

Working PAPER

BY BRIAN P. GILL (MATHEMATICA POLICY RESEARCH); JENNIFER S. LERNER (HARVARD UNIVERSITY); AND
PAUL MEOSKY (HARVARD UNIVERSITY)

Re-Imagining Accountability in K-12 Education: A Behavioral Science Perspective

May 2015

Abstract

Since the passage of the No Child Left Behind Act (NCLB) over a decade ago, the primary lever that American policymakers have used to improve K-12 school performance is “accountability.” Accountability in this sector is understood to mean a regime of regular standardized testing of students coupled with concrete consequences, including public reporting of school-level results and sanctions for schools and staff for failure to achieve test-score targets. Accountability in the form of high-stakes testing is now pervasive, even as its efficacy is hotly contested. The design of accountability regimes in schooling merits particular attention now, as Congress considers policy changes in re-authorizing the federal law that became, in its most-recent iteration, NCLB. We argue that the policy debate can be informed by an extensive literature from behavioral science on accountability (Lerner & Tetlock, 1999). That literature makes clear, first, that accountability comes in many forms that activate different mechanisms, and second, that accountability can produce positive or negative effects, depending on the accountability type, the decision context, and the nature of the task. We discuss the implications of the behavioral literature for schooling, where the mix of public and private purposes suggests the need for accountability to multiple constituencies, including public officials, parents, and students. We conclude that an effective accountability regime will involve (1) multiple forms of accountability; (2) multiple measures of educational practice and educational outcomes; and (3) feedback mechanisms to promote the improvement of practice. Moreover, a multi-pronged accountability approach should specifically increase the use of professional accountability, which has historically been underutilized in schools.

Contact: bgill@mathematica-mpr.com

Introduction

Since the passage of the No Child Left Behind Act (NCLB) over a decade ago, the primary lever that American policymakers have used to improve school performance is “accountability.” Building on previously passed laws in a handful of states, NCLB established a regime of public reporting of every public school’s level of student proficiency in math and reading as measured by standardized assessments, along with automatic sanctions for failure to achieve test-score targets. Schools that fall short of state-determined test-score goals—for all students or for particular subgroups of students—became subject to a series of mandatory interventions, progressively ratcheted up if student achievement failed to improve. Chronically failing schools could be re-staffed, taken over by the state, or shut down. In education policy, “accountability” came to be synonymous with the use of student tests for high-stakes purposes for schools.

The Obama Administration has been critical of some elements of NCLB, but its policies endorsed high-stakes testing and even expanded the scope of the stakes. The administration used the carrot of billions of dollars of funding made available in the American Recovery and Reinvestment Act to establish the Race to the Top (RTT) grant program. Recognizing that student achievement levels are affected by many factors outside the control of schools, RTT shifted the accountability measures from achievement *levels* to achievement *growth*, but it doubled down on using student test results for high stakes purposes, making not only schools but also individual teachers and principals accountable for student achievement growth. Subsequently, with efforts to reauthorize NCLB stalled in Congress, Secretary of Education Arne Duncan reasserted the administration’s commitment to high-stakes testing by offering states flexibility to waive some of the most onerous aspects of NCLB if they adopted several favored policies, notably including the incorporation of student achievement growth as an explicit component in the evaluation of teachers and principals. Over 40 states have applied for flexibility and are now beginning to include student test scores in educator evaluations. In short, accountability in the form of high stakes testing is now pervasive in American education policy.

As Congress now considers re-authorization of the Elementary and Secondary Education Act (the federal law that became known as NCLB in its most recent iteration), the efficacy of high-stakes testing is hotly contested. Calls for its refinement or abolition have come with increasing frequency from scholars, pundits, and stakeholders. An “opt out” movement has enlisted parents who refuse to have their children take the standardized tests. Even Secretary Duncan recently announced his concern about excessive testing. Scholars of education policy have begun to call for refining or replacing the current accountability regime (e.g., Tucker, 2014; Darling-Hammond, Wilhoit, and Pittenger, 2014; Center on Reinventing Public Education and Fordham Institute, 2014).

There is an extensive experimental literature in psychological science and behavioral economics concerning the effects of different forms of accountability, identifying the conditions under which specialized forms of accountability have the potential to improve outcomes, have no effect on outcomes, or make matters worse. The policy debate in education has been conducted almost entirely without reference to that literature, which might otherwise provide valuable

insights for the design of school accountability regimes.¹ Our aims in this paper are to harness insights from the behavioral-science literature on accountability to the consideration of accountability regimes in education policy, and—more specifically—to propose ways that policymakers might use a range of accountability tools to promote school performance.

We begin with Lerner and Tetlock’s (1999) comprehensive review of the behavioral literature, which makes two broad points that motivate much of our discussion. First, accountability comes in many forms, involving different mechanisms and (potentially) different behavioral responses. Outcome-based accountability (such as the high-stakes testing that is usually meant in discussions of accountability in education policy) is only one of these forms. Second, the effects of accountability on judgments and decisionmaking are not uniformly positive. Accountability can produce positive or negative effects (or both), depending on the accountability type, the decision context, and the nature of the task.

We also begin with the recognition that schooling has multiple purposes, both public and private, with multiple constituencies, including parents and students as well as the public. The classic principal-agent problem that accountability policies often seek to address is complicated by this fact. The K-12 accountability system needs to attend to these multiple purposes and multiple constituencies.

Reviewing the behavioral literature and relevant studies in education policy, we conclude that continuous improvement in schools would be best served by (1) employing multiple forms of accountability that (2) use multiple measures of student outcomes and multiple measures of professional practice and (3) incorporate feedback mechanisms to promote the improvement of practice. More specifically, policymakers should consider increasing the use of robust forms of professional accountability, which has historically been underutilized in schools.

Defining accountability

Lerner and Tetlock’s review of the behavioral literature identifies four mechanisms that make people feel accountable: the *mere presence of another*, involving the simple awareness that someone else is watching; *identifiability*, involving the expectation that an action or outcome will be personally attributable to the accountable party; *reason-giving*, involving the expectation that the accountable party will need to explain actions to another; and *evaluation*, involving the assessment of the accountable party’s performance by another “according to ... normative ground rules and with ... implied consequences” (1999, p.255).

The high-stakes testing regime inaugurated by NCLB and expanded by the Obama/Duncan initiatives incorporates the identifiability and evaluation components described in the behavioral literature. But the generality of the four mechanisms suggests many other ways to employ accountability in schooling. In this paper, we describe four kinds of accountability systems that are used in other policy contexts and are potentially relevant in schools, involving different

¹ A National Academies report (Hout and Elliot, 2011) examined some of the behavioral literature specifically related to the use of test-based incentives, but its scope did not include an examination of other forms of accountability.

elements of the four mechanisms identified by Lerner and Tetlock. Accountability approaches that are—or could be—used in schooling include:

- *Outcome-based accountability*, under which decisionmakers are responsible for achieving particular results, with positive or negative consequences following from achieving or failing to achieve targeted results. High-stakes testing is the prominent example of outcome-based accountability in education.
- *Rule-based accountability*, under which decisionmakers are required to act in accordance with rules or regulations that delineate mandated or forbidden activities.
- *Market-based accountability*, under which decisionmakers are responsible to clients or customers who implicitly hold them accountable by choosing among different providers.
- *Professional accountability*, under which decisionmakers are responsible for following standards of practice determined by the profession, and under which they have access to resources to help them improve their practice.

Most of these accountability mechanisms are used in other professions, which establish rules and standards for minimal performance, operate in markets that allow their clients and customers a choice of providers, and expect practitioners to use state-of-the-art practices. Professional accountability mechanisms in other fields also typically include substantial educational expectations for entry into the profession and resources for practitioners continuing learning. Outcome-based accountability is not prominent in most professions (with the notable exception of tort law, in which plaintiffs' attorneys are typically paid only if they win), perhaps because market-based accountability serves the same function. Outcome-based accountability is increasingly being attempted in health care, however (Gold, 2010), which resembles K-12 schooling in that consumers have limited market power.

