ANNUAL REVIEWS

Annual Review of Law and Social Science

Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement

Daniel E. Ho and Sam Sherman

Stanford Law School, Stanford University, Stanford, California 94305;
email: dho@law.stanford.edu, ssherma2@stanford.edu

Annu. Rev. Law Soc. Sci. 2017. 13:251–72

First published as a Review in Advance on July 28, 2017

The *Annual Review of Law and Social Science* is online at lawsocsci.annualreviews.org

https://doi.org/10.1146/annurev-lawsocsci-110316-113608

Keywords

administrative law, quality assurance, quality improvement, performance measurement, performance management

Abstract

Decentralized decisions among government officials can cause dramatic inconsistencies in bureaucratic decision making. This article provides a synthetic review of the evidence base for improving the quality of bureaucratic decisions and reducing such street-level arbitrariness. First, we offer a typology to unify quality assurance management techniques often treated in distinct scholarly literatures. This synthesis reveals common challenges but also points to novel hybrid solutions that borrow across management techniques. Second, although empirical evidence is limited, our review suggests that ongoing management techniques, such as monitoring, peer review, and pay-for-performance, are more successful than ex post techniques, such as audits and appeals. Third, performance measurement and pay exacerbate the quantity–quality trade-off long opined about in public administration. We offer suggestions for future directions—most importantly, the vital role of academic-agency research collaborations in crafting quality improvement efforts—to address this endemic challenge to bureaucracy and rule of law.

1. INTRODUCTION

The ideal Weberian bureaucracy is impersonal (Weber 1968, pp. 979–80). Decisions should follow rules, not the whims of an official. Yet how can government assure the quality and consistency of bureaucratic decisions, given the modern reality of decentralized, complex decisions dispersed among many line-level officials?[1]

The challenge of guaranteeing bureaucratic rationality among such officials (a.k.a. street-level bureaucrats) has consumed scholars for decades (Bardach & Kagan 1982, Lipsky 1983, Wilson 1989). Documenting dramatic disparities in social security disability adjudication, Mashaw (1973) argued that due process of law mandated better management of employees. The problem is not isolated to social security. The evidence is overwhelming that inconsistency in decision making is endemic across the administrative state, from immigration adjudication to nuclear safety inspections and from child welfare determinations to food safety inspections (Ho 2017, pp. 5–12). Discretion remains in the face—and even perhaps because—of complex rules (Bardach & Kagan 1982). Call this the problem of "street-level arbitrariness."[2] Mashaw (1973, pp. 790–91) proposed that due process values pointed to something like a quality assurance program—the evaluation of employee performance against quality standards—to remedy this pervasive problem of street-level arbitrariness.

Nearly 45 years after Mashaw sounded the alarm, what have we learned about how government can effectively assure quality and consistency? This article synthetically reviews empirical evidence about the dominant techniques to manage and improve the quality of line-level decision making. We proceed as follows. Section 2 briefly discusses theoretical and political developments in public-sector quality management and performance measurement. Section 3 presents a novel typology that elucidates institutional design differences among seven dominant management techniques. Although discrete literatures have studied these techniques, no review has synthesized them in a common framework that provides insights about common challenges as well as novel, hybrid solutions. We then proceed by reviewing the evidence base for these techniques. Section 4 discusses training occurring before officials formally produce decisions (ex ante). Section 5 discusses the evidence for ongoing initiatives, namely peer review, monitoring, disclosure, and pay-for-performance. Section 6 discusses the two common ex post initiatives: auditing and internal agency appeals. Section 7 concludes.

2. THEORY AND CONTEXT

In their popular best-selling book, *Reinventing Government*, Osborne & Gaebler (1993) argued that performance measurement and management would transform bureaucracies into nimble, self-learning, customer-oriented organizations. New Governance was hailed by many as a general solution to cure public sector agency problems, and *Reinventing Government* served as the blueprint for Clinton reforms to reinvent government. The Government Performance and Results Act [Pub. L. No. 103-62], passed during the Clinton administration, mandated annual performance plans for agencies and linking budgeting decisions to performance measures.

[1] We use the term line-level officials to refer to officials who have decision-making capacity in the first instance of the relevant administrative agency. Throughout the piece, we use the terms frontline, line-level, and street-level interchangeably to refer to such decisions.

[2] The problem is not confined to street-level bureaucrats as conventionally envisioned (e.g., police). However, because of the familiarity with street-level bureaucrats, we use street-level arbitrariness as shorthand to encompass the general problem of line-level bureaucratic arbitrariness.

The scholarly reception was more critical (Kravchuk & Schack 1996). Simple principal-agent models were recognized to be inadequate for explaining public-sector organizations, owing to multiple service goals and complex agency relationships (Dixit 2002, Holmstrom & Milgrom 1991). Dixit (2002, p. 697) argued that these characteristics "make inappropriate the naïve application of magic bullet solutions." Behn (2003, p. 586) cautioned public managers against seeking "the one magic performance measure." Radin (2006, pp. 3, 33–51) pointed out a "misfit between expectations and practice" and noted that "one size fits all" tendencies have historically undermined the effectiveness of performance management schemes. This skepticism of performance management has long-standing roots. In a classic piece, Kerr (1975, pp. 779–80) articulated the quantity–quality trade-off, arguing that "simple, quantifiable standards against which to measure and reward performance . . . may be successful in highly predictable areas . . . but are likely to cause goal displacement when applied anywhere else." Prioritizing caseload production among administrative law judges (ALJs), for example, might sacrifice the quality of decisions, a theme recurring across the management techniques described below.

How well have such initiatives actually worked? Ideally, any quality improvement initiative would be subjected to rigorous evaluation. In practice, agencies themselves rarely offer strong evidence. The Program Assessment Rating Tool, for instance, encouraged evaluation using experimental methods, but agencies continued to offer largely qualitative evidence of impact (Heinrich 2012). To conduct this review, we searched for any published works empirically studying the efficacy of prevailing management techniques to improve the quality and consistency of bureaucratic decision making. In one sense, the literature turns out to be remarkably thin: Few rigorously designed studies of quality assurance programs exist. Yet in another sense, many social scientific insights can be culled together from discrete literatures in administrative law, public administration, personnel economics, and public economics.

3. A TYPOLOGY

Given the limits of rules,[3] our review focuses on seven predominant management techniques to address street-level arbitrariness: training, peer review, monitoring, disclosure, pay-for-performance, auditing, and internal appeals.[4] Before we review the evidence, we offer a typology of ideal types in **Figure 1** to understand the conceptual relationship between these techniques. Each row represents one technique (e.g., training), sorted by the timing of when it occurs in the administrative process (ex ante, ongoing, or ex post), and the columns classify whether the feature is a critical (*black cells*), potential (*gray cells*), or uncharacteristic (*white cells*) dimension of the management technique.

The typology offers three broad lessons. First, **Figure 1** highlights critical distinctions between ideal types. Disclosure and pay-for-performance share many typical features, but the critical

[3]We do not examine as a management technique the impact of reducing discretion by crafting rules. The sentencing context provides evidence that rules can reduce interjudge disparities (Yang 2014), and evidence from the agency context also suggests that rules can be suboptimally complex and exacerbate differences between line-level officials (Braithwaite & Braithwaite 1995, Ho 2012). Instead, we begin our review from the premise of Bardach & Kagan (1982) and Lipsky (1983, pp. 14–16) that line-level bureaucrats retain tremendous discretion. We hence focus on techniques that manage street-level arbitrariness given a set of rules, even though some management techniques seek to clarify rules (e.g., peer review). In addition, although there may be promise to behavioral economics approaches (e.g., Thaler & Sunstein 2008), checklists (Gawande 2010), and predictive analytics (Ray & Lubbers 2014), we do not include them in our review here, as they have not been primarily focused on improving the quality and consistency of line-level officials in the public sector. Section 6.2 briefly touches on predictive analytics in the context of promising improvements of the Social Security appeals process.

[4]See Sections 3.1–3.7 for descriptions of each of these techniques.

