Quality Review of Mass Adjudication: A Randomized Natural Experiment at the Board of Veterans Appeals, 2003–16

Daniel E. Ho*
Stanford University

Cassandra Handan-Nader
Stanford University

David Ames
Bergmann and Moore, LLC

David Marcus
University of California, Los Angeles

We study a unique natural experiment, during which 5–10% of draft opinions by judges of the Board of Veterans Appeals (BVA) were randomly selected for "quality review (QR)" by a team of full-time staff attorneys. The express goals of this performance program were to measure accuracy and reduce remands on appeal. In cases of legal error, the QR team wrote memoranda to judges for correction of draft opinions. We use rich internal administrative data on nearly 600,000 cases from 2003 to 2016 to conduct the first rigorous evaluation of this program. With precise estimates, we show that QR had no appreciable effects on appeals or remands. Based on internal records, we demonstrate that this inefficacy is likely by design, as meeting the performance measure of "accuracy" conflicted with error correction. These findings inform longstanding questions of law, organization, and bureaucracy, including performance management, standards of review, and institutional design of adjudication. (*JEL* K23, K40, H11)

1. Introduction

A hallmark feature of the American administrative state is mass adjudication. Each year, more than 1500 administrative law judges (ALJs) in the

We thank Anne McDonough, Oluchi Mbonu, and Reid Whitaker for terrific research assistance and Michael Asimow, Jacob Goldin, David Hausman, Mark Krass, Daryl Levinson, Rob MacCoun, Anne Joseph O'Connell, Nick Parrillo, James Ridgway, Sam Sherman, Bill Simon, Angela Teuscher, Zac Townsend, and participants at the faculty workshop at the Northwestern Pritzger School of Law, the faculty workshop at Stanford Law School, and the Munro Distinguished Lecture in Stanford's political science department for helpful comments and conversations.

*Stanford Law School, Stanford University, Stanford, CA, USA. Email: dho@law.stanford.edu.

The Journal of Law, Economics, and Organization, Vol. 35, No. 2
doi:10.1093/jleo/ewz001
Advance Access published March 29, 2019

Social Security Administration (SSA) resolve over 600,000 disability and social security appeals, more than the caseload of all US federal district courts combined. In the Executive Office for Immigration Review, roughly 300 immigration judges process over 270,000 cases, grappling with a backlog of nearly 700,000 cases. And in the Board of Veterans' Appeals (BVA or Board), some 90 veterans law judges (VLJs) decide 50,000 cases, with over 1000 cases docketed per VLJ, annually.

Decades of scholarship have pointed to severe challenges in the effectiveness, accuracy, and consistency of such mass administrative justice (see, e.g., Mashaw et al. 1978; Mashaw 1983; Braithwaite and Braithwaite 1995; Ramji-Nogales et al. 2007; Noonan et al. 2009; Asimow 2016; Ho 2017). In landmark studies, Mashaw et al. (1978) and Mashaw (1983) documented dramatic disparities in how SSA ALJs adjudicated comparable cases. Mashaw (1973a, 1983) argued these failures amounted to a constitutional due process problem, requiring an internal *management* systems for quality assurance and performance management. Mashaw pointed to the VA's "statistical quality assurance control" system as one positive example (Mashaw 1973a: 793–6) and famously argued that agencies could and should internally develop such mechanisms for bureaucratic rationality (Mashaw 1983). As a theoretical matter, performance measurement and monitoring may address principal–agent problems in bureaucratic delegation (Barnow 2000; Brignall and Modell 2000; Dixit 2002; Gilmour and Lewis 2006; Ho and Sherman 2017). A quality review (QR) program, for instance, may address the information asymmetry between the principal (Congress) and agent (the agency), therefore reducing agency costs particularly for multidimensional tasks (Holmstrom and Milgrom 1991). The then-General Accounting Office (GAO), in response, called for better management of administrative adjudication (US General Accounting Office 1978; Lubbers 1993). The Administrative Conference of the United States similarly recommended using management techniques to address inter-judge disparities (Administrative Conference of the United States 1978). Yet in spite of continuing challenges in mass adjudication (Verkuil 1991, 2017; Krent and Morris 2013; Gelbach and Marcus 2016, 2018; Hausman 2016; Ho 2017) and much writing about law and management of the bureaucracy (e.g., Chassman and Rolston 1979; Simon 1983, 2006; Wilson 1991; Brodkin 2006; Metzger 2014), there exists little rigorous evidence about the effectiveness of quality management systems in adjudication specifically or the public sector generally (Margetts 2011; Greiner and Matthews 2016; Ho and Sherman 2017). To date, the evidence consists exclusively of useful but limited qualitative case studies (e.g., Brodkin and Lipsky 1983; Koch and Koplow 1990), with no systematic evidence about the causal effect of quality assurance programs (Brennan 1998; Cable 2001; Simon 2012).