Table 1 suggests how these accountability types employ one or more of the accountability mechanisms identified in the behavioral literature.

Table 1. Accountability types in schooling and psychological accountability mechanisms

		Accountability types in schooling			
		Outcome-based	Rule-based	Market-based	Professional
	Mere presence of another		✓		✓
Psychological accountability mechanisms	Identifiability	✓	✓	✓	✓
	Reason-giving			✓	✓
	Evaluation	✓	✓		✓

The accountability types differ not only in the psychological mechanisms they invoke, but also in the constituencies to which educators are accountable. Outcome-based and rule-based accountability implicitly involve the public officials who define the outcomes and establish the rules. Market-based accountability makes parents or students the relevant constituency. And

professional accountability directs educators' attention to colleagues and peers. All of these constituencies are potentially relevant in the design of an accountability system in education, because schooling involves professional practice in the service of both private and public goods. Table 2 shows the constituencies for the different accountability types.

Table 2. Accountability types and constituencies: Accountable to whom?

		Accountability types in schooling			
		Outcome-based	Rule-based	Market-based	Professional
Constituencies	Public officials	✓	✓		
	Parents			✓	✓
	Students			✓	✓
	Peers				✓

In the pages below, we describe applications and potential applications of each of these four accountability types in schools, illuminate how they encompass different psychological mechanisms, and consider relevant findings from the behavioral literature regarding the effects of particular accountability types.

Outcome-based accountability

Outcome-based accountability is relatively new in education policy, but its increasing prominence has made “accountability” synonymous with high-stakes testing. Prior to the 1990s, outcome-based accountability was almost unknown in K-12 schooling. The education standards movement that gained steam during the 1990s posited that schools should aim to bring their students to “proficiency” in reading and math, which required defining expectations at each grade level and measuring students’ success in reaching those expectations.

NCLB is the latest re-authorization of the Elementary and Secondary Education Act (ESEA). The preceding version of ESEA, passed in 1994, required states to set proficiency standards, assess their students in multiple grades, and publicly report school-by-school proficiency levels. Thus, the primary accountability mechanism implicit in the 1994 federal law was, in Lerner and Tetlock’s terms, *identifiability*: student outcomes for each individual school were publicly reported, but in most cases there were no explicit sanctions applied to the schools. Eight years later, NCLB raised the stakes, adding explicit sanctions to schools falling short of state-defined proficiency targets, thereby incorporating the evaluation mechanism alongside the *identifiability* mechanism.

Behavioral evidence on outcome-based accountability

The behavioral literature includes few examples in which outcome-based accountability has produced positive effects in randomized experiments. Outcome accountability has been hypothesized to motivate people to find creative new strategies when standard procedures are inadequate (Tetlock and Mellers, 2011; Wilson, 1989). Because it does not constrain decisionmakers with rules and standard processes, outcome accountability might be more effective than other forms of accountability at promoting innovation (Patil, Vieider, and Tetlock,

2012). But the existing empirical literature has not tested this possibility in any depth. One study comparing outcome accountability to process accountability found that process accountability produced better results (i.e., more accurate predictions) in situations in which results depended on simple additive functions of existing information that could be easily intuited by the decisionmakers, but that the advantage of process accountability over outcome accountability disappeared when the available information could not be combined in a simple linear way to produce an accurate prediction of outcomes (de Langhe, van Osselaer, and Wierenga, 2011). Patil, Vieider, and Tetlock (2012) hypothesize that the absence of experimental evidence on positive effects of outcome-based accountability may be partly due to the difficulty of adequately simulated complex real-world circumstances in laboratory settings.

Behavioral science provides a number of examples in which outcome-based accountability has unintended consequences that can undermine its effectiveness. In particular, it has been shown to exacerbate sunk-cost bias, making decisionmakers more likely to continue undertaking courses of action simply because of prior investments, even if they have a low probability of success (Simonson & Staw 1992). In addition, in many contexts (though not always), using tangible rewards as performance incentives has been shown to undermine intrinsic motivation (Deci, Koestner, and Ryan, 1999).

Outcome-based accountability may impair the decision-making process through the elicitation of stress and negative emotion (Siegel-Jacobs & Yates 1996). Outcome-based accountability increases the decision's difficulty and may amplify negative emotions (Zhang & Mittal 2005). Stress burdens the decision maker's cognitive load, leaving less working memory free for the complex and adaptive mental processes teaching demands (Mendl, 1999; Hockey, 1997; Hancock & Warm, 1989; Lazarus & Folkman, 1984). Perhaps for this reason, outcome accountability sometimes fails to increase the use of strategies that require substantial effort (Lerner & Tetlock 1999).

In addition, the behavioral literature shows that accountability regimes that are viewed as illegitimate can be counterproductive (though this finding is not specific to outcome-based accountability) (Lerner & Tetlock, 1999). In the current policy environment, for example, "value-added" measures (VAMs) that aim to quantify individual teachers' contributions to student achievement are viewed with suspicion by many teachers. "VAM is a sham," according to Randi Weingarten, president of the American Federation of Teachers. VAM improves on cruder outcome-based accountability regimes that rely on student achievement *levels* and create an implicit bias against educators serving disadvantaged students (who are likely to have lower achievement even if their schools are substantially improving their achievement). VAM seeks to eliminate this bias by accounting for factors outside the teacher's control. VAM may in fact remove most of the bias (Kane et al., 2013; Kane and Staiger, 2008; Chetty, Friedman, and Rockoff, 2013), but even if it is methodologically sound, perceptions of its flaws may undermine its ability to promote improvements in performance.

Moreover, even unbiased measures of educators' contributions to student achievement can be problematic in a high-stakes accountability regime because the underlying achievement measures are imperfect. Achievement tests never fully capture all of the skills and knowledge that society expects schools to impart and students to learn; recent evidence suggests that instructional practices that are effective in raising test scores are not the same as those that

promote students' effort and long-term goals (Ferguson and Danielson, 2014). Using the tests for accountability encourages schools to focus on the elements of learning that are tested to the detriment of elements that are not tested, as demonstrated in other contexts by the psychology literature on conformity and the availability heuristic (Tetlock, 1983; Tetlock, Skitka, & Boettger, 1989). This phenomenon illustrates Campbell's law, which notes that attaching stakes to any social indicator can undermine the validity of the indicator as the accountable actors seek ways to "game" the measure (Campbell, 1976).

Evidence from the field on outcome-based accountability

Whether these various limitations of outcome-based accountability are outweighed by potential benefits in an empirical question. In fact, the effects of NCLB in particular and high-stakes testing in general are a matter of heated debate. Most of the existing studies suggest positive effects in at least some grades and subjects (Dee & Jacob, 2011; Carnoy & Loeb, 2002; Hanushek & Raymond, 2005; Ahn & Vigdor, 2014; Deming et al., 2013 finds mixed results). Few studies have yet examined the empirical effects of the recent expansion of high-stakes testing to individual educators. An exception is a recent paper by Dee & Wyckoff (2015) examining the effects of the teacher evaluation system that the public schools of Washington, DC began implementing several years ago; they find evidence of favorable effects for low-rated and high-rated teachers alike. A few studies have examined the effects of performance-pay incentives for teachers on student achievement, finding mixed results (Springer et al., 2012a and 2012b; others cited in Springer; Glazerman & Seifullah 2012).