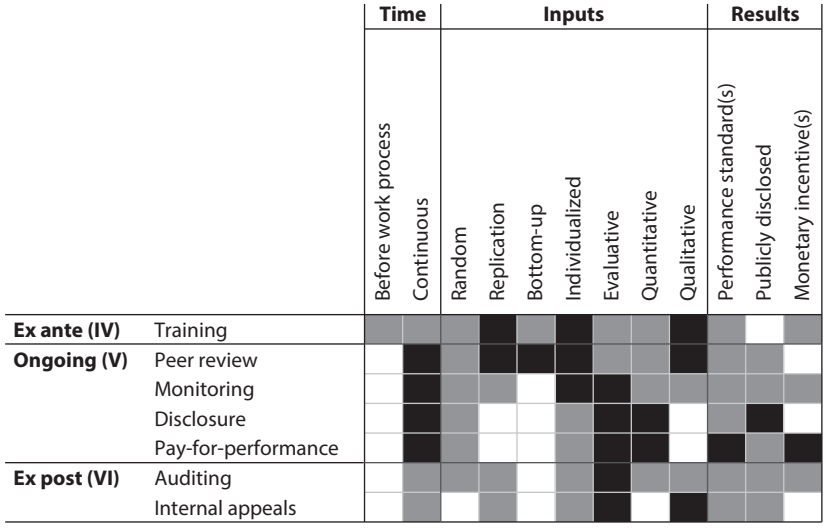

Figure 1

Typology of ideal types. Each cell is coded for whether the feature is uncharacteristic (*white*), potential (*gray*), or critical (*black*) to the management technique. Categories: Time, timing and frequency of the technique; Inputs, process and substance of information collection that a technique entails; Results, how the collected information is used. Headers: Before work process, whether the intervention occurs before any work process occurs (e.g., orientation training to a new employee); Continuous, whether the intervention occurs on a continuing basis or only a discrete number of times, but does not refer to whether the intervention occurs over a continuous interval of time; Random, whether a random sample of employees and/or work products is involved; Replication, whether the judgment in the intervention attempts to replicate the steps used to produce the work being evaluated, giving due consideration to the same factors that were considered by the unit/employee who originally produced the work; Bottom-up, whether frontline employees are a key driver to quality improvement, perhaps by facilitating or assisting in carrying out the intervention (e.g., frontline employees reviewing each other's work), but does not mean that the quality improvement process was driven exclusively by frontline workers; Individualized, individual information, such as training or feedback, is provided to workers, but this does not exclude more general information from being provided in addition to the individualized information; Evaluative, whether the purpose is to formally evaluate a worker, group of workers, or organization; Quantitative, whether work products and/or employees are evaluated using quantitative indicators (e.g., caseload statistics); Qualitative, whether work products and/or employees are evaluated qualitatively (e.g., through a qualitative evaluation based on supervisorial observation); Performance standard(s) intervention involves collecting performance measures and comparing those measures against a set standard; Publicly disclosed, some information collected during the intervention is disclosed to the public; Monetary incentive(s), results are directly tied to at least one monetary incentive and/or sanction affecting budget, promotion, and bonus/salary decisions.

distinction is that the latter entails a direct monetary incentive for meeting a standard. Second, although much research has focused on discrete techniques, this typology reveals that there is much more fluidity across categories than conventionally recognized. Many features can potentially be incorporated (*gray cells*), such that one technique can blend into another. Monitoring in the form of performance reviews that prioritize quantitative indicators (e.g., a 5% bonus if an employee conducts 300 inspections) can approximate a formal pay-for-performance scheme. Third, the typology also suggests novel combinations. Training, for instance, could be coupled with direct monetary incentives, such as the receipt of a bonus for meeting a performance standard early in the training period. With 12 binary design features, 4,096 unique combinations of quality

arrangements exist, pointing to many unexplored ways to solve the quality conundrum. We now sketch out the 7 predominant techniques.

3.1. Training

Training is the ex ante education of line-level employees in how to deliver government services. For instance, a food safety program may have a training period before inspectors engage in formal inspections. Training typically occurs before the delivery of formal work product and involves individualized feedback that attempts to replicate frontline decisions. While qualitative assessment is a critical feature of training, **Figure 1** also suggests novel hybrids: Although conventionally seen as qualitative, training could be combined with quantitative performance indicators; although typically occurring ex ante, training efforts could be structured on an ongoing basis; and although ongoing training efforts for the full staff can be costly, random selection of employees and/or work products could be deployed to design more cost-effective training.

3.2. Peer Review

Peer review is the evaluation of work product by another frontline employee. For instance, Seattle–King County instituted a peer review process during which health inspectors were randomly paired up to conduct inspections and discuss differences in health code interpretation (Ho 2017). **Figure 1** illustrates that although peer review shares many features with training, it is distinct in two respects: It occurs continuously (rather than ex ante), and learning is bottom up (i.e., driven by line-level employees, not by senior staff). Peer review typically does not deploy monetary incentives, prioritizing collaboration over competition. **Figure 1** also points to novel combinations: Peer review could be combined with performance standards, such as calibrating the intensity of the program based on measures of arbitrariness (e.g., peer disagreement), and peer review results could be publicly disclosed and/or be used in a more evaluative sense, as was the case in the peer review process designed in response to structural litigation in child welfare (Noonan et al. 2009, pp. 533–51). Sabel & Simon (2017, pp. 74–78) conceptualize peer review as the hallmark for an alternative organizational form (i.e., to ensure fidelity to not only rules but goals).

3.3. Monitoring

Monitoring consists of the evaluation of frontline personnel, typically by supervisors. Conventional monitoring in the public sector consists of top-down performance appraisals, but many other design choices exist. Managers may prioritize quantitative or qualitative indicators, each varying in complexity: Duflo et al. (2012), for instance, studied the effect of monitoring a simple indicator of work attendance, whereas Jacob & Lefgren (2005) studied monitoring based on value-added teaching measures. Although monitoring can be time-consuming, **Figure 1** also shows that random sampling can facilitate monitoring and reduce costs. To ensure the quality of translation in immigration asylum interviews, for instance, phone monitoring could be completed for a random quality assurance sample, as is routinely conducted in the private sector (see the literature on statistical process control, e.g., Oakland 2007).

3.4. Disclosure

Disclosure involves public dissemination of information regarding frontline work product or personnel. Baltimore, for instance, publishes performance indices for municipal agencies (Perez &

Rushing 2007, p. 11), and the Social Security Administration (SSA) publishes caseload statistics of each ALJ. A more concerted effort to use disclosure to improve management by public pressure is ranking jurisdictions on the quality of election administration (Gerken 2009, pp. 86–92). Performance-based budgeting consists of making appropriations based on such performance disclosures, but when managers link this information to salary or bonus decisions for individual employees, disclosure converges to pay-for-performance. A common criticism is that disclosure causes fixation on easily available quantitative measures, but the typology reveals that random sampling can be deployed to measure more dimensions of quality in cost-effective ways.

3.5. Pay-For-Performance

Pay-for-performance schemes pay frontline personnel for meeting a performance standard, typically quantitative. For instance, health inspectors might be paid a bonus for conducting a target number of inspections. **Figure 1** shows that disclosure and pay-for-performance are close cousins, with the critical distinction of direct monetary incentives. As **Figure 1** shows, pay-for-performance can blend into other techniques. Monitoring in the form of performance appraisals that tie teacher salaries to measures of student achievement (Jacob & Lefgren 2005, p. 1), for instance, can resemble pay-for-performance. If third parties replicated decisions and measured outcomes, performance measures could be more robust strategic gaming (e.g., performance pay for health inspectors based on third-party inspection results).

3.6. Auditing

Auditing consists of the evaluation of frontline work products or personnel, typically by an external party. Whereas monitoring conventionally evaluates frontline personnel on an individual basis, auditing is typically conducted on a program basis. Again, **Figure 1** illustrates novel configurations. Audits can, in principle, be ongoing and based on random samples of work product. Indeed, this is one way to understand Cuéllar's (2006) proposal for executive auditing to improve bureaucratic decisions. Similarly, the Food and Drug Administration's standardization process essentially entails an audit of each frontline inspector at three-year intervals, showing that audits can be designed in an individualized and ongoing manner, even if typically programmatic and episodic (US Food Drug Adm. 2015, p. 11).

3.7. Internal Appeals

Regulated parties can often appeal line-level decisions within the agency. For instance, restaurateurs may appeal an inspection result to a supervisor (or separate agency tribunal). Because parties decide whether to appeal, such cases are neither random nor representative, but appeals may yield lessons to improve frontline decisions. **Figure 1** again highlights interesting hybrids. First, appeal outcomes could be tied to a performance standard; e.g., ALJs whose decisions are reversed at too high a rate may be seen as falling below a reversal performance standard.[5] Second, not all systems described as appeals are in fact appeals. The SSA's "pre-effectuation review," which consists of reviewing a random sample of ALJ decisions (Off. Insp. Gen. 2012, p. 7), basically functions as an audit system.

[5] Social Security disability claimants can appeal a disability determination to an ALJ. In the 1980s, the SSA briefly set per-month caseload requirements for ALJs (Cofer 1985a, p. 231).

Before we turn to the evidence base for each discrete technique, we should articulate the criteria for our literature review. First, we sought to identify the rigorously designed studies of the efficacy of management techniques. We hence prioritized randomized controlled trials (RCTs) and natural experiments over observational and case studies. In areas where well-designed studies were lacking, we resorted to the best available evidence. Second, although we focused on the US context, we included well-designed studies from the comparative context where US evidence was lacking. Third, although our focus was on the public sector, we included private sector studies when these added important and potentially generalizable insights not studied in the public sector literature. As a result of these criteria, our review remains necessarily selective. Many well-designed studies are centered on healthcare, education, and labor training, showing that much work remains ahead in understanding the efficacy of quality initiatives across the administrative state.[6]

4. EX ANTE: TRAINING

Training is the prototypical way in which agencies prepare employees prior to producing a formal decision. Training theoretically benefits employees by exposing them to situations that they might encounter in their work, but in a less consequential or simulated environment. Although there are a myriad of questions in the design of ex ante training (How long? What mix of lectures, reading, and exercises? Should trainees shadow current employees?), we find virtually no systematic evidence assessing the impact of ex ante training per se. The closest literature is on government job training, with modest positive results (Greenberg et al. 2003, 2006), and teacher training, with mixed evidence (Bouguen 2016, pp. 92–93).

A more expansive evidence base exists for in-service training. For example, Banerjee et al. (2012) found that randomly enrolling Indian police officers into in-service training consisting of skill modules lasting a few days had a positive and significant impact on crime victims' satisfaction with the police. They attribute training's success over other failed interventions to the fact that it could largely be run without continuous commitment by ground staff, such as police station supervisors (p. 36).