We study a unique randomized natural experiment that offers rich insight into this central issue in bureaucracy, organization, and administrative law. For over 15 years, the BVA used a computer to randomly sample 5% of draft

(original) decisions by judges, subjecting these decisions to a time- and resource-intensive QR process by an independent team of full-time attorneys. For decisions remanded by the US Court of Appeals for Veterans Claims (CAVC), which hears appeals of BVA decisions, the BVA randomly sampled 10% of decisions. Attorneys analyzed draft opinions, identified legal errors, and wrote memoranda to VLJs to enable judges to correct opinions before being issued. Roughly 76% of decisions appealed to CAVC result in a remand of at least one issue, and the program was expressly designed with the dual goals of (a) reducing the remand/reversal rate of BVA decisions from CAVC and (b) measuring the accuracy of BVA decisions.

We secure internal administrative data on nearly 600,000 cases from 2003 to 2016, never before used by outside researchers, to provide the first rigorous study of the effects of this internal management system. First, we show that we are able to replicate the random case selection process for QR with high fidelity. Our ability to replicate the randomization process stems largely from the fact that we are using the same, rich internal dataset that BVA used to carry out this process. At the same time, we also rely on public records and information act requests to ensure that we are replicating the process exactly, as the GAO documented imperfections in the randomization scheme in early years (US General Accounting Office 2002).[1] The administrative data contain rich covariate information, and we show balance on over 80 dimensions, including legal representation, timing, number of legal issues, age and gender of appellant, service period, issue type, medical diagnostic codes, and disposition. The randomization hence provides a credible research design to compare "treatment" decisions subjected to QR with "control" decisions.

We study whether the program had effects on the probability that claimants appeal to the CAVC and the probability that CAVC reversed or remanded (conditional on appeal). We find that both for original and CAVC-remanded decisions, there is no appreciable benefit of QR. This is so notwithstanding the BVA's commitment of significant resources to QR. Cases that underwent QR have indistinguishable appeal, reversal, and remand rates from cases that did not. We then study whether the program affected inter-judge variability. We test and find no evidence for heterogeneous VLJ-specific treatment effects.

We then investigate the mechanism for the lack of effectiveness. We rule out that VLJs defied the advice of the QR team. Our evidence also does not support the possibility that the results are explained by arbitrariness of CAVC decisions or the QR team. To the contrary, we show that conditional on QR, the presence of an error flagged by the QR process is associated with a higher risk of a remand. This shows that the QR team was in fact able to identify low-quality *types* of opinions. But even for opinions that the QR deemed to have *no errors*, the remand rate remained a stunning 74%. As a

1. We also document a lesser known design choice of the selection process, which is the exclusion of cases by senior management from review. See Appendix B.

result, the limited corrections had no substantive effect on how a case, which typically presents numerous issues, fared on appeal.

This evidence also points to the best explanation for the program's ineffectiveness: divergence between CAVC's and BVA's standards of review. Formally, the standards were announced as the same: BVA should identify issues which would "result in the reversal or remand of a Board decision by [CAVC]." Yet internal documents secured through FOIA requests reveal that the review team in fact deployed a significantly more lenient standard. Errors were identified only when there were no "legitimate differences of opinion." A later revision of the training manual affirmed what had effectively become the QR program's operational standard: an error should only be called when "undebatable." We demonstrate this divergence empirically by comparing the rate at which the QR team called errors with CAVC's remand rate for the same error in the same cases. For the most common error, namely the failure to adequately explain a decision, we find CAVC remands at six times the rate that QR calls the error. We also show that more stringent quality reviewers are more likely to agree with CAVC's disposition in a case.