Meanwhile, since the passage of NCLB, there is evidence that many schools have narrowed the curriculum (Dee, Jacob, & Schwartz, 2013) and spent increasing time preparing for state assessments (Hamilton, Stecher, & Yuan, 2012; Jennings and Sohn, 2014). Scores on low-stakes assessments have improved more slowly than scores on high-stakes assessments, and sometimes not at all (Koretz & Barron, 1998; Jacob, 2005; Jennings and Sohn, 2014; Holcombe, Jennings, and Koretz, 2013). In extreme cases, educators have been caught cheating on high-stakes assessments. Corruption of the measure is a particular challenge with teacher-developed "student learning objectives" (SLOs), which are increasingly used as outcome-based accountability measures for teachers of students in non-tested grades (Gill et al., 2014). SLOs essentially ask teachers to grade themselves, creating an inherent conflict of interest. Not surprisingly, a recent paper by Balch & Springer (2015) found no correlation between teachers' success in achieving SLOs and their value-added as measured by standardized assessments.

In sum, the evidence on outcome-based accountability is mixed: it can produce both favorable and unfavorable results. To promote more favorable effects of outcome-based accountability, we recommend application of the following measurement guidelines, working together as an ecological system (see Hernandez, Hodges, and Cascardi, 1998). Recognizing that measurement is costly and has the inevitable effect of altering behavior, an outcome-based accountability regime should:

- Include measures selected based on input from key stakeholders coupled with empirical evidence that they predict long-term student outcomes within and/or beyond school.
- Incorporate multiple, complementary measures to better capture the broad aims of schooling and to reduce the ability to artificially inflate scores;

- Provide feedback that is made constructive by coupling it with resources and tools for the improvement of practice;
- Be regularly updated by ongoing behavioral science on best measurement practices.

Rule-based accountability

Historically, schools have relied on rule-based accountability for defining some of the structural conditions of schooling and setting minimal standards for teaching (Tyack, 1974). Rule-based accountability has included traditional teacher evaluations, in which a principal observes a teacher once or twice a year and issues a judgment about whether the teacher's performance is satisfactory—where 99 percent or more of teachers are typically judged satisfactory (Weisberg et al. 2009). It includes state-level decisions about allowable textbooks. It includes contractual rules about working conditions, hours, and class sizes. It includes federal and state regulations about how money is spent. Apart from these kinds of constraints, however, teachers have typically had wide discretion about instructional decisions in their classrooms (Tyack and Cuban, 1995). This was the dominant approach in American public schools for much of the twentieth century. It allowed enormous variation in school and teacher quality, provided little opportunity for teachers to scrutinize and improve their practice, and lacked a mechanism for evaluating and disseminating effective practices. Over time, the accumulation of rules and protocols may have raised minimum standards and reduced the most egregious inequities, but it also reduced opportunities for innovation and may have sacrificed efficiency for regularity (see Tyack, 1974; Chubb & Moe, 1990; Wilson, 1989).

Recently, some states, districts, and school management organizations have become more directive about elements of instruction and school operations that had previously been left largely to the discretion of principals and teachers, pursuing a “maximal” version of rule-based accountability. Some districts and school management organizations now expect all teachers of the same courses are expected to cover the same material at the same pace (see Snipes et al., 2002). Principals have been asked to facilitate curriculum management by taking on greater responsibilities as instructional leaders rather than just building managers. Pacing guides are commonly used, and instruction is expected to be explicitly tied to state standards. In some cases, lesson plans are scripted down to the minute (Beatty, 2011). Rule-based accountability regimes implicitly rely on the mechanism of identifiability. When teaching is monitored to ensure compliance, the presence of another is also used as an accountability mechanism.

There is limited evidence in support of a maximal rule-based approach inside schools. Scripted “direct instruction” has been found to promote student achievement in elementary grades (Borman et al., 2003). But it seems far less likely to work at the high school level, in courses designed to do more than improve basic skills. Maximal rule-based accountability could easily become counterproductive: Individuals who are closely monitored to control their performance often lose their intrinsic motivation in response to their perceived loss of autonomy (Enzle & Anderson, 1993). Intense procedural scrutiny can also exacerbate sunk-cost bias (Ross & Staw, 1993). Strict adherence to rules can undermine innovation and entrench practices that may not be optimal (Patil, Vieider, & Tetlock, 2012). The perception of rules as illegitimate can produce a boomerang effect (Baer, Hinkle, Smith, & Fenton, 1980; Brehm, 1966; Heilman & Toffler, 1976), just as the perception of illegitimacy can undermine an outcome-based accountability regime.

Indeed, aggressive rule-based accountability may be especially unsuited to schooling because teaching is an inherently complex task that requires daily adjustments and judgments to do well (Duffy et al., 1986; Clark & Peterson, 1986; O’Keefe & Johnston, 1989; Borko & Shavelson, 1990). State and federal officials, in particular, are very far removed from classrooms. As Rick Hess (2013) notes, a bureaucrat can mandate implementation, but cannot mandate effective implementation.

Market-based accountability

Market-based accountability is based on principles from traditional neoclassical economics rather than newer behavioral approaches. As the prominent behavioral economists George Loewenstein and Peter Ubel (2010) have noted, even in light of the wealth of new knowledge about how human judgments can deviate from rational self-interest, neoclassical economic principles remain valuable guides to public policy in many contexts.

Historically, market-based accountability did not play a substantial role in policies related to public education in the United States, despite being the primary accountability system in most other sectors of the economy. Tuition-based private schools have always existed alongside the public education system, but the public system has been based on the “common school” model, which assumes that each community will educate its children together in the same schools, operated by public officials in school districts that maintain local monopolies on the provision of publicly supported education (Tyack & Hansot, 1982; Glenn, 1988).

In the last few decades, policymakers have shown increasing interest in incorporating market-based accountability into public policies related to education, motivated partly by the belief that local monopolies controlled by school boards may not be the most effective way to provide schooling (Friedman, 1955; Chubb & Moe, 1990), and partly by the view that giving families choices in schooling is inherently valuable (Coons, 1992). Market-based accountability is accountability to the families that are implicitly the clients of the educators.

The most prominent examples of market-based K-12 education reform are charter schools—publicly funded schools of choice that are autonomously operated, outside the direct control of local school districts (Gill et al., 2007). The first charter schools opened only two decades ago, but today over 6000 are operating in more than 40 states, representing about six percent of all schools across the country (NAPCS) and much higher percentages in cities such as New Orleans, Washington DC, and Kansas City. Meanwhile, a smaller number of states have adopted “voucher” programs that permit some students to attend private schools at public expense.

Implicitly, market-based accountability involves the identifiability and reason-giving mechanisms. Schools that enroll their students through the choices of their families must be identifiable, and unlike schools that are enrolled via residential assignment, they must offer reasons for parents to enroll their children. The reasons that schools of choice provide to attract students are widely varied, from general academic quality to excellence in a particular subject area to personalized attention to mundane issues such as safety and location.

Empirically, the evidence on the effects of market-based schools on students’ test scores and longer-term educational attainment suggests promise but is not definitive. The average effect of charter schools—by far the most prominent form of publicly supported market-based schools—

remains a matter of some debate. In some contexts and locations, charter schools are producing substantial positive effects on their students (Abdulkadiroglu et al., 2011; Furgeson et al., 2012; Booker et al., 2011; Tuttle et al., 2013). But the variation in the performance of charter schools is wide (Gill & Nichols-Barrer, 2014; Center for Research on Educational Outcomes, 2013). Meanwhile, a few studies of small-scale voucher programs have found positive educational impacts, particularly for African-American students (Wolf et al., 2013; Chingos & Peterson, 2015).

Nonetheless, K-12 schooling differs from many other services in ways that might make exclusive reliance on the market less than optimal. First, the classic principal-agent² problem—aligning the interests of agents (educators) and clients—is complicated by the involvement of multiple clients (students, parents, and the public) whose interests may not be 100 percent aligned. Moreover, the first set of clients—the students—are under-age and presumed to be not fully competent to know their own interests. These are circumstances that can produce market failures.

