washington
1399 New York Avenue NW, Suite 500
Washington, DC 20005
202-760-3200

HOOVER INSTITUTION