A meta-analysis (Mansouri & Lockyer 2007) and a Cochrane review (Forsetlund et al. 2009)[7] of "continuing medical education" found small, positive effects on physician performance and patient outcomes. Mansouri & Lockyer (2007) found that the most effective interventions (*a*) were interactive, (*b*) used multiple training methods, (*c*) were longer, and (*d*) were designed for a small group of physicians from a single discipline. Forsetlund et al. (2009) found that higher attendance at educational meetings and mixed interactive/didactic sessions were associated with better practice outcomes.[8] In education, the evidence of the effect of professional development on student achievement is largely mixed, with many null results in RCTs (Hill et al. 2013, pp. 476–78). Hill et al. (2013) found no difference between online and in-person training but found that observing an expert may be more beneficial than observing video of oneself.

[6]The closest work is that by Finan et al. (2015), who review personnel economics of the state. Their scope differs in that they cover selection and retention of public-sector employees, focus largely on the development context, and rely nearly exclusively on RCTs. Our review overlaps in the evidence for monitoring, pay-for-performance, and auditing.

[7]Cochrane reviews are systematic literature reviews of empirical studies related to healthcare.

[8]Forsetlund et al. (2009, p. 13), however, also found evidence suggesting publication bias (bias in published effect sizes compared to effect sizes of all studies conducted).

5. ONGOING INITIATIVES

5.1. Peer Review

Because peer review leverages the power of mutual observation and learning from one's colleagues, many scholars have suggested peer review as an antidote to the problem of street-level arbitrariness. Until recently, the evidence in the public sector consisted principally of observational case studies. Braithwaite & Braithwaite (1995) found that an exit conference for nursing home inspectors could promote consistency in Australia. In child welfare, Noonan et al. (2009) argued that caseworker peer review improved the quality of case management. Ho (2017) conducted the only RCT of peer review in a government agency, which randomized food safety inspectors into weekly peer review inspections and meetings to clarify challenging code items and found that the intervention increased the accuracy and consistency of inspections, even when inspectors were out alone.

In the teaching context, randomized experiments (Bowman & McCormick 2000) and multiple baseline studies (Mallette et al. 1999), one with a nonrandom control group (Morgan et al. 1992), have generally suggested positive effects for peer coaching of teacher trainees on teaching behavior and/or student achievement. Goldstein (2007) studied a school district before and after implementing teacher peer review and found that it improved accountability among teachers.

A related literature studies peer coaching principally of medical doctors, finding generally favorable effects (Schwellnus & Carnahan 2014). Gattellari et al. (2005) conducted the only RCT of peer coaching per se, documenting increased knowledge among general practitioners about screening for prostate cancer compared with the control group. A Cochrane review found two RCTs generally suggesting positive effects for physician feedback from peer physicians (Ivers et al. 2012, p. 25). Several RCTs also examined the effect of peer teaching (or peer-assisted learning) on learning outcomes for medical students, finding peer teachers (i.e., medical students) comparable to professional teachers (Secomb 2008, Yu et al. 2011). Randomized experiments on life coaching, however, found that external or professional coaching outperformed peer coaching among MBA students (Sue-Chan & Latham 2004) and general adults (Spence & Grant 2007).

Overall, the literature suggests that peer review can have positive effects, but much more study is needed in the public sector. One particular promise of peer review is that its more collaborative orientation can overcome political resistance to top-down quality assurance efforts (Ho 2017), such as the fierce debates over performance appraisal of ALJs.

5.2. Monitoring

If a major impediment lies in the ability to observe the quality of employee work, then monitoring could improve quality. The literature on monitoring suggests that external monitoring (i.e., by a party outside of the organization) is better than direct, in-person monitoring and that managerial commitment is critical. Evidence for use of financial incentives, types of monitoring measures, relative performance feedback (RPF), and group versus individual feedback remains inconclusive.

Callen et al. (2015) found that smartphone monitoring of health clinics nearly doubled clinic inspection rates and decreased physician absence in politically competitive districts in Pakistan, and that increasing the salience of physician absence to senior officials significantly decreased absences. Dhaliwal & Hanna (2014) found that thumb-scan and phone-monitoring devices improved health outcomes in Indian public-sector healthcare, though effects on medical staff attendance were heterogeneous, and staff expressed lower job satisfaction. Impersonal, electronic mechanisms may be effective, consistent with evidence outside of the public sector (Pierce et al. 2015, Staats et al. 2016). Nagin et al. (2002) studied a monitoring scheme of employees at a telephone call

center, consisting of calling back a nontrivial percentage of their pledges. Monitors were at central headquarters, not the immediate supervisors (p. 851), and the study found workers cheated less when the probability of detection increased, but treatment effects were heterogeneous, as workers responded differentially to manipulations in the monitoring rate. One RCT suggested that in-person, external monitoring was effective when the monitors were students taking daily pictures of their teachers to verify attendance (Duflo et al. 2012). Similarly, Banerjee et al. (2012, p. 30) found that decoy observers improved police behavior, even when results would not be conveyed to supervisors.

Other evidence suggests that direct, in-person monitoring by supervisors may be undesirable. In a private sector experiment, Bernstein (2012) found that direct monitoring may induce employees to go by the book, decreasing learning of techniques that increased frontline job productivity but were unknown to supervisors. Direct monitoring can also produce gaming. Chen & Kremer (2002) found that school headmasters tasked as monitors in a teacher attendance bonus program falsely reported attendance.

As Chen & Kremer (2002) demonstrate, a lack of sustained managerial commitment can cause the positive effect of a monitoring program to attenuate over time. Banerjee et al.'s (2008) experiment regarding Indian government nurse monitoring combined with financial incentives to promote attendance similarly found that local administrators were complicit in weakening the monitoring program, letting nurses claim an increasing number of "exempt days" (for more on gaming, see Section 5.4). Dhaliwal & Hanna (2014, p. 6) found that supervisors approved most exemption requests even in the face of attempts to restrict such exemptions. Staats et al. (2016) found that the positive effects of electronic monitoring gradually attenuated, and notes that "ongoing managerial interventions [were needed] to sustain the benefits of monitoring" (p. 1583). The positive effects of electronic monitoring reported by Callen et al. (2015, p. 23) and Dhaliwal & Hanna (2014, p. 5) also appear to have declined over time.

The evidence for combining monitoring with financial incentives is generally mixed (Banerjee et al. 2008, Chen & Kremer 2002, Nagin et al. 2002). Monitoring is often coupled with performance feedback, but it is unclear how feedback interacts with financial incentives. One RCT found that physician feedback combined with a financial incentive increased immunization coverage (Fairbrother et al. 1999), but two RCTs of physician feedback and incentives failed to find an effect on cancer screening rates and immunizations, respectively (Hillman et al. 1998, 1999). However, a private-sector meta-analysis of 72 empirical studies suggested that supervisorial feedback could amplify the effects of financial incentives (Stajkovic & Luthans 2003, p. 174).

It is also unclear whether monitoring should use quantitative or qualitative performance indicators, or both. Two studies supporting the use of quantitative indicators are presented by Jacob & Lefgren (2005), who found that teacher value-added measures did a better job of predicting future student achievement than subjective principal ratings, and Rockoff et al. (2012), who found that principals increasingly incorporated objective performance data when the data were more precise and their priors were less precise. Conversely, a laboratory experiment suggests that supervisors may use quantitative indicators in asymmetric or distortionary ways (Bol & Smith 2011).[9]

RPF, consisting of informing employees of their individual performance ranking relative to their peers, does not appear to work well. Cochrane reviews reported mixed experimental evidence for peer-comparison feedback on outcomes such as asthma management and quality of care for diabetic patients (Ivers et al. 2012, pp. 24–25) and found RPF's effect on these outcomes comparable

[9]The authors found that supervisorial evaluations of sales staff were higher when an uncontrollable factor decreased sales but not lower when an uncontrollable factor increased sales. For more on the quantity–quality trade-off, see Section 5.4.

to other performance feedback (Jamtvedt et al. 2007, p. 12). One RCT found a negative effect of RPF on trainee exam scores in a health worker training program (Ashraf et al. 2014, p. 50), whereas another RCT found a positive effect of removing such feedback on sales performance (Barankay 2012). Quasi-experimental private-sector studies found benefits on worker productivity (Blanes i Vidal & Nossol 2011), particularly for public disclosure of RPF throughout an organization (Song et al. 2015). Lastly, a meta-analysis of RCTs in healthcare found similar effects of group and individualized feedback on healthcare quality-of-care measures but also found that more frequent feedback augmented effectiveness (Hysong 2009).

5.3. Disclosure

Managers of government agencies may publicly disclose performance measures to improve the effectiveness of line-level service delivery. Disclosure's appeal is that public pressure could improve work product by shaming or competition. Disclosure often includes performance rankings, such as Gerken's (2009) Democracy Index for ranking local- and state-level administration of elections, or Mason et al.'s (2014) disclosure of performance rankings of local police. Mason et al.'s (2014) pre-post study found that showing citizens real performance data of British police increased citizens' trust significantly across several performance domains. However, studies of disclosure as a technique to manage public officials are scarce, in part because disclosure is often coupled with other techniques (e.g., performance-based budgeting). Evidence on the efficacy of disclosure outside of the bureaucratic management context, however, can be informative.[10]

School accountability regimes assign letter grades or ratings to teachers or schools based on standardized test results. Much evidence demonstrates that this has induced gaming behavior, such as increasingly classifying students into categories exempted from school grading (Cullen & Reback 2006, Figlio & Getzler 2006); selective teacher behavior focusing on marginal students that most affect the school's grade (Krieg 2008, Neal & Schanzenbach 2010, Reback 2008); and, in the extreme, teacher cheating (Jacob & Levitt 2003). Incentives to game the grading scheme by overexempting students may be mitigated when there is a financial cost, such as a student voucher, to increasing exemptions (Chakrabarti 2013) (for more on gaming, see Section 5.4).