The likely inadequacy of markets as complete solutions for accountability is also suggested by the fact that schooling creates externalities. Students' educational experiences are affected not only by the quality of the inputs from the school, but also by the characteristics of other students (Betts & Morell, 1999; Sacerdote, 2001; Linkenbach & Perkins, 2003; Kremer & Levy, 2008). Moreover, many parents perceive the peer environment of the school to be very important. In consequence, an unfettered market in schooling may produce segregated schools, as parents with high levels of knowledge, wealth, or motivation seek out schools with other families like their own (Bifulco & Ladd, 2006; Booker et al., 2008).

In addition, students in conventional public schools may be affected by the existence of the market-based schools. Whether the externalities are positive or negative is a matter of heated debate. Skeptics worry that market-based schools will drain conventional public schools of funding and motivated families, potentially harming the students left behind. Supporters argue that breaking the local monopoly on public schooling produces healthy competitive pressure that will benefit all students. It is much harder to measure the systemic effects of market-based schools than to measure their direct effects on their own students. Scholars have attempted to measure systemic effects using a variety of approaches. Findings are far short of definitive, but they provide some reason for optimism: although several studies have found neither positive or negative effects of charter schools on achievement in nearby conventional schools (Bettinger, 2005; Bifulco & Ladd, 2006; Zimmer et al., 2009), several others have found evidence of small positive effects (Jinnai, 2013; Zimmer et al., 2009; Winters, 2012), and only one has found evidence of a negative effect (Imberman, 2011).

Another externality relates to the original rationale for public education: society as a whole benefits from the inculcation of the knowledge, skills, and attitudes necessary for effective citizenship in a democracy. Historically, this formed a key rationale for the common-school model, which implicitly assumed that effective education for citizenship required not only public funding for education but also public operation of the schools (Tyack & Hansot, 1982; Glenn, 1988). Almost no empirical evidence exists on whether market-based schools can effectively

² We use the term “client” rather than “principal” to avoid confusing principal-clients with principal-school leaders.

serve the original public purpose of education (Gill et al., 2007). The extent to which market-based schools successfully educate for citizenship is largely unstudied (though the existence of charter-school organizations such as Democracy Prep is an encouraging anecdotal indicator; see Lake & Miller, 2012). To be sure, empirical evidence on the effectiveness of conventional public schools in serving this goal is equally sparse; but the fact that the education of citizens is a public good argues against relying entirely on market mechanisms to serve this purpose.

Indeed, in recognition of the public purposes of schooling, existing school choice programs rarely rely exclusively on market accountability. Charter schools are exempt from some forms of regulation, but like conventional public schools, they are subject to a substantial amount of rule-based and outcome-based accountability: They are not permitted to charge tuition, they are required to submit their students to the same high-stakes tests taken in conventional public schools, and they are (typically) required to admit all applicants as long as they have space available. Moreover, charter schools require the approval of a publicly empowered authorizer to open, and they can be closed by the authorizer if they fall short of expectations. Even the private schools that participate in voucher programs typically are required to submit to some public regulation if they are to receive public funds. Milwaukee's program, the longest-operating publicly funded voucher program in the country, imposes requirements for instructional time, forbids the charging of tuition, requires students to take state assessments, and does not allow schools to impose admissions criteria.

Professional accountability

Prominent voices in education research are calling for an increased reliance on professional accountability, under which teachers would have more support, more collaboration, more training, and higher expectations (Tucker, 2014; Darling-Hammond et al., 2014). Professional accountability can take many forms, involving all four of the accountability mechanisms identified by Lerner and Tetlock. Licensing, standards, and reviews by organizations representing the teaching profession involve evaluation; observations and assistance by supervisors, instructional coaches, peers, or mentors involve identifiability, reason-giving, and sometimes evaluation; collaboration and co-teaching involve the presence of another and reason-giving.

It is worth noting that professional accountability is not synonymous with professionalism. In many fields, professionalism implies an ethic of meeting professional standards even when no one is watching and there are no consequences. Professionalism may be an important motivator of effective practice, but we do not consider it (in isolation) a form of accountability, which by definition involves an external observer. Professional accountability involves some form of transparency: lawyers write briefs and contracts that are scrutinized by other lawyers; surgeons practice their craft in operating rooms full of other professionals; architects, engineers, and programmers develop reputations based on their finished products.

Traditional and novel versions of professional accountability in education

Standards and requirements for acquiring a license to teach have long been applied by states. These typically include requirements for coursework, student teaching, and passing exams. Retaining a license may require minimal ongoing professional development. Teacher contracts generally reward master's degrees and years of teaching experience as proxies for professional

skill. Research has repeatedly found, however, that these characteristics explain only a small part of the variation in teacher effectiveness (e.g. Kane, Rockoff, and Staiger, 2008). Master's degrees have little or no relationship to teachers' ability to improve their students' achievement (Chingos & Peterson, 2011; Clotfelter, Ladd, & Vigdor, 2007). Teachers improve in their first few years of teaching, on average, but they often plateau early (Chingos & Peterson, 2011; Rice, 2013). And most studies of professional development find no effect on student achievement (e.g., Garet, Porter, Desimone, Birman, & Yoon, 2001; Hawley & Valli, 1999). Licensing and professional requirements at a high enough level might be useful in identifying especially effective teachers. Teachers completing the rigorous certification process of the National Board for Professional Teaching Standards (NBPTS) may be more effective than average (Cowan & Goldhaber, 2015; Cavalluzzo, Barrow, & Henderson, 2015; Clotfelter, Ladd, & Vigdor, 2007).

Forms of professional accountability that are more robust and ambitious than traditional licensing requirements may have more promise for improving instructional performance. Although most of the variation in teacher quality is not explained by licenses, educational credentials, or experience, the variation is nonetheless large: Teacher quality is the most influential school-controlled factor in student achievement growth (Rivkin et al., 2005). The fact that the variation is large suggests that it should be possible to identify and support the teaching practices that best promote student achievement. Indeed, recognition of the variation in teacher effectiveness provides much of the motivation for the Obama Administration's aggressive promotion of more rigorous teacher evaluations. In response, states and school districts are adopting new and extensive rubrics for the observation and rating of teaching practice, and creating additional distinctions in performance ratings beyond a simple satisfactory/unsatisfactory judgment. Teaching practice is sometimes rated on 20 or more dimensions, often including criteria for teacher-student responsiveness and classroom environment (as with Charlotte Danielson's Framework for Teaching, perhaps the most widely used system for measuring professional practice).

If not implemented well, the ambitious new systems for observing and evaluating teachers' professional practice could deteriorate into compliance exercises that resemble traditional rule-based accountability mechanisms. But if they function as intended, as rich measures of practice that provide guidance for the improvement of teaching, they may have the potential to promote substantial improvement in practice. Robust professional accountability systems—unlike outcome-based, rule-based, and market-based accountability—include tools and resources to help teachers improve their practice. One of the complaints about high-stakes testing is that it provides no information to educators about how to do better. The same criticism can be made of market-based accountability approaches, which likewise provide no direct feedback relevant to the improvement of practice (perhaps explaining the wide variation in the performance of charter schools). As Lerner and Tetlock found in their review of the behavioral literature, neither process nor outcome accountability can, on their own, eliminate errors that derive from lack of knowledge. Professional accountability systems—if taken beyond screening and compensation reform—have the unique advantage of coupling accountability with support for improvement.

Novel forms of professional accountability might include restructuring traditional job descriptions and training methods. Some school districts have recently created teacher residency programs modeled on medical residencies, in which aspiring teachers spend much more time in the classroom than traditional teacher education programs provide (Hallberg and Green, 2015).

Other districts and schools are finding ways to give teachers leadership opportunities in “hybrid” roles that exploit their knowledge in developing the skills of their colleagues (e.g., Barnwell, 2015). Borrowing wisdom from other fields (see Gawande, 2011), instructional coaching has become a prominent professional development strategy in school districts and charter schools, giving expert teachers opportunities to aid their colleagues and giving classroom teachers opportunities to improve their skills.