We explore the reasons for this functional divergence of standards of review. One important factor is the desire to meet BVA's performance goal of "accurate" decisions. BVA would regularly report accuracy rates of 93–95% in its Annual Reports (e.g., Board of Veterans' Appeals 2014, 2016), which were defined as a key performance measure under the Government Performance and Results Act (GPRA) and scrutinized in congressional oversight hearings (Senate Committee on Veterans' Affairs 2005; House Committee on Veterans' Affairs 2007a, 2008). The divergence hence illustrates the potential conflict when an agency can define its own performance measure under conflicting objectives.

Our setting has several virtues. First, methodologically, our study is the first to leverage randomization and large-scale administrative data to provide credible inferences about the causal effect of a QR program in the administrative state. The internal data, used by BVA to run the QR program, allow us to cleanly replicate the randomization scheme. Due to the sheer scale of the program, our estimates are also quite precise, allowing us to rule out effects of any substantial magnitude. Second, the BVA QR program exemplifies the kind of program scholars and policymakers have envisioned as curing the due process problems of mass adjudication (Mashaw 1973a; Administrative Conference of the United States 1978; US Government Accountability Office 2005; Gelbach and Marcus 2016). The review process was resource-intensive, involving a team of four to six full-time staff attorneys, with a case load exceeding that of most US district courts. By leveraging the insight of peers, our natural experiment is also related to the idea of Mashaw et al. (1978), which used simulation to calculate reversal rates if appeals were decided by panels, and Ho (2017), which found evidence in a randomized controlled trial that

peer review reduced the inter-inspector citation variance. Last, the BVA context allows us to focus on a fairly well-defined, if complex, area of law. Approximately 95% of appeals pertain to disability issues. This substantive focus means that the QR team would seem well-positioned to identify systematic errors in VLJ decision-making. And while many have pointed to the parallels between the SSA, immigration courts, and the BVA (e.g., Verkuil 1991, 2017; Congressional Research Service 2012; Asimow 2016; Sabel and Simon 2017; Gelbach and Marcus 2018), few studies have empirically examined decision-making in veterans adjudication.[2]

One potential limitation to our study is that while the design allows us to rigorously assess the impact of QR on cases, it does not allow us to cleanly assess the impact of the program as a whole. We hence consider time series evidence of whether the implementation of the modern system in 1998 reduced appeals to CAVC or remands by CAVC. Combined with institutional knowledge, the evidence does not suggest that the implementation of the program alone improved performance in any substantial way. Yet regardless of the behavioral effect the creation of the QR program, our paper addresses why QR of cases fails to prevent remands and reversals relative to the control group.

Our paper also informs several other strands of scholarly literature. First, our findings illustrate the difficulty of performance measurement in the public sector when a principal's objective may not be contractible and when there are heterogeneous objectives (Holmstrom and Milgrom 1991; Baker 1992; Barnow 2000; Dixit 2002; Bevan and Hood 2006; Duflo et al. 2013). Our findings underscore the difficulty of monitoring bureaucratic and judicial quality, which is central to questions of presidential and congressional oversight of agencies (McCubbins and Schwartz 1984; Cuéllar 2006; Boyd and Driscoll 2013), and can be conceived of as an example of supervisor–agent collusion (gaming performance targets) in the agency framework of Tirole (1986). Second, scholars have long debated whether an appeals process can serve as a form of "error correction" (Shavell 1995), with administrative law scholars expressing more skepticism in the mass adjudicatory context, particularly given nonrandom selection of appeals (Mashaw 1980; Simon 2015; Hausman 2016). Our paper shows limitations to an agency's ability to reduce reversal rates even *with* random selection of judicial decisions. We also find that non-appealed cases continue to have high rates of errors flagged even under the lenient standard deployed by the QR team, suggesting imperfect selection of errors for appeal. Third, our study also provides evidence of the causal effect of standards of review, a core topic of administrative law (see, e.g., Breyer et al. 2011). Our setting enables us to examine how the same set of cases fared under two divergent standards of review. This helps overcome conventional selection challenges in observational studies of the impact of standards of review (e.g., Schuck and Elliott 1990; Miles and

2. Notable exceptions are Ridgway and Ames (2018) and Ridgway et al. (2016).

Sunstein 2006). Last, these findings address the question of whether institutions can be reformed from within (Banerjee et al. 2012), particularly in the development of an "internal administrative law," a topic of increasing scholarly focus (Metzger and Stack 2016; Parrillo 2017; Sabel and Simon 2017).[3] Our evidence is consistent with Blanes i Vidal and Leaver (2015), who find that favoritism bias leads judges to reverse peers less frequently when reviewing the quality of judicial decisions. Such potential for favoritism and conflicts may be a substantial challenge to developing QR—and administrative law—from within an agency.