In the medical context, efforts have focused on the public release of hospital performance data. Yet the evidence base for such healthcare disclosures remains flimsy (Marshall et al. 2000).[11] One review found "[a]lmost no evaluations . . . have used controlled experimental designs" (Hibbard et al. 2003, p. 84), and a Cochrane review of all experimental, quasi-experimental, controlled before-after, and interrupted time series studies found no consistent indication that disclosure changed consumer or professional behavior, or that it improved care (Ketelaar et al. 2011). Prior literature reviews came to similar conclusions (Fung et al. 2008, Shekelle et al. 2008).

Several studies used pre-post comparisons or regression analyses to examine the effect of consumer report cards that graded the risk-adjusted performance of doctors on health outcomes. The tenor of these studies was positive on both health outcomes, such as risk-adjusted mortality (Hannan et al. 1994, Peterson et al. 1998), and hospital behavior, such as medical services offered (Longo et al. 1997). However, Dranove et al. (2003, p. 560) warned that provider selection, particularly with respect to sicker patients, may confound these effects. Though report card adherents

[10] For a review of disclosure policies, see Fung et al. (2007); for a more critical perspective, see Ben-Shahar & Schneider (2014).

[11] Several studies from the United Kingdom find positive effects of disclosed waiting time targets for hospitals (Besley et al. 2009, Propper et al. 2010) and school performance tables (Burgess et al. 2013), but the control group is cross-national, making inferences limited.

assert that the data do not support an out-migration of patients (Peterson et al. 1998) and that the models adjust for high-risk patients (Hannan et al. 1997), the debate has not been conclusively settled (for a detailed review of the controversy, see Marshall et al. 2000, pp. 51–66).

Although Dranove et al. pointed to powerful perverse incentives on disclosers, the effect on disclosees may be quite limited. Consumers do not appear to rely on report cards in medical care decisions (Faber et al. 2009, p. 7). Schneider & Epstein (1998) found that only 12% of patients were aware of the Pennsylvania hospital performance report on cardiac surgery prior to surgery. A consumer survey of large employers found that "the variety and amount of performance information . . . is a barrier to effective decision making" (Hibbard et al. 1997, p. 172). Of six studies analyzing the impact of healthcare disclosure, only one suggested that disclosure had some impact on consumer behavior (Marshall et al. 2000, p. 62). As a result, Hibbard et al. (2003, p. 93) argued that disclosed information should be "highly evaluable" and make "immediately obvious who the top and bottom performers are." A systematic review found a positive effect of easy-to-read presentation formats in laboratory studies (Faber et al. 2009, p. 5).

Public disclosure is intuitively appealing. The impact of stigma, sometimes called naming and shaming, has been demonstrated in both India (Pattanayak et al. 2009) and the United States (Figlio & Rouse 2006). But although both studies compared shaming or stigma with direct financial incentives, neither study could distinguish the mechanism from one of market discipline owing to the disclosure.[12] Some disclosures also institute a system of direct financial sanctions/rewards—e.g., in many school accountability regimes (Figlio & Rouse 2006, p. 239)—making it difficult (*a*) to empirically isolate the effect of disclosure and (*b*) to conceptually distinguish disclosure from pay-for-performance.

Last, the literature also dispels a common myth: Disclosure is not costless. Casalino et al. (2016) found that physician practices in four common specialties spend 785 hours per physician on average and more than $15.4 billion on reporting quality measures annually. Just as physician grading can shift resources away from the sickest patients and school grading can shift resources away from the lowest-performing students, Ho (2012) found that restaurant grading in New York and San Diego perversely shifted resources away from higher health hazards to resolve grade disputes.

5.4. Pay-For-Performance

Given that officials may lack the profit motive, some posit that quality can be improved by paying based on measured levels of performance. Pay-for-performance has the most extensive evidence base of techniques studied, albeit with less of a civil service focus. Hasnain et al. (2012, p. 39) review the literature, finding 20 field RCTs in OECD (Organisation for Economic Co-operation and Development) countries and 6 field RCTs in developing countries, but only one high-quality study of a context analogous to the civil service. That study (Dowling & Richardson 1997) found that a minority of managers within a public healthcare trust reported exerting more effort after performance pay was implemented. Hasnain et al. (2012, p. 30) note, "Most glaringly, the role of politicized bureaucracies has not been addressed properly." The Department of Defense implemented a pay-for-performance scheme in the National Security Personnel System, but "widespread employee dissatisfaction" (Haga et al. 2010, p. 211) and claims of discrimination (p. 223) contributed to its

[12] Indeed, the market disciplining effect would have been impossible to demonstrate in the context of work by Pattanayak et al. (2009), who dealt with inducing rural villages to switch from open defecation to the use of latrines through a community education program involving aspects of shaming.

repeal in 2010 (p. 223). In the private sector, a meta-analysis found pay-for-performance produced net benefits on productivity, moderated by the interestingness of the task (Weibel et al. 2010).

Several lessons can be distilled from the literature. First, pay-for-performance is most successful when the quantity–quality trade-off is least acute.[13] In some areas of healthcare, for example, morbidity rates or immunization rates are arguably closer to reflecting end goals. Thus, the trade-off between healthcare quality (the end goals) and measurable quantitative indicators (morbidity or immunization rates) is arguably less acute than in other sectors. Although reviews on physician pay-for-performance find the evidence base limited (Eijkenaar 2013, Scott et al. 2011), relative to other fields, the experimental evidence in healthcare mostly supports (albeit in some cases very weakly) the effectiveness of pay-for-performance (An et al. 2008, Basinga et al. 2011, Kouides et al. 1998, Roski et al. 2003). In a Cochrane review of financial incentives for primary care physicians, six of seven studies illustrated modestly positive effects on quality of care for some primary outcome measures, though not all (Scott et al. 2011, p. 2). Some of the modestly positive effects of pay-for-performance in healthcare may stem from a less acute quantity–quality trade-off.

Where the quantity–quality trade-off is most acute, performance measures can undermine desired outcomes. In the Job Training Partnership Act (JTPA) [or its successor, the Workforce Investment Act (WIA)], one performance measure, for instance, consisted of employment rates of individuals in the thirteenth week after completing the labor training program (Barnow & Smith 2004, pp. 24–25). Evidence suggests that training center employees gamed JTPA/WIA employment measures by altering the mix of services provided (Marschke 2002) and manipulating program graduation of participants to induce better performance on employment performance measures (Courty & Marschke 1997, 2004). Focusing only on a short-run employment rate incentivized training center employees to "cream-skim" applicants who appeared more employable prior to enrolling (Courty et al. 2011, Heckman & Smith 2003). The extent of cream-skimming may be overstated, however (Heckman et al. 2002). Heckman et al. (1996) found contrasting evidence, namely, strong negative selection on predicted earnings and weak evidence for positive selection on expected impacts. They attribute these findings to a "social worker mentality," suggesting that frontline attitudes can moderate the impact of perverse incentives. Gaming behavior has also been documented among Navy recruiters (Asch 1990) and healthcare providers (Chen et al. 2011, McDonald & Roland 2009) operating under pay-for-performance systems. In school accountability schemes, randomized evaluations found mixed evidence of teaching to the test [e.g., Glewwe et al. 2010 (positive evidence) and Muralidharan & Sundararaman 2011 (no evidence)]. A quasi-experimental study found no evidence of teacher manipulation of test scores (Lavy 2009).

Second, performance measurement does not equal program impact.[14] In an important study, Barnow (2000) used data from the JTPA randomized experiment in 16 job training sites from 1987 to 1989 and compared performance measures (e.g., employment rates postgraduation) with gold-standard experimental evidence of impact. He found only a weak positive correlation. Among job training programs, numerous analyses have found at best a weakly positive (and at worst a negative) relationship between performance measures and long-run program impacts, with one exception (Barnow & Smith 2004, pp. 33–40; Heckman et al. 2011).

[13] For the definitive historical treatment of the American shift from fee-based government service toward salarization, with many similar distortionary dynamics of fee-based government services, see Parrillo (2013).

[14] This point also underscores limitations of program evaluations that classify improvement in the program's performance measures as true success. Ideally, an impact evaluation can circumvent these challenges. For more on the difference between performance measurement and evidence-based policy evaluation, see Heinrich (2007). Also, some studies found benefits of pay-for-performance, such as increased mentoring (Glazerman & Seifullah 2012, p. xiv) and better documentation (Fairbrother et al. 1999), at least partially through qualitative methods.

Third, carefully designed performance standard adjustments (or risk adjustments) might more closely approximate program impact by adjusting for local conditions (Eijkenaar 2013, pp. 119–20). Yet a review of performance measurement adjustments found limited evidence of such alignment, and most performance regimes have not contemplated adjustments (Barnow & Heinrich 2010). Schochet & Fortson (2012) found that regression adjustment "somewhat" changed Job Corps training center performance measures, but the measures were not correlated with impact estimates.