In most professions, professional accountability includes answerability to clients (Newton, Hodges, and Keith, 2013). K-12 schooling traditionally involves little direct accountability of educators to students, but in higher education, student evaluations of teachers have been commonly used for professional accountability. School districts such as the Pittsburgh Public Schools are now including student surveys in new teacher evaluation systems that incorporate multiple measures of teacher performance (Chaplin, Gill, Thompkins, and Miller, 2014). Meanwhile, some schools are using surveys of teachers as part of the process of principal evaluation, following the example of the “360” evaluations that have become prominent in the business world. A teacher survey is a core component of one of the most popular new evaluation rubrics for principals, the Vanderbilt Assessment of Leadership in Education (VAL-ED) (Porter et al., 2010).

Another professional accountability system is an intensive review of school quality conducted by independent, expert educators, like the reviews that have long been conducted in British schools. As advocated by its supporters (e.g. Tucker, 2014; Darling-Hammond et al., 2014), a school quality review would involve an extended school visit by outside experts, who might observe instruction, interview teachers, students, and parents, and closely examine data on school performance. Most importantly, they would conclude their review not just with a judgment of the school’s performance, but also with clear recommendations for improvement.

Coaching, peer observation, collaboration, student surveys, and school quality reviews have in common an element of making teaching more transparent (activating all four accountability mechanisms: presence of another, identifiability, reason-giving, and evaluation). This is a departure from the traditional organization of American schools, in which most of teachers’ work occurs behind closed doors. Indeed, these kinds of rich professional accountability systems emphatically reject the notion that professionalism is equivalent to allowing individual teachers complete discretion to practice as they choose. Rules such as those found in the Chicago Public Schools (2014), explicitly prohibiting the use of classroom video recordings for evaluation purposes, are anathema to this vision of professional accountability. Meanwhile, schools like the Kauffman Charter School in Kansas City are literally making teaching transparent by giving classrooms interior windows that make them visible to adults in the hallways.

Greater transparency in teaching is consistent with educational practice in some countries that consistently outperform the U.S. in international comparisons of student achievement. A recent international study of educational practice (OECD, 2014) found that although American teachers are more likely than average to receive feedback from principals, they are much less likely to receive feedback from mentors or other teachers. Only 11 percent of responding American teachers received feedback from mentors, versus 39 percent in Japan, 38 percent in Singapore, and 24 percent in Australia. Only 27 percent of responding American teachers received feedback from colleagues, versus 84 percent in Korea, 57 percent in the Netherlands,

and 43 percent in Finland. All of those countries significantly outperformed the U.S. in all three subjects (math, reading, and science) in the most recent study of the Program for International Assessment (PISA). American teachers were also far less likely than their counterparts in other countries to receive feedback from student surveys. And they were less likely to report that the feedback they received on their teaching led to public recognition, career advancement, or increased compensation.

Behavioral evidence related to professional accountability

The behavioral literature includes a wide range of findings potentially relevant to forms of professional accountability that are richer and more extensive than traditional licensing. Specifically, the literature includes many randomized, experimental studies demonstrating the value of a requirement to justify a decision to others—a common expectation of professional accountability regimes. Ashton (1992), for example, found that requiring subjects to justify their decisions encouraged the use of high-effort strategies that were sensitive to available evidence, thereby alleviating mistakes and inconsistencies. Similarly, Lerner et al. (1998) found that the need to justify decisions stimulated systematic thinking and attention to evidence. Bodenhausen, Kramer, and Susser (1994) demonstrated that a justification requirement reduced reliance on stereotypic thinking. Tetlock (1985) found that subjects who had to justify their judgments were less likely to be led astray by the fundamental attribution error (i.e., the tendency to over-attribute responsibility to individuals and ignore the importance of situations). Siegel-Jacobs and Yates (1996) found that subjects who were accountable for the quality of their judgment process made more accurate judgments.

In group contexts, accountability for process has been found to increase the group's demand for information, induce more sharing of information, and produce better decision outcomes (Scholten, van Knippenberg, Nijstad, and De Dreu, 2007).

Professional accountability has the potential to be structurally more inclusive of employees and to allow flexibility to react to evolving situations (Sutton & Galunic, 1996). In encouraging educators to take initiative, professional accountability may give them greater sense of control, which has been shown to improve performance on attention-demanding tasks, promote more considerate decisionmaking, and assist memory formation (Sherman et al., 2012; Elovania et al., 2006; Nabi et al., 2013; Hancock, 1989; Mendl, 1999).

More generally, professional accountability may best provide the systematic, effortful, and self-critical thinking that is associated with even-handed reasoning and increased accuracy in judgment and choice using systematic rather than heuristic processing (as described in Kahneman, 2011). Professional accountability might offer flexibility for innovation while disallowing the acceptability heuristic—the adoption of the cognitively lazy but easily defensible decision preferred by the audience (Tetlock, 1991).

Professional accountability also finds some support from political science. In the context of a discussion of different approaches to accountability in political representation, Mansbridge (2009) argues that a “selection model” is superior to a sanctions-based approach to the principal-agent problem when the aims and motivations of principal (client) and agent are well aligned. Although Mansbridge is writing about representative democracy, the general point is also relevant in the schooling context, where educators' primary aims are consonant with the aims of

policymakers, parents, and students. When goals are aligned, the selection model allows greater freedom of action by the agent (the professional educator), reducing the need for, and potential harms of, detailed rules or extrinsic motivators. Similar arguments have been made about accountability in business (Pratt and Zeckhauser 1985). As implied by the term, the selection model depends on “accurate selection and sorting as well as the capacity to ‘de-select’” (Mansbridge, 2009, p. 381)—suggesting, in the education context, the need for transparent professional standards, rigorous evaluation, and perhaps a complementary element of market-based accountability.

Evidence on effects of professional accountability in schooling

As yet, little empirical evidence is available on the types of job-embedded professional accountability that involve day-to-day interactions of teachers in schools and classrooms, perhaps because they were not frequently the subject of policy interventions in the past. That is changing. Teacher residencies are showing promise in producing high-performing teachers and keeping them in the classroom (Hallberg and Green, 2015). Early evidence on the effects of instructional coaching on student achievement is encouraging (Furgeson et al., 2012; Marsh, McCombs, & Martorell, 2010; also Lockwood, McCombs, and Marsh, 2010). Some of the highest-performing charter-school networks make intensive instructional coaching an integral part of their educational approaches (Lake et al., 2012). Two randomized experimental studies have found substantial effects of instructional coaching on teaching practice (Neuman and Cunningham, 2009; Knight and Cornett, 2009).

Meanwhile, several of the recently-developed rubrics for observing and evaluating instructional practice have been shown to produce scores that are significantly correlated with teachers’ value-added (Kane, 2012; Chaplin et al., 2014; Walsh & Lipscomb, 2013), suggesting that careful observation has the potential to produce feedback that can lead to improved teaching and improved student outcomes.

In higher education, student evaluations of teachers have been found to have some positive effects on teaching (L’Hommedieu, Menges, & Brinko, 1990). Recent studies examining student surveys in K-12 schools have found that student survey results are (modestly) correlated with measures of teachers’ contributions to student achievement (Kane, 2012; Chaplin et al., 2014). Whether student surveys can produce improvements in K-12 student achievement is as yet unknown.

Promising evidence on job-embedded professional accountability used formally comes from a recent study of an intensive, peer-based teacher evaluation system used in Cincinnati. A rigorous quasi-experimental study of the peer evaluation program found that teachers substantially increased their effectiveness in raising student achievement during and after the year in which they were evaluated (Taylor & Tyler, 2012). Interestingly, the evaluation criteria were based entirely on professional practice, not on student achievement results—the intervention did not include a high-stakes testing (outcome-based accountability) component.

Conclusion: Increasing professional accountability in a multi-mode approach

The preceding review makes clear that the usual usage of the term “accountability” in K-12 education policy reflects an unnecessarily narrow understanding of accountability mechanisms.