Our paper proceeds as follows. Section 2 provides institutional background to veterans adjudication and the QR process. Section 3 describes our unique BVA dataset and demonstrates that we are able to replicate the randomized case selection process for QR. Section 4 presents results, including analyses of the overall (intention-to-treat) effect and the (complier) effect on the subgroup of cases for which QR triggered a memorandum to the VLJ. Section 5 considers the possibility that the implementation of the modern QR system had a program effect that is not manifested in case-specific review. Section 6 discusses other limitations and Section 7 concludes with implications.

2. Institutional Background

Each year, the Department of Veterans Affairs (VA) administers benefits amounting to roughly $90 billion per year, covering over 6.5 million veterans and dependents (Figure 1). The majority of Board cases involve disability benefits. A disability benefits claimant first files an application online, in writing, or in person at one of the Veterans Benefits Administration's (VBA) regional offices.[4] If the claimant is dissatisfied with the initial decision, she can file a notice of disagreement, which occurs for roughly 11–12% of initial decisions. The regional office then reexamines the application. Upon its decision, the claimant can then appeal to the BVA. Roughly one-third of initial notices of disagreement reach the Board.[5]

In 2015, the BVA's annual budget was around $94 million (Board of Veterans Appeals 2015: BVA-1), with most of it allocated for the personnel of 63 VLJs and 450 staff attorneys (Board of Veterans' Appeals 2016).[6] VLJs are appointed by the President and removable only for

3. In a companion paper, we expand on the implications of theories of internal administrative law (Ames et al. 2020).

4. In fiscal year 2015, 98% of BVA cases originated from the VBA, and less than 2% originated from the Veterans Health Administration or the National Cemetery Administration.

5. For more extensive background on the administrative process at the VA, see Asimow (2018).

6. The VLJ figure includes only frontline VLJs. Including Chief VLJs and Deputy Vice Chairmen, the number would be roughly 90. Since 2017, BVA has undergone significant expansion to increase case output. As of May 2017, the Board had 630 staff attorneys and 83 VLJs, and was planning to hire an additional 100 staff attorneys and 8 additional VLJs by the end of 2018 fiscal year (Department of Veterans Affairs 2018: 23).

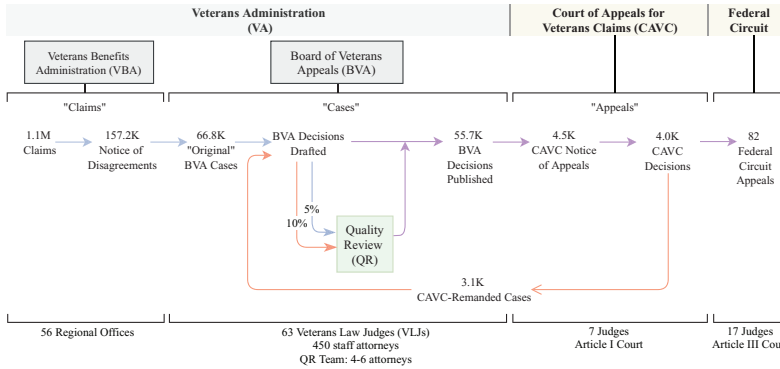

Figure 1. Diagram of Appeals Flow for Veterans Disability Claims.
Notes: Case counts are provided for illustrative purposes for the fiscal year 2015, and may differ slightly from appeals rates presented in the text, as those are averaged across the observation period. "Original appeals" are estimated by subtracting CAVC-remanded cases and also include appeals from other benefits program. Between 94% and 98% of appeals are typically for disability compensation. Blue arrows indicate "original appeals" and red arrows indicate "CAVC-remanded" cases.