Fourth, whether financial group–based or individual rewards are preferable depends on institutional context. One review of the physician pay-for-performance literature concluded that group-based incentives were preferable (Eijkenaar 2013, pp. 122–23). Two recent RCTs found positive effects of facility-level financial incentives on institutional baby deliveries, preventive care visits for infants, and newborn health outcomes such as birth weight for Rwandan healthcare facilities (Basinga et al. 2011, Gertler & Vermeersch 2012). One RCT reported a positive effect of team-level incentives on tax revenue collected for Pakistani tax collectors, though rates of tax bribery also increased (Khan et al. 2014). Difference-in-difference evidence similarly found positive effects for a team-based incentive scheme among UK tax collectors (Burgess et al. 2010), whereas evidence using matching methods to examine team-based incentives in a UK job training center found positive effects on productivity in smaller offices and negative effects in larger offices (Burgess et al. 2012). By contrast, recent evidence in teaching weakly suggests that individual incentives may be preferable. An experiment of school-level pay-for-performance in New York City found no effect on student outcomes (Fryer 2011), and one in Kenya found evidence of teaching to the test (Glewwe et al. 2010). Another experiment in India found that students in individual-incentive schools outperformed those in group-incentive schools after two years (Muralidharan & Sundararaman 2011, p. 41). Further, two RCTs in the United States differed on the effect of individual teacher incentives on student achievement (Fryer et al. 2012, Springer et al. 2011). Interestingly, Fryer et al. (2012) attributed the positive effect of pay-for-performance to designing the incentive function around loss aversion (i.e., paying teachers in advance, but requiring a refund if students failed to improve sufficiently).

6. EX POST INITIATIVES

6.1. Auditing

Audits may also facilitate observation of quality. Assessments of the effectiveness of audits are limited to qualitative case studies in the United States. An Inspector General review found that most SSA ALJs were not notified of the quality review results of their decisions, and that the SSA did not maintain data on the review results by ALJ or hearing office (Off. Insp. Gen. 2012). Gelbach & Marcus (2016, pp. 118–21) documented that some regional offices have implemented more effective deliberation over remand orders, although such practices may reflect preexisting differences in quality and management. Schwartz (2010, pp. 1067–71) documented case studies suggesting that when police departments affirmatively incorporate information from audits, lawsuits, and officer misconduct, police conduct appears to improve.

The most rigorous evidence of audits stems from the development context. We caution that generalizability to the United States may be particularly questionable in this context, owing to substantial cross-national differences in the potential bribability of civil servants. Nonetheless, two lessons seem worth considering from the development context. First, stakeholder involvement in audits appears questionable. Three RCTs differed on the effectiveness of community monitoring, which is akin to auditing by stakeholders in our framework (Banerjee et al. 2010, 2012; Svensson & Björkman 2009). A field experiment by Bjorkman Nyqvist et al. (2014) examining the long-run

impact of community monitoring in Uganda found that information disclosure to stakeholders played a key role in the program's success, suggesting that auditing can blend into disclosure.

Olken (2007) conducted RCTs with over 600 Indonesian village road projects, finding that increased government audits reduced missing expenditures by 8%, but that adding stakeholder participation had little effect. Olken, however, did not examine a long time horizon and speculated that repeat relationships with auditors may facilitate bribing, and hence suggested auditor rotation (p. 244). Randomizing inspectors to establishments was implemented in New York City's health department after repeat relationships led to the indictment of over half of the health department for extortion in the 1980s (Kurtz 1988).

Second, numerous studies show that corruption plagues auditing, leading to underreporting of violations. Duflo et al. (2013a) found that overall auditing quality of environmental auditors in the Indian state of Gujarat was very low owing to rampant underreporting, although some auditors performed well. Duflo et al. (2013b) showed in a two-year RCT that in lieu of payment by firms, payment of auditors from a central money pool reduced underreporting, forcing more treatment plants to reduce pollution. However, an RCT by Duflo et al. (2014) found that increasing audit frequency only marginally increased compliance with pollution standards; overall, pollution emissions did not significantly reduce, and the only polluters affected were those near the regulatory threshold (p. 3).

6.2. Internal Appeals

In theory, appeals provide a mechanism by which a claimant can directly contest a determination. Yet internal agency appeal systems have long been criticized. Qualitative studies have focused on the SSA appeals system in particular. Cofer (1985b, p. 13) concluded, "The arguments are persuasive that the $18 million a year expense of the [Appeals Council] could be put to better use." Koch & Koplow (1989, p. 296) found that the Appeals Council system was "broken," had "deep, permanent flaws" and was "wholly unsatisfactory." Gelbach & Marcus (2016 , p. 16) studied SSA hearing offices and found that (*a*) dissemination of information in remand orders to ALJs can vary dramatically and (*b*) many federal judges handling SSA appeals exhibit little awareness of the agency process. In one hearing office, ALJs reported an improvement in decision making owing to semiannual memoranda summarizing district court social security decisions (pp. 119–20).

Over the last decade, empirical studies have uncovered evidence of inconsistency and/or ineffectiveness in internal appeals processes. Hausman found that the Social Security Appeals Council promoted consistency only in a limited sense: Although claimants were more likely to appeal the decisions of harsher judges, decisions by outlier judges were not disproportionately more likely to be reversed (D. Hausman, unpublished manuscript). In the regulatory context, Ho (2012, pp. 667–70) found no appreciable difference in New York food safety inspection score consistency prior to the appeal hearing or posthearing. In immigration adjudication, Hausman (2016) found that the Board of Immigration Appeals (BIA) (as well as judicial review) failed to promote uniformity because cases selected for review were unrepresentative. Miller et al. (2014, p. 139) found that higher BIA workload was associated with greater deference to immigration judges (IJs). They also note, "The easiest way for IJs to insulate their decisions from review is to simply grant some form of relief to the applicant" (p. 149).

Some argue, however, that internal appeals still hold more promise than judicial review. Despite reservations ("the impact of the Appeals Council . . . has been seriously compromised by the low esteem in which it is held by the ALJ corps"), Mashaw et al. (1978), for instance, opposed eliminating the Appeals Council (pp. 106–7). Because inconsistency also plagues judicial review (Hausman 2016, Ramji-Nogales et al. 2007), the real question is one about relative institutional

competence. For instance, Ray & Lubbers (2014) discuss one promising case study of internal appeals, describing a more data-driven effort by the SSA to identify errors and improve the quality of disability case review.

7. IMPLICATIONS

The empirical literature on performance measurement and quality management in the public sector is growing fast. We hope to have provided a useful overview of these areas, and we conclude with several implications.

First, our typology reveals that there is much synergy between conventionally separate bodies of scholarship. Barnow's (2000) evidence that performance measurement says little about program impact, for instance, almost surely travels to structurally similar settings, such as disclosure, monitoring, and auditing. In addition, the typology opens substantial policy space for agencies to blend and tailor quality initiatives to local constraints; for instance, if auditing is untenable owing to staff resistance, borrow the collaborative elements of peer review.

Second, ongoing management techniques, such as monitoring, peer review, and pay-for-performance, appear to be more successful than ex post techniques. A related theme is that proactive managerial intervention and commitment are critical to the success of quality efforts. Just as judicial review can sound the alarm too late, ex post review can be too little, too late. The most effective interventions involve managerial commitment to address street-level arbitrariness.

Third, despite New Governance's allure, the quantity–quality trade-off continues to pervade each of these management techniques. In particular, techniques that prioritize quantitative indicators (e.g., disclosure, pay-for-performance)—particularly when the dimensionality of work product may not be captured by extant data—are most seriously prone to perverse effects. One alternative is to shift to management techniques that pose a less acute trade-off (e.g., monitoring, training, auditing, peer review). Another alternative consists of better quality measurement for purposes of disclosure and pay schemes. The results from peer review and audits, for instance, can themselves constitute indicators of the level of consistency. But such informed measurement may be more involved than "the simple act of defining measures" that Osborne & Gaebler (1993, p. 147) envisioned.[15]

Fourth, although scholarship would ideally provide empirically grounded answers to the policy maker interested in effectively promoting quality of service delivery, the evidence base remains sorely limited in two ways. Methodologically, well-designed studies are lacking. As a potential indicator of publication bias, we also note that RCTs, which owing to scale and cost are far less likely to remain in a file drawer, exhibit far more null results than observational studies. More attention to rigorous policy evaluation is desperately needed in the field. Substantively, the best-designed studies are limited to a small number of domains: job training, healthcare, and education. This is worrisome, because the lessons distilled in this article may not generalize to the myriad of regulatory domains—nuclear safety, child welfare, occupational safety, disability adjudication, and environmental safety, to name just a few. Many of the best-designed studies are in the development context. As a result, when it comes to understanding institutional design to address street-level arbitrariness, as a relative matter, US administrative law remains in the dark.

Last, one of the most promising paths forward lies in academic-agency collaborations to identify, craft, and evaluate the most effective interventions for improving quality of decisions. The

[15] Osborne & Gaebler (1993, pp. 349–59) do anticipate problems such as cream-skimming, perverse incentives, resistance, and the need to balance quantitative and qualitative assessment, but the popularization of these techniques has often glossed over these challenges.

current state of affairs is that quality improvement initiatives may be lauded on government websites and in glossy annual reports, but little attention is paid to serious evaluation. As one colleague would put it, current quality efforts amount to massive nonrandomized experiments without control groups and an effective sample size of one. The result is that these efforts stifle learning in the administrative state.