The *outcome-based accountability* that is implicit in high-stakes testing is only one type of accountability. And although there is evidence that outcome-based accountability can produce positive results in schooling, relying on it exclusively is likely to lead to a variety of unintended and undesirable consequences.

Rule-based accountability, long included in American education policy, has surely been useful in setting minimal standards and expectations, but it has not been designed to promote high performance. And the effectiveness of more-aggressive rule-based approaches is likely to decline with the decision maker's increasing distance from the classroom. Detailed mandates from distant officials are especially susceptible to being perceived by educators as illegitimate.

Market-based accountability likewise can play a productive role in improving school performance, but even most supporters of school choice do not advocate an educational market that is unconstrained by other forms of accountability. Given the effects of student sorting and the public purposes of schooling, rule-based modifications to market-based accountability make good sense.

Professional accountability is the most under-utilized form of accountability in K-12 schooling in America. Barriers to entering the profession are low. Evaluation standards have historically been very low. Compensation and career advancement have not been based on any measure of performance. And there has been little expectation that teachers will continually improve their practice, or even that their practice will be regularly observed by peers. This is changing, with a variety of initiatives related to rigorous educator evaluation and job-embedded professional development (coaching) now underway.

We propose that the evidence from the behavioral-science and field literature on accountability suggests that accountability systems to improve the performance of K-12 schools should include three key elements:

1. Multiple forms of accountability
2. Multiple measures of educational practice and educational outcomes
3. Mechanisms to promote the improvement of practice

First, the diversity of advantages and disadvantages of different forms of accountability suggests that *multiple forms of accountability* might be usefully employed in complementary ways. In a recent paper examining the evidence on process and outcome accountability, Patil, Vieider, and Tetlock (2012) point out that outcome-based accountability may better promote innovation while process-based accountability (including forms of professional accountability) better promotes use of identified best practices. They further suggest that the disadvantages of both process and outcome accountability might be counteracted by systems that promote the “empowerment” of decisionmakers: “agents who feel empowered under process accountability are likely to resist conformity to deficient standard practices as well as attend to outcomes whereas those who feel empowered under outcome accountability are likely to attend to processes as well as outcomes, thereby facilitating organizational learning” (p.79). Empowerment is implicit in professional accountability, but it can also be incorporated in outcome accountability, they argue, when the outcome accountability regime communicates a

desire to achieve shared objectives, achieving legitimacy. Similarly, a National Academies report (Hout and Elliott, 2011) argues that external rewards are most likely to be effective when they are well aligned with educators' intrinsic aims, thereby promoting "autonomous motivation." Conversely, an accountability system that is viewed as illegitimate is unlikely to produce positive effects regardless of whether it is based on outcomes or processes.

The Equity Project (TEP) Charter School in New York City provides a compelling example of using multiple accountability approaches in tandem to produce positive effects (Furgeson, McCullough, Wolfendale, and Gill, 2014). The school achieved a measure of fame before it even opened, when it announced its plan to pay its teachers \$125,000 plus bonuses that are based on schoolwide improvements in student achievement. Although it was TEP's compensation package that made front-page news, important components of its educational approach go beyond compensation to include intensive screening, collaboration, observation, and professional development. Aided by a large pool of applicants, TEP imposes a rigorous screening process in hiring teachers, and avoids hiring novices (unlike many charter schools). TEP teachers participate in a six-week-long, collaborative professional development session each summer. TEP teachers are expected to observe each other in the classroom at least twice a week, providing written feedback to their colleagues. In short, TEP's model includes professional accountability alongside market accountability (it is a school of choice) and a substantial dose of outcome accountability. It embodies Rick Hess' (2015) vision that "what works" in school reform includes "transparency," "empowering talented professionals," and "rewarding excellence." And it has worked: in its first four years of operation, TEP has produced substantial positive effects on student achievement (Furgeson et al., 2014).

Second, accountability regimes should incorporate *multiple measures of educational practice and educational outcomes*. As we described earlier, one of the weaknesses of outcome-based accountability is that there is no single student outcome measure that fully captures all of the skills, knowledge, and attitudes that students should be learning in school. Incorporating multiple measures of student outcomes makes it less likely that schools will strategically focus on a single outcome while neglecting others that are also important; and it may mitigate the Campbell's law problem, since it is harder to "game" many outcomes than a single outcome. The U.S. Department of Education's guidelines for granting flexibility from elements of NCLB points in this direction by allowing states to use a broader range of measures for school accountability than have been employed in the past. California's CORE districts, for example, have taken advantage of this flexibility to include in their accountability regime not only test scores, but also attainment measures (such as high school graduation), and social-emotional measures (such as absenteeism, suspensions, and measures of non-cognitive skills).³

Including multiple measures of practice in the professional component of an accountability regime is likewise important, for some of the same reasons. As Ron Ferguson and Charlotte Danielson (2014) have found, instructional practices associated with growth in students' test scores are not identical to instructional practices associated with some favorable social-emotional student outcomes. Moreover, multiple measures of practice permit the accountability regime to be responsive to multiple constituencies. Student and parent perspectives on teachers provide additional accountability that can complement peer and supervisor perspectives. Similarly,

³ See <http://www2.ed.gov/policy/eseaflex/approved-requests/corerequestfullredacted.pdf>

evaluations of school principals are incomplete without systematic input from teachers like that provided by the VAL-ED principal evaluation instrument.

Third, employing multiple forms of accountability and multiple measures of practice and outcomes helps to create a complete system that provides *mechanisms to promote the improvement of practice*. Transparency of practice—permitting feedback from colleagues, coaches, supervisors, parents, and students—creates the opportunity for educators to learn how to get better. Rich data on student outcomes creates the opportunity to diagnose needs of students. And rewards for success create the opportunity for educators to innovate in productive ways.

In sum, the findings of behavioral science and the record from the field together suggest that the optimal approach to improve educational effectiveness will use each type of accountability—ideally in a package that empowers educators in the service of achieving outcomes viewed as legitimate. Given the existing structures of American public schools, achieving that optimal approach will require shifting the balance toward a larger role for professional accountability.

References

- Abdulkadiroglu, 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. *Quarterly Journal of Economics*, *126*(2), 699–748.
- Ahn, T., & Vigdor, J. (2014). The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina (Working Paper no. 20511). Cambridge, MA: National Bureau of Economic Research.
- Balch, R., & Springer, M. (2015). Performance pay, test scores, and student learning objectives. *Economics of Education Review*, *44*, 114–125.
- Barnwell, P. (2015, February 18). Why schools need more “hybrid” teaching roles. *Education Week*.
- Beatty, B. (2011). The dilemma of scripted instruction: Comparing teacher autonomy, fidelity, and resistance in the Froebelian Kindergarten, Montessori, Direct Instruction, and Success for All. *Teachers' College Record*, *113*(3), 395–430.
- Bettinger, E. P. (2005). The effect of charter schools on charter students and public schools. *Economics of Education Review*, *24*, 133–147.
- Betts, J., & Morell, D. (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of Human Resources*, *34*(2), 268–293.
- Bifulco, R., & Ladd, H. (2006). The impacts of charter schools on student achievement: Evidence from North Carolina. *Education Finance and Policy*, *1*(1), 50–90.
- Bodenhausen, G. V., Kramer, G. P., & Susser, K. (1994). Happiness and stereotypic thinking in social judgment. *Journal of Personality and Social Psychology*, *66*(4), 621–632.
- Booker, K., Gilpatric, S., Gronberg, T., & Jansen, D. (2008). The effect of charter schools on traditional public school students in Texas: Are children who stay behind left behind? *Journal of Urban Economics*, *64*, 123–145.
- Booker, K., Sass, T. R., Gill, B., & Zimmer, R. (2011). The effects of charter high schools on educational attainment. *Journal of Labor Economics*, *29*, 377–415.
- Borko, H., & Shavelson, R. J. (1990). Teacher decision making. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 311–346). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230.
- Campbell, D. T. (1976). Assessing the impact of planned social change (Occasional paper no. 8). Hanover, NH: Dartmouth College Public Affairs Center.

- Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4): 305–331.
- Center on Reinventing Public Education. (2014). Designing the next generation of state education accountability systems. Retrieved from www.crpe.org
- Center for Research on Education Outcomes (CREDO). (2013). *National charter school study*. Stanford, CA: Stanford University.
- Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value added: Multiple measures of teacher effectiveness in the Pittsburgh public schools* (REL 2014-024). Washington, DC: U.S. Department of Education, Mid-Atlantic Regional Educational Laboratory.
- Chicago Public Schools. (2014). *REACH Students Educator Evaluation Handbook 2014-15*.
- Chingos, M. M., & Peterson, P. E. (2010). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review*, 30, 449–465.
- Chingos, M. M., & Peterson, P. E. (2015). Experimentally estimated impacts of school vouchers on college enrollment and degree attainment. *Journal of Public Economics*, 122, 1–12.
- Chubb, J., & Moe, T. (1990). *Politics, markets, and America's schools*. Washington, DC: Brookings Institution.
- Clark, C., & Peterson, P. (1986). Teachers' thought processes. In M. Wittrock (Ed.), *Handbook of research on teaching*. New York, NY: Macmillan.
- Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement?* (Working paper no. 12828). Cambridge, MA: National Bureau of Economic Research.
- Coons, J. E. (1992). School choice as simple justice. *First Things*, 22, 193–200.
- Cornett, J., & Knight, J. (2009). Research on coaching. In J. Knight (Ed.), *Coaching: Approaches and perspectives*. London: Sage Publications.
- Cvetkovich, G. (1978). Cognitive accommodation, language, and social responsibility. *Social Psychology*, 41, 149–155.
- Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives*, 22(86).
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–688.

- Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30(3), 418–446.
- Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effect of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, 35(2), 252–279.
- Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2), 267–297.
- De Langhe, B., van Osselaer, S. M. J., and Wierenga, B. (2011). The effects of process and outcome accountability on judgment process and performance. *Organizational Behavior and Human Decision Processes*, 115, 238–252.
- Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2013). *School accountability, postsecondary attainment, and earnings* (Working paper no. 19444). Cambridge, MA: National Bureau of Economic Research.
- Duffy, G., & Roebler, L. (1986). Constraints on teacher change. *Journal of Teacher Education*, 37(1), 55–58.
- Dusek, J., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology*, 75(3), 327–46.
- Elovainio, M., Ferrie, J. E., Singh-Manoux, A., Gimeno, D., De Vogli, R., Shipley, M. J., ... Kivimaki, M. (2009). Cumulative exposure to high-strain and active jobs as predictors of cognitive function: The Whitehall II study. *Occupational and environmental medicine*, 66(1), 32–37.
- Ferguson, R. F., with Danielson, C. (2014). How Framework for Teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. Hoboken, NJ: Jossey-Bass Publishers.
- Fogarty, J., Wang, M., & Creek, R. (1983). A descriptive study of experienced and novice teachers' interactive instructional thoughts and actions. *The Journal of Educational Research*, 77(1), 22–32.
- Friedman, M. (1955). The role of government in education. In R. A. Solo (Ed.), *Economics and the Public Interest*. Piscataway, NJ: Rutgers University Press.
- Furgeson, J., Gill, B., Haimson, J., Killewald, A., McCullough, M., Nichols-Barrer, I., ... Lake, R. (2012). *Charter-school management organizations: Diverse strategies and diverse student impacts*. Cambridge, MA: Mathematica Policy Research.
- Furgeson, J., McCullough, M., Wolfendale, C., & Gill, B. (2014). *The Equity Project Charter School: Impacts on student achievement*. Cambridge, MA: Mathematica Policy Research.

- Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, 38, 915–945.
- Gawande, A. (2011, October 3). Personal best. *The New Yorker*.
- Gill, B., Timpane, P. M., Ross, K. E., Brewer, D. J., & Booker, K. (2007). *Rhetoric versus reality: What we know and what we need to know about vouchers and charter schools*. Santa Monica, CA: The RAND Corporation.
- Gill, B., English, B., Furgeson, J., & McCullough, M. (2014). *Alternative student growth measures for teacher evaluation: profiles of early-adopting districts*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.
- Gill, B., & Nichols-Barrer, I. (2014). Charter schools. In D. Brewer & L. Picus (Eds.), *Encyclopedia of Education Economics and Finance*. Thousand Oaks, CA: Sage Publications, Inc.
- Glenn, Jr., C. L. (1988). *The myth of the common school*. Amherst, MA: University of Massachusetts Press.
- Gold, M. (2010). *Accountable care organizations: Will they deliver?* Princeton, NJ: Mathematica Policy Research.
- Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics*, 89(1), 134–150.
- Hallberg, K., & Green, G. (2015, March 11). How can we hire and keep high quality teachers in struggling schools? InformED Blog, American Institutes for Research Education Policy Center.
- Hamilton, L. S., Stecher, B. M., & Yuan, K. (2012, June). Standards-based accountability in the United States: Lessons learned and future directions. *Education Inquiry* 3(2), 149–170.
- Hargreaves, A. P., & Shirley, D. L. (2009). *The fourth way: The inspiring future of educational change*. Thousand Oaks, CA: Corwin.
- Harris, D. N., & Sass, T. R. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management*, 28(1), 55–80.
- Hawley, W., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 151–180). San Francisco, CA: Jossey Bass.

- Hernandez, M., Hodges, S., & Cascardi, M. (1998). The ecology of outcomes: System accountability in children's mental health. *The Journal of Behavioral Health Services & Research*, 25(2).
- Hess, F. (2013, Fall). The missing half of school reform. *National Affairs*, 17, 19–35.
- Hess, F. (2015, March 13). Thinking about “what works.” *Education Week* blog. Retrieved from http://blogs.edweek.org/edweek/rick_hess_straight_up/2015/03/thinking_about_what_works.html
- Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states’ tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation* (pp. 163–189). Greenwich, CT: Information Age Publishing.
- Housner, L. D., & Griffey, D. C. (1985). Teacher cognition: Differences in planning and interactive decision making between experienced and inexperienced teachers. *Research Quarterly for Exercise and Sport*, 56, 45–53.
- Hout, M., & Elliott, S. W. (Eds.). (2011). *Incentives and test-based accountability in education*. Washington, DC: National Academies Press.
- Imberman, S. (2011). The effect of charter schools on achievement and behavior of public-school students. *Journal of Public Economics*, 95, 850–863.
- Jacob, B. (2005). Accountability, incentives, and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, 89, 761–796.
- Jennings, J., & Sohn, H. (2014). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. *Sociology of Education*, 87(2), 125–141.
- Jinnai, Y. (2013). *The impact of charter schools’ entry on traditional public schools: New evidence from North Carolina* (Job market paper). Rochester, NY: University of Rochester.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
- Kane, T. J. (2012, Fall). Capturing the dimensions of effective teaching. *Education Next*, 12(4).
- Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.
- Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? *Economics of Education Review*, 27(6), 615–631.
- Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working paper no. 14607). Cambridge, MA: National Bureau of Economic Research.