cause (31 U.S.C. §7101A). While appointments are made by the President, the process is traditionally pro forma. VLJs are career civil servants and in many respects comparable to ALJs.[7] VLJs do not have fixed terms, and turnover has historically been low, with most VLJ departures stemming from retirements. For example, among VLJs promoted in 2001, all remained in the role past 2005, two-thirds remained past 2010, and one-third remained past 2015. The majority of VLJs started their careers as BVA staff attorneys. Staff attorneys are hired and subject to the federal civil service (General Schedule, GS) system. Of staff attorneys hired in 2001, 67% remained past 2005, 38% remained past 2010, and 29% remained past 2015. VLJs hear appeals, hold hearings, and issue opinions under a *de novo* review standard, applied to the full claims file. The volume of benefits determinations at the VA and caseloads at the BVA are high. In January 2018, over 157,000 cases were pending with the Board, and the Board resolves over 50,000 cases annually.[8] Veterans who received a resolution by the Board in 2017 waited an average of 7 years from filing the Notice of Disagreement (Department of Veterans Affairs 2018).

7. For instance, 31 U.S.C. §7101A(b) provides that VLJs are paid according to the same salary scale as ALJs and 31 U.S.C. §7101A(e) provides that VLJs may be removed under the same provisions that govern removal of ALJs, although VLJs are subject to performance reviews.

8. These output figures do not reflect the most recent push for greater production. The Board projects output of over 80,000 cases per year in fiscal years 2018–19, but also projects case receipts to exceed 90,000 in fiscal year 2018 and 115,000 in fiscal year 2019 (Department of Veterans Affairs 2018).

Until 1988, BVA decisions were final. With the Veterans' Judicial Review Act of 1988, Congress added another layer of appeal, creating the US CAVC. CAVC is statutorily comprised of three to seven judges, with a current temporary expansion to nine judges, who are appointed by the President, subject to advice and consent of the Senate. Unlike Article III judges, CAVC judges are not lifetime appointees and serve fixed 15-year terms (38 U.S.C. §7253). CAVC judges have the authority to decide cases in three-judge panels or by a single member of the court. The majority are veterans and one criticism has been that a minority have substantial prior experience with veterans' law issues. For instance, no CAVC judge has ever worked as an adjudicator at BVA, nor been a member of the private bar practicing before the CAVC (Hennings et al. 2016). Because historically BVA was the final appeal, the terminology can be confusing. For clarity, we will refer to matters decided at BVA as "decisions" and matters decided at CAVC as "appeals."

Roughly 6% of all BVA decisions are appealed to CAVC, which reviews findings of fact under a "clearly erroneous" standard and findings of non-factual issues under an "arbitrary and capricious" standard (38 U.S.C. §7261). It is worth noting that VA does not have the statutory authority to appeal grants of benefits, so the 6% overall appeals rate translates into a 14% appeals rate of cases with at least one issue denied, which is comparable to other administrative contexts.[9] CAVC remands BVA decisions frequently (Ridgway 2009). Roughly 76% of all cases appealed to the CAVC result in a remand (on at least one issue) to the BVA. In 14% of CAVC-remanded cases, the resulting BVA decision is again appealed to CAVC. CAVC decisions may also be appealed to the US Court of Appeals for the Federal Circuit and then the US Supreme Court, but such appeals are exceedingly rare.

The creation of CAVC and its high remand rate led BVA to develop a more systematic QR program. Describing the initiative in 1998, the BVA Chairman wrote:

> Quality in appellate decision-making is one of several ways to measure how well the Board is fulfilling its statutory mission ... It is also the Board's *single most important goal* in fulfilling that mission because timely delivery of appellate decisions is meaningless if the underlying adjudication is fundamentally flawed.[10]

9. For instance, the appeals rate from the Social Security Appeals Council to district courts, which is institutionally most comparable to an appeal from BVA to CAVC as that requires securing an attorney to an adversarial court, is roughly 13–15% (Social Security Administration 2018).