Academic-agency collaborations (e.g., Duflo et al. 2013a, Ho 2017) provide a path forward: Academics have the research capacity both to draw on the existing evidence base to design interventions and to evaluate them in cost-effective ways; agencies in the face of such stark disparities in decision making have a myriad of opportunities to implement pilots for quality improvement. Although the evidence base for managing the quality of bureaucratic decisions has developed rapidly, the data revolution has also barely scratched the surface of agency decision-making processes. The tools, resources, and academic partners exist to build a firmer basis for managing street-level arbitrariness. Failing to capitalize on these opportunities would leave unanswered the due process question Jerry Mashaw raised so vividly 45 years ago. And agencies would remain again confined to muddle through (Lindblom 1959).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Thanks to David Hausman, Dave Marcus, and Chuck Sabel for helpful comments and conversations.

LITERATURE CITED

An LC, Bluhm JH, Foldes SS, Alesci NL, Klatt CM, et al. 2008. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. *Arch. Intern. Med.* 168(18):1993–99

Asch BJ. 1990. Do incentives matter? The case of Navy recruiters. *Ind. Labor Relat. Rev.* 43(3):89S–106S

Ashraf N, Bandiera O, Lee SS. 2014. Awards unbundled: evidence from a natural field experiment. *J. Econ. Behav. Organ.* 100:44–63

Banerjee AV, Banerji R, Duflo E, Glennerster R, Khemani S. 2010. Pitfalls of participatory programs: evidence from a randomized evaluation in education in India. *Am. Econ. J. Econ. Policy* 2(1):1–30

Banerjee AV, Chattopadhyay R, Duflo E, Keniston D, Singh N. 2012. *Can institutions be reformed from within? Evidence from a randomized experiment with the Rajasthan police*. Work. Pap. 12-04, Mass. Inst. Technol., Cambridge, MA

Banerjee AV, Duflo E, Glennerster R. 2008. Putting a Band-Aid on a corpse: incentives for nurses in the Indian public health care system. *J. Eur. Econ. Assoc.* 6(2–3):487–500

Barankay I. 2012. *Rank incentives: evidence from a randomized workplace experiment*. Discuss. Pap., Univ. Penn., Philadelphia, PA

Bardach E, Kagan RA. 1982. *Going by the Book: The Problem of Regulatory Unreasonableness*. Philadelphia: Temple Univ. Press

Barnow BS. 2000. Exploring the relationship between performance management and program impact: a case study of the Job Training Partnership Act. *J. Policy Anal. Manag.* 19(1):118–41

Barnow BS, Heinrich CJ. 2010. One standard fits all? The pros and cons of performance standard adjustments. *Public Adm. Rev.* 70(1):60–71

Barnow BS, Smith JA. 2004. Performance management of US job training programs. In *Job Training Policy in the United States*, ed. CJ O'Leary, RA Straits, SA Wandner, pp. 21–56. Kalamazoo, MI: WE Upjohn Inst.

Basinga P, Gertler PJ, Binagwaho A, Soucat AL, Sturdy J, Vermeersch CM. 2011. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *Lancet* 377(9775):1421–28

Behn RD. 2003. Why measure performance? Different purposes require different measures. *Public Adm. Rev.* 63(5):586–606

Ben-Shahar O, Schneider CE. 2014. *More Than You Wanted to Know: The Failure of Mandated Disclosure*. Princeton, NJ: Princeton Univ. Press

Bernstein ES. 2012. The transparency paradox: a role for privacy in organizational learning and operational control. *Adm. Sci. Q.* 57(2):181–216

Besley TJ, Bevan G, Burchardi K. 2009. *Naming & shaming: the impacts of different regimes on hospital waiting times in England and Wales*. Discuss. Pap., Cent. Econ. Policy Res., London

Bjorkman Nyqvist M, De Walque D, Svensson J. 2014. *Information is power: experimental evidence on the long-run impact of community based monitoring*. Policy Res. Work. Pap. 7015, World Bank, Washington, DC

Blanes i Vidal J, Nossol M. 2011. Tournaments without prizes: evidence from personnel records. *Manag. Sci.* 57(10):1721–36

Bol JC, Smith SD. 2011. Spillover effects in subjective performance evaluation: bias and the asymmetric influence of controllability. *Account. Rev.* 86(4):1213–30

Bouguen A. 2016. Adjusting content to individual student needs: further evidence from an in-service teacher training program. *Econ. Educ. Rev.* 50:90–112

Bowman CL, McCormick S. 2000. Comparison of peer coaching versus traditional supervision effects. *J. Educ. Res.* 93(4):256–61

Braithwaite J, Braithwaite V. 1995. The politics of legalism: rules versus standards in nursing-home regulations. *Soc. Legal Stud.* 4:307–41

Burgess S, Propper C, Ratto M, Scholder K, von Hinke S, Tominey E. 2010. Smarter task assignment or greater effort: the impact of incentives on team performance. *Econ. J.* 120(547):968–89

Burgess S, Propper C, Ratto M, Tominey E. 2012. *Incentives in the public sector: evidence from a government agency*. Discuss. Pap. 6738, Inst. Study Labor, Bonn, Ger.

Burgess S, Wilson D, Worth J. 2013. A natural experiment in school accountability: the impact of school performance information on pupil progress. *J. Public Econ.* 106:57–67

Callen M, Gulzar S, Hasanain A, Khan Y, Rezaee A. 2015. *Personalities and public sector performance: evidence from a health experiment in Pakistan*. Work. Pap. 21180, Natl. Bur. Econ. Res., Cambridge, MA

Casalino LP, Gans D, Weber R, Cea M, Tuchovsky A, et al. 2016. US physician practices spend more than $15.4 billion annually to report quality measures. *Health Aff.* 35(3):401–6

Chakrabarti R. 2013. Accountability with voucher threats, responses, and the test-taking population: regression discontinuity evidence from Florida. *Education* 8(2):121–67

Chen D, Kremer M. 2002. *Interim report on a teacher attendance incentive program in Kenya*. Manuscr., Harvard Univ., Cambridge, MA. **http://users.nber.org/~dlchen/papers/Interim_Report_on_a_Teacher_Attendance_Incentive_Program_in_Kenya.pdf**

Chen T-T, Chung K-P, Lin I, Lai M-S. 2011. The unintended consequence of diabetes mellitus pay-for-performance (P4P) program in Taiwan: Are patients with more comorbidities or more severe conditions likely to be excluded from the P4P program? *Health Serv. Res.* 46(1):47–60

Cofer DP. 1985a. The question of independence continues: administrative law judges within the Social Security Administration. *Judicature* 69:228–35

Cofer DP. 1985b. *Judges, Bureaucrats, and the Question of Independence: A Study of the Social Security Administration Hearing Process*. Westport, CT: Greenwood

Courty P, Kim DH, Marschke G. 2011. Curbing cream-skimming: evidence on enrolment incentives. *Labour Econ.* 18(5):643–55

Courty P, Marschke G. 1997. Measuring government performance: lessons from a federal job-training program. *Am. Econ. Rev.* 87(2):383–88

Courty P, Marschke G. 2004. An empirical investigation of gaming responses to explicit performance incentives. *J. Labor Econ.* 22(1):23–56

Cuéllar M-F. 2006. Auditing executive discretion. *Notre Dame Rev.* 82:227–311

Cullen JB, Reback R. 2006. *Tinkering toward accolades: school gaming under a performance accountability system*. Work Pap. 12286, Natl. Bur. Econ. Res., Cambridge, MA

Dhaliwal I, Hanna R. 2014. *Deal with the devil: the successes and limitations of bureaucratic reform in India*. Work. Pap. 20482, Natl. Bur. Econ. Res., Cambridge, MA

Dixit A. 2002. Incentives and organizations in the public sector: an interpretative review. *J. Hum. Resour.* 37(4):696–727

Dowling B, Richardson R. 1997. Evaluating performance-related pay for managers in the National Health Service. *Int. J. Hum. Resour. Manag.* 8(3):348–66

Dranove D, Kessler D, McClellan M, Satterthwaite M. 2003. Is more information better? The effects of "report cards" on health care providers. *J. Political Econ.* 111(3):555–88

Duflo E, Greenstone M, Pande R, Ryan N. 2013a. What does reputation buy? Differentiation in a market for third-party auditors. *Am. Econ. Rev.* 103(3):314–19

Duflo E, Greenstone M, Pande R, Ryan N. 2013b. Truth-telling by third-party auditors and the response of polluting firms: experimental evidence from India. *Q. J. Econ.* 128(4):1499–545

Duflo E, Greenstone M, Pande R, Ryan N. 2014. *The value of regulatory discretion: estimates from environmental inspections in India*. Work. Pap. 20590, Natl. Bur. Econ. Res., Cambridge, MA

Duflo E, Hanna R, Ryan SP. 2012. Incentives work: getting teachers to come to school. *Am. Econ. Rev.* 102(4):1241–78

Eijkenaar F. 2013. Key issues in the design of pay for performance programs. *Eur. J. Health Econ.* 14(1):117–31