- Knight, J., & Cornett, J. (2009). *Studying the impact of instructional coaching* (Unpublished paper). University of Kansas. Retrieved from www.instructionalcoaching.org
- Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.
- Kremer, M., & Levy, D. (2008). Peer effects and alcohol use among college students. *The Journal of Economic Perspectives*, 22(3), 189–206.
- Lake, R., Bowen, M., Demeritt, A., McCullough, M., Haimson, J., & Gill, B. (2012). *Learning from charter school management organizations: Strategies for student behavior and teacher coaching*. Cambridge, MA: Mathematica Policy Research.
- Lake, R., & Miller, C. (2012). *Strengthening the civic mission of charter schools*. Washington, DC: American Enterprise Institute.
- Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125, 255–275.
- L’Hommedieu, R., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology*, 82(2), 232–241.
- Linkenbach, J., & Perkins, H. W. (2003). Most of us are tobacco free: An eight-month social norms campaign reducing youth initiation of smoking in Montana. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse: A handbook for educators, counselors, and clinicians* (pp. 224–234). San Francisco: Jossey-Bass.
- Lockwood, J. R., McCombs, J. S., & Marsh, J. (2010). Linking reading coaches and student achievement: Evidence from Florida middle schools. *Educational Evaluation and Policy Analysis*, 32(3), 372–388.
- Loewenstein, G., & Ubel, P. (2010, July 14). Economics behaving badly. *The New York Times*.
- Mansbridge, J. (2009). A “selection model” of political representation. *The Journal of Political Philosophy*, 17(4), 369–398.
- Marsh, J. A., McCombs, J. S., & Martorell, F. (2010). How instructional coaches support data-driven decision making. *Educational Policy*, 24(6), 872–907.
- McCullough, M., English, B., Angus, M. H., & Gill, B. (2015). *Alternative student growth measures for teacher evaluation: Implementation experiences of early-adopting districts*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

- Mendl, M. (1999). Performing under pressure: Stress and cognitive function. *Applied Animal Behaviour Science*, 65(3), 221–244.
- Mero, N. P., & Motowidlo, S. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology*, 80(4), 517–524.
- National Association of Public Charter Schools. Charter School Dashboard. Retrieved from <http://dashboard.publiccharters.org/dashboard/home>
- Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal*, 46(2), 532–566.
- Newton, L. H., Hodges, L., & Keith, S. (2013). Accountability in the professions: Accountability in journalism. *Journal of Mass Media Ethics*, 19, 166–190.
- OECD. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. Paris: OECD Publishing.
- O’Keefe, P., & Johnston, M. (1989). Perspective taking and teacher effectiveness: A connecting thread through three developmental literatures. *Journal of Teacher Education*, 40(3), 20–26.
- Patil, S., Vieider, F., & Tetlock, P. E. (2012). *Process and outcome accountability: Oxford handbook of public accountability*. New York, NY: Oxford University Press.
- Patil, S. V., & Tetlock, P. E. (2014). Punctuated incongruity: A new approach to managing trade-offs between conformity and deviation. *Research in Organizational Behavior*, 34, 155–171.
- Porter, A. C., Polikoff, M. S., Goldring, E. B., Murphy, J., Elliott, S. N., & May, H. (2010). Investigating the validity and reliability of the Vanderbilt Assessment of Leadership in Education. *The Elementary School Journal*, 111(2), 282–313.
- Pratt, J. W., & Zeckhauser, R. J. (Eds.). (1985). *Principals and agents: The structure of business*. Boston, MA: Harvard Business School Press.
- Rice, J. K. (2013). Learning from experience? Evidence on the impact and distribution of teacher experience and the implications for teacher policy. *Education Finance and Policy*, 8(3), 332–348.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417–458.
- Ross, J., & Staw, B. M. (1993). Organizational escalation and exit: Lessons from the Shoreham nuclear power plant. *The Academy of Management Journal*, 36(4), 701–732.
- Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics*, 116(2), 681–704.

- Scholten, L., van Knippenberg, D., Nijstad, B. A., & De Dreu, C. K.W. (2007). Motivated information processing and group decision-making: Effects of process accountability on information processing and decision quality. *Journal of Experimental Social Psychology*, 43, 539–552.
- Shavelson, R., Cadwell, J., & Izu, T. (1977). Teachers' sensitivity to the reliability of information in making pedagogical decisions. *American Educational Research Journal*, 14 (2), 83–97.
- Sherman, G. D., Lee, J. J., Cuddy, A. J. C., Renshon, J., Oveis, C., Gross, J. J., & Lerner, J. S. (2012). Leadership is associated with lower levels of stress. *Proceedings of the National Academy of Sciences*, 109(44), 17903–17907.
- Siegel-Jacobs, K., & Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes*, 1, 1–17.
- Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes*, 51(3), 416–446.
- Snipes, J., Doolittle, F., & Herlihy, C. (2002). Foundations for success: Case studies of how urban school systems improve student achievement. New York, NY: MDRC.
- Springer, M. G., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D. F., ... Stecher, B. M. (2012a). *Final report: Experimental evidence from the Project on Incentives in Teaching (POINT)*. Nashville, TN: National Center on Performance Incentives at Peabody College, Vanderbilt University.
- Springer, M. G., Pane, J. F., Le, V., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012b). Team pay for performance: experimental evidence from the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*, 34(4), 367–390.
- Sutton, R. I., & Galunic, D. C. (1996). Consequences of public scrutiny for leaders and their organizational image and its management. *Administrative Science Quarterly*, 36, 245–268.
- Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), 3628–3651.
- Tetlock, P. E. (1983). Accountability and perseverance of first impressions. *Social Psychology Quarterly*, 46, 285–292.
- Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology: Attitudes and Social Cognition*, 45, 74–83.
- Tetlock, P. E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*, 48(3), 227–236.
- Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in Experimental Social Psychology*, 25, 331–376.

- Tetlock, P. E., & Kim, J. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, 52(4), 700–709.
- Tetlock, P. E., & Mellers, B. A. (2011). Structuring accountability systems in organizations: Key trade-offs and critical unknowns. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence analysis: Behavioral and social scientific foundations*. Washington, DC: National Academies Press.
- Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology*, 57(4), 632–640.
- Tucker, M. S. (2014). *Fixing our national accountability system*. Washington, DC: National Center on Education and the Economy.
- Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013, February). *KIPP middle schools: Impacts on achievement and other outcomes*. Washington, DC: Mathematica Policy Research.
- Tyack, D. (1974). *The one best system: A history of American urban education*. Cambridge, MA: Harvard University Press.
- Tyack, D., & Hansot, E. (1982). *Managers of virtue: Public school leadership in America, 1820–1980*. New York, NY: Basic Books.
- Walsh, E., & Lipscomb, S. (2013). *Classroom observations from Phase 2 of the Pennsylvania Teacher Evaluation Pilot: Assessing internal consistency, score variation, and relationships with value added*. Cambridge, MA: Mathematica Policy Research.
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project.
- Wilson, J. Q. (1989). *Bureaucracy: What government agencies do and why they do it*. New York, NY: Basic Books.
- Winters, M. A. (2012). Measuring the competitive effect of charter schools on public school student achievement in an urban environment: Evidence from New York City. *Economics of Education Review*, 31(2), 293–301.
- Wolf, P., Kisida, B., Gutmann, B., Puma, M., Eissa, N., & Rizzo, L. (2013). School vouchers and student outcomes: Experimental evidence from Washington, DC. *Journal of Policy Analysis and Management*, 32(2), 246–270.
- Zimmer, R., Gill, B., Booker, T. K., Lavertu, S., Sass, T. R., & Witte, J. (2009). Charter schools in eight states: Effects on achievement, attainment, integration, and competition. Santa Monica, CA: RAND.

About the series

Policymakers and researchers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers access to our most current work.

For more information about this paper, contact Brian Gill, senior fellow, at bgill@mathematica-mpr.com.

Suggested citation: Gill, Brian P., Lerner, Jennifer S., and Meosky, Paul. "Re-Imagining Accountability in K-12 Education: A Behavioral Science Perspective." Working Paper 39. Cambridge, MA: Mathematica Policy Research, May 2015.

www.mathematica-mpr.com

**Improving public well-being by conducting high-quality,
objective research and surveys**

PRINCETON, NJ - ANN ARBOR, MI - CAMBRIDGE, MA - CHICAGO, IL - OAKLAND, CA - WASHINGTON, DC

MATHEMATICA
Policy Research

Mathematica® is a registered trademark
of Mathematica Policy Research, Inc.