Faber M, Bosch M, Wollersheim H, Leatherman S, Grol R. 2009. Public reporting in health care: How do consumers use quality-of-care information? A systematic review. *Med. Care* 47(1):1–8

Fairbrother G, Hanson KL, Friedman S, Butts GC. 1999. The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates. *Am. J. Public Health* 89(2):171–75

Figlio DN, Getzler LS. 2006. Accountability, ability and disability: Gaming the system? *Adv. Appl. Microecon.* 14:35–49

Figlio DN, Rouse CE. 2006. Do accountability and voucher threats improve low-performing schools? *J. Public Econ.* 90(1):239–55

Finan F, Olken BA, Pande R. 2015. *The personnel economics of the state*. Work. Pap. 21825, Natl. Bur. Econ. Res., Cambridge, MA

Forsetlund L, Bjorndal A, Rashidian A, Jamtvedt G, O'Brien MA, et al. 2009. Continuing education meetings and workshops: effects on professional practice and health care outcomes. *Cochrane Database Syst. Rev.* 2(2):CD003030

Fryer RG. 2011. *Teacher incentives and student achievement: evidence from New York City public schools*. Work. Pap. 16850, Natl. Bur. Econ. Res., Cambridge, MA

Fryer RG Jr., Levitt SD, List J, Sadoff S. 2012. *Enhancing the efficacy of teacher incentives through loss aversion: a field experiment*. Work. Pap. 18237, Natl. Bur. Econ. Res., Cambridge, MA

Fung A, Graham M, Weil D. 2007. *Full Disclosure: The Perils and Promise of Transparency*. Cambridge, UK: Cambridge Univ. Press

Fung CH, Lim Y-W, Mattke S, Damberg C, Shekelle PG. 2008. Systematic review: the evidence that publishing patient care performance data improves quality of care. *Ann. Intern. Med.* 148(2):111–23

Gattellari M, Donnelly N, Taylor N, Meerkin M, Hirst G, Ward JE. 2005. Does "peer coaching" increase GP capacity to promote informed decision making about PSA screening? A cluster randomised trial. *Fam. Pract.* 22(3):253–65

Gawande A. 2010. *The Checklist Manifesto: How to Get Things Right*. London: Macmillan

Gelbach JB, Marcus D. 2016. *A study of social security disability litigation in the federal courts*. Final Rep., Adm. Conf. US, Washington, DC. **https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2821861**

Gerken HK. 2009. *The Democracy Index: Why Our Election System Is Failing and How to Fix It*. Princeton, NJ: Princeton Univ. Press

Gertler PJ, Vermeersch C. 2012. *Using performance incentives to improve health outcomes*. Policy Res. Work. Pap. 6100, World Bank, Washington, DC

Glazerman S, Seifullah A. 2012. *An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years*. Final Rep., Math. Policy Res., Washington, DC

Glewwe P, Ilias N, Kremer M. 2010. Teacher incentives. *Am. Econ. J. Appl. Econ.* 2(3):205–27
Goldstein J. 2007. Easy to dance to: solving the problems of teacher evaluation with peer assistance and review. *Am. J. Educ.* 113(3):479–508
Greenberg DH, Michalopoulos C, Robins PK. 2003. A meta-analysis of government-sponsored training programs. *Ind. Labor Relat. Rev.* 57(1):31–53
Greenberg DH, Michalopoulos C, Robins PK. 2006. Do experimental and nonexperimental evaluations give different answers about the effectiveness of government-funded training programs? *J. Policy Anal. Manag.* 25(3):523–52
Haga BI, Richman R, Leavitt W. 2010. System failure: implementing pay for performance in the Department of Defense's National Security Personnel System. *Public Pers. Manag.* 39(3):211–30
Hannan EL, Kilburn H, Racz M, Shields E, Chassin MR. 1994. Improving the outcomes of coronary artery bypass surgery in New York State. *JAMA* 271(10):761–66
Hannan EL, Siu AL, Kumar D, Racz M, Pryor DB, Chassin MR. 1997. Assessment of coronary artery bypass graft surgery performance in New York: Is there a bias against taking high-risk patients? *Med. Care* 35(1):49–56
Hasnain Z, Pierskalla JH, Manning N. 2012. *Performance-related pay in the public sector: a review of theory and evidence*. Policy Res. Work. Pap. 6043, World Bank, Washington, DC
Hausman D. 2016. The failure of immigration appeals. *Univ. Pa. Law Rev.* 164(5):1177–238
Heckman J, Heinrich C, Smith J. 2002. *The performance of performance standards*. Work. Pap. 9002, Natl. Bur. Econ. Res., Cambridge, MA
Heckman JJ, Heinrich CJ, Smith J. 2011. Do short-run performance measures predict long-run impacts? In *The Performance of Performance Standards*, ed. JJ Heckman, CJ Heinrich, pp. 273–304. Kalamazoo, MI: WE Upjohn Inst.
Heckman JJ, Smith JA. 2003. *The determinants of participation in a social program: Evidence from a prototypical job training program*. Work. Pap. 9818, Natl. Bur. Econ. Res., Cambridge, MA
Heckman JJ, Smith JA, Taber C. 1996. *What do bureaucrats do? The effects of performance standards and bureaucratic preferences on acceptance into the JTPA program*. Work. Pap. 5535, Natl. Bur. Econ. Res., Cambridge, MA
Heinrich CJ. 2007. Evidence-based policy and performance management challenges and prospects in two parallel movements. *Am. Rev. Public Adm.* 37(3):255–77
Heinrich CJ. 2012. How credible is the evidence, and does it matter? An analysis of the Program Assessment Rating Tool. *Public Adm. Rev.* 72(1):123–34
Hibbard JH, Jewett JJ, Legnini MW, Tusler M. 1997. Choosing a health plan: Do large employers use the data? *Health Aff.* 16(6):172–80
Hibbard JH, Stockard J, Tusler M. 2003. Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff.* 22(2):84–94
Hill HC, Beisiegel M, Jacob R. 2013. Professional development research consensus, crossroads, and challenges. *Educ. Res.* 42(9):476–87
Hillman AL, Ripley K, Goldfarb N, Nuamah I, Weiner J, Lusk E. 1998. Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care. *Am. J. Public Health* 88(11):1699–701
Hillman AL, Ripley K, Goldfarb N, Weiner J, Nuamah I, Lusk E. 1999. The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care. *Pediatrics* 104(4):931–35
Ho DE. 2012. Fudging the nudge: information disclosure and restaurant grading. *Yale Law J.* 122(3):522–688
Ho DE. 2017. Does peer review work? An experiment of experimentalism. *Stanford Law Rev.* 69:1–119
Holmstrom B, Milgrom P. 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *J. Law Econ. Organ.* 7:24–52
Hysong SJ. 2009. Meta-analysis: audit & feedback features impact effectiveness on care quality. *Med. Care* 47(3):356–63
Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, et al. 2012. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst. Rev.* 2012(6):CD000259. **https://doi.org/10.1002/14651858.CD000259.pub3**
Jacob BA, Lefgren L. 2005. *Principals as agents: Subjective performance measurement in education*. Work. Pap. 11463, Natl. Bur. Econ. Res., Cambridge, MA

Jacob BA, Levitt SD. 2003. *Rotten apples: an investigation of the prevalence and predictors of teacher cheating*. Work. Pap. 9413, Natl. Bur. Econ. Res., Cambridge, MA

Jamtvedt G, Young JM, Kristoffersen DT, O'Brien MA, Oxman AD. 2007. Audit and feedback: effects on professional practice and health care outcomes. *Cochrane Database Syst. Rev.* 2007(1):CD000260

Kerr S. 1975. On the folly of rewarding A, while hoping for B. *Acad. Manag. J.* 18(4):769–83

Ketelaar NA, Faber MJ, Flottorp S, Rygh LH, Deane KH, Eccles MP. 2011. Public release of performance data in changing the behaviour of healthcare consumers, professionals or organisations. *Cochrane Database Syst. Rev.* 2011(11):CD004538

Khan AQ, Khwaja AI, Olken BA. 2014. *Tax farming redux: experimental evidence on performance pay for tax collectors*. Work. Pap. 20627, Natl. Bur. Econ. Res., Cambridge, MA

Koch CH Jr., Koplow DA. 1989. The fourth bite at the apple: a study of the operation and utility of the Social Security Administration's Appeals Council. *Fla. State Univ. Law Rev.* 17:199–324

Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, et al. 1998. Performance-based physician reimbursement and influenza immunization rates in the elderly. *Am. J. Prev. Med.* 14(2):89–95

Kravchuk RS, Schack RW. 1996. Designing effective performance-measurement systems under the Government Performance and Results Act of 1993. *Public Adm. Rev.* 56(4):348–58

Krieg JM. 2008. Are students left behind? The distributional effects of the No Child Left Behind Act. *Education* 3(2):250–81

Kurtz H. 1988. 28 New York City restaurant inspectors accused of extortion. *Washington Post*, March 25. **https://www.washingtonpost.com/archive/politics/1988/03/25/28-new-york-city-restaurant-inspectors-accused-of-extortion/7cceaf7f-659b-4f27-b6f2-694322701276/?utm_term=.5adc593c9baf**

Lavy V. 2009. Performance pay and teachers' effort, productivity, and grading ethics. *Am. Econ. Rev.* 99(5):1979–2021

Lindblom CE. 1959. The science of "muddling through." *Public Adm. Rev.* 19(2):79–88

Lipsky M. 1983. *Street-Level Bureaucracy: The Dilemmas of the Individual in Public Service*. New York City: Russell Sage Found.

Longo DR, Land G, Schramm W, Fraas J, Hoskins B, Howell V. 1997. Consumer reports in health care: Do they make a difference in patient care? *JAMA* 278(19):1579–84

Mallette B, Maheady L, Harper GF. 1999. The effects of reciprocal peer coaching on preservice general educators' instruction of students with special learning needs. *Teach. Educ. Spec. Educ.* 22(4):201–16

Mansouri M, Lockyer J. 2007. A meta-analysis of continuing medical education effectiveness. *J. Contin. Educ. Health Prof.* 27(1):6–15

Marschke G. 2002. *Performance incentives and organizational behavior: evidence from a federal bureaucracy*. Work. Pap., Univ. Albany, NY

Marshall M, Shekelle P, Brook R, Leatherman S. 2000. *Dying to Know: Public Release of Information about Quality of Health Care*. London: Nuffield Trust

Mashaw JL. 1973. Management side of due process: some theoretical and litigation notes on the assurance of accuracy fairness and timeliness in the adjudication of social welfare claims. *Cornell Rev.* 59:772–824

Mashaw JL, Goetz CJ, Goodman FI, Schwartz WF, Verkuil PF, Carrow MM. 1978. *Social Security Hearings and Appeals: A Study of the Social Security Administration Hearing System*. Lanham, MD: Lexington Books

Mason D, Hillenbrand C, Money K. 2014. Are informed citizens more trusting? Transparency of performance data and trust towards a British police force. *J. Bus. Ethics* 122(2):321–41

McDonald R, Roland M. 2009. Pay for performance in primary care in England and California: comparison of unintended consequences. *Ann. Fam. Med.* 7(2):121–27

Miller B, Keith LC, Holmes JS. 2014. *Immigration Judges and U.S. Asylum Policy*. Philadelphia: Univ. Pa. Press

Morgan RL, Gustafson KJ, Hudson PJ, Salzberg CL. 1992. Peer coaching in a preservice special education program. *Teach. Educ. Spec. Educ.* 15(4):249–58

Muralidharan K, Sundararaman V. 2011. Teacher performance pay: experimental evidence from India. *J. Political Econ.* 119(1):39–77

Nagin DS, Rebitzer JB, Sanders S, Taylor LJ. 2002. Monitoring, motivation, and management: the determinants of opportunistic behavior in a field experiment. *Am. Econ. Rev.* 92(4):850–73

Neal D, Schanzenbach DW. 2010. Left behind by design: proficiency counts and test-based accountability. *Rev. Econ. Stat.* 92(2):263–83

Noonan KG, Sabel CF, Simon WH. 2009. Legal accountability in the service-based welfare state: lessons from child welfare reform. *Law Soc. Inq.* 34(3):523–68

Oakland JS. 2007. *Statistical Process Control*. Abingdon, UK: Routledge

Off. Insp. Gen. 2012. *The Social Security Administration's review of administrative law judges' decisions*. Rep. A-07-12- 21234, Soc. Secur. Adm., Baltimore, MD

Olken BA. 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *J. Political Econ.* 115(2):200–49

Osborne D, Gaebler T. 1993. *Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector*. New York: Plume

Parrillo NR. 2013. *Against the Profit Motive: The Salary Revolution in American Government, 1780–1940*. New Haven, CT: Yale Univ. Press

Pattanayak SK, Yang J-C, Dickinson KL, Poulos C, Patil SR, et al. 2009. Shame or subsidy revisited: social mobilization for sanitation in Orissa, India. *Bull. World Health Organ.* 87(8):580–87

Perez T, Rushing R. 2007. *The CitiStat model: how data-driven government can increase efficiency & effectiveness*. Rep., Cent. Am. Prog., Washington, DC

Peterson ED, DeLong ER, Jollis JG, Muhlbaier LH, Mark DB. 1998. The effects of New York's bypass surgery provider profiling on access to care and patient outcomes in the elderly. *J. Am. Coll. Cardiol.* 32(4):993–99

Pierce L, Snow DC, McAfee A. 2015. Cleaning house: the impact of information technology monitoring on employee theft and productivity. *Manag. Sci.* 61(10):2299–319

Propper C, Sutton M, Whitnall C, Windmeijer F. 2010. Incentives and targets in hospital care: evidence from a natural experiment. *J. Public Econ.* 94(3):318–35

Radin BA. 2006. *Challenging the Performance Movement: Accountability, Complexity, and Democratic Values*. Washington, DC: Georgetown Univ. Press

Ramji-Nogales J, Schoenholtz AI, Schrag PG. 2007. Refugee roulette: disparities in asylum adjudication. *Stanford Law Rev.* 60:295–411

Ray GK, Lubbers JS. 2014. A government success story: how data analysis by the Social Security Appeals Council (with a push from the Administrative Conference of the United States) is transforming social security disability adjudication. *George Wash. Law Rev.* 83:1575–608

Reback R. 2008. Teaching to the rating: school accountability and the distribution of student achievement. *J. Public Econ.* 92(5):1394–415

Rockoff JE, Staiger DO, Kane TJ, Taylor ES. 2012. Information and employee evaluation: evidence from a randomized intervention in public schools. *Am. Econ. Rev.* 102(7):3184–213

Roski J, Jeddeloh R, An L, Lando H, Hannan P, et al. 2003. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. *Prev. Med.* 36(3):291–99

Sabel CF, Simon WH. 2017. The management side of due process in the service-based welfare state. In *Administrative Law from the Inside Out: Essays on Themes in the Work of Jerry L. Mashaw*, ed. NR Parillo, pp. 63–86. Cambridge, UK: Cambridge Univ. Press

Schneider EC, Epstein AM. 1998. Use of public performance reports: a survey of patients undergoing cardiac surgery. *JAMA* 279(20):1638–42

Schochet PZ, Fortson J. 2012. *Do regression-adjusted performance measures for workforce development programs track longer-term program impacts? A case study for job corps*. Work. Pap., Math. Policy Res.

Schwartz JC. 2010. Myths and mechanics of deterrence. *UCLA Law Rev.* 57:1023–94

Schwellnus H, Carnahan H. 2014. Peer-coaching with health care professionals: What is the current status of the literature and what are the key components necessary in peer-coaching? A scoping review. *Med. Teach.* 36(1):38–46

Scott A, Sivey P, Ait Ouakrim D, Willenberg L, Naccarella L, et al. 2011. The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database Syst. Rev.* 9(9):CD008451

Secomb J. 2008. A systematic review of peer teaching and learning in clinical education. *J. Clin. Nurs.* 17(6):703–16

Shekelle P, Lim Y-W, Mattke S, Damberg C. 2008. *Does public release of performance results improve quality of care*. Syst. Rev. Health Found., Lond.

Song H, Tucker AL, Murrell KL, Vinson DR. 2015. *Public relative performance feedback in complex service systems: improving productivity through the adoption of best practices*. Res. Pap. Ser. 16–043, Harvard Bus. School, Harvard Univ., Boston, MA

Spence GB, Grant AM. 2007. Professional and peer life coaching and the enhancement of goal striving and well-being: an exploratory study. *J. Posit. Psychol.* 2(3):185–94

Springer MG, Ballou D, Hamilton L, Le V-N, Lockwood JR, et al. 2011. *Teacher pay for performance: experimental evidence from the Project on Incentives in Teaching (POINT)*. Rep., Soc. Res. Educ. Eff., Evanston, IL

Staats BR, Dai H, Hofmann D, Milkman KL. 2016. Motivating process compliance through individual electronic monitoring: an empirical examination of hand hygiene in healthcare. *Manag. Sci.* 63:1563–85

Stajkovic AD, Luthans F. 2003. Behavioral management and task performance in organizations: conceptual background, meta-analysis, and test of alternative models. *Pers. Psychol.* 56(1):155–94

Sue-Chan C, Latham GP. 2004. The relative effectiveness of external, peer, and self-coaches. *Appl. Psychol.* 53(2):260–78

Svensson J, Björkman M. 2009. Power to the people: evidence from a randomized field experiment of a community-based monitoring project in Uganda. *Q. J. Econ.* 124(2):735–69

Thaler RH, Sunstein CR. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale Univ. Press

US Food Drug Adm. 2015. *Standardization of retail food safety inspection personnel*. Regul., US Food Drug Adm., Silver Spring, MD. **https://www.fda.gov/Food/GuidanceRegulation/RetailFoodProtection/Standardization/default.htm**

Weber M. 1968. *Economy & Society: An Outline of Interpretive Sociology*, Vol. 3. New York: Bedminister

Weibel A, Rost K, Osterloh M. 2010. Pay for performance in the public sector—benefits and (hidden) costs. *J. Public Adm. Res. Theory.* 20(2):387–412

Wilson JQ. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books

Yang CS. 2014. Have interjudge sentencing disparities increased in an advisory guidelines regime-evidence from *Booker*. *N.Y. Univ. Law Rev.* 89:1268–342

Yu T-C, Wilson NC, Singh PP, Lemanu DP, Hawken SJ, Hill AG. 2011. Medical students-as-teachers: a systematic review of peer-assisted teaching during medical school. *Adv. Med. Educ. Pract.* 2:157–72

Annual Review of Law and Social Science

Volume 13, 2017

Contents