10. Richard B. Standefer, Acting Chairman, Memorandum No. 01-98-15 (May 14, 1998) (emphasis added).

The Office of Quality Review, comprised of four to seven attorneys, reviewed draft opinions for specific errors. During the time of this study, QR attorneys were competitively selected to serve 2-year terms. QR attorneys were drawn exclusively from BVA staff attorneys (at the GS-13 and GS-14 grades), who were drafting decisions for VLJs. Following completion of a QR detail, most QR attorneys returned to drafting decisions for VLJs. These details were generally coveted positions, and the majority of QR attorneys have eventually been promoted to GS-15 Senior Counsel or VLJ positions. For most of its existence, the QR Office has been led by a permanent, competitively selected GS-15 Senior Counsel, internally known as the "Chief." All QR Chiefs previously served as QR attorneys. In 2016, salaries (excluding benefits) for the seven-member QR team amounted to roughly $780k.

While the QR program was revised in the early years, in part due to criticism by the GAO (US General Accounting Office 2002), it remained formally unchanged from November 1, 2002 to November 15, 2016. The program randomly selected 5% of "original" appeals (i.e., those not on remand from the CAVC) and 10% of appeals on remand from the CAVC. Random selection was made by computer after an opinion was drafted by a VLJ, but before the opinion was issued, so as to enable VLJs to make corrections. The QR team determined whether the opinion (a) addressed all relevant issues, (b) accounted for all evidence, (c) addressed relevant laws and regulations, (d) provided a clear explanation of the "reasons and bases" for the decision, (e) addressed due process, and (f) was properly formatted (e.g., spelling, grammar, and structure). Each QR team member coded these categories along with a more exhaustive subcategory coding.[11] Formally, the QR team's standard of review was equated with CAVC's: the QR team should "call" a substantive error (i.e., errors excluding formatting errors) when the opinion exhibited "a deficiency that would be outcome determinative, that is, result in the reversal or remand of a Board decision by [CAVC]" (Board of Veterans' Appeals 2002: 7). In instances of legal error, the QR team would draft a memorandum to be circulated to the VLJ. VLJs were then given the chance to revise the opinion before it was issued. When a VLJ disagreed with the memorandum, the VLJ was permitted to make an informal challenge to the BVA's Chief Counsel for Policy and Procedure. In practice, VLJs typically revised opinions and made very few challenges to QR memoranda. In addition to these memoranda, the QR team conducted training to address common errors and circulated monthly reports on changes in the law, quality concerns, and errors identified.

One of BVA's strategic performance goals was "to make deficiency-free decisions 95 percent of the time" (US General Accounting Office 2002).[12] VLJs are subject to regular performance reviews and re-certification (38

11. As we detail in Appendix C, the subcategories were refined over time.

12. This performance goal itself changed over time. In its budget request of the 2008 fiscal year, for instance, BVA published a target of 92%.

U.S.C. §7101A). While the Board described the goal of the QR data "to measure performance in the area of quality for the Board as a whole," it also formally permitted its use in performance reviews.[13] Yet the Board also noted that "data obtained as a result of this process is statistically significant at the Board level, rather than at the individual Board member level." In practice, QR information was hence used on occasion to compare performance of "decision teams" of dozens of VLJs, but it was never used in performance reviews of individual staff attorneys or VLJs.

3. Descriptive Statistics and Balance

3.1 Data

We secure data, never before analyzed by outside researchers, on all BVA decisions from October 1, 1999 to January 26, 2018. Originally designed to physically locate files, the scope of the "Veterans Appeals Control and Locator System" (VACOLS) was expanded over time to manage, track, and measure all relevant dimensions of BVA appeals. For each case, we obtain a rich set of variables, including the BVA disposition (e.g., whether relief was granted) on each issue, prior procedural history (e.g., hearing information), appellant information (e.g., age, gender, service period), issues disputed (e.g., whether the disability had a service connection), diagnostic categories for each issue (e.g., musculoskeletal disease), whether the case was selected for QR, and all error codes the QR team identified for that case, whether the case was appealed to CAVC, CAVC's disposition on each issue (e.g., affirmed, remanded), and BVA's coding of the reason for a CAVC remand. We clean and restructure the database, resulting in 2,727,418 appeals, 6,157,531 unique issues, 459,628 hearings, and 39,528 appeals selected for QR.

3.2 QR-Eligible Cases

In order to identify the causal effect of QR on a case, we need to be able to replicate the 5% selection rule for original decisions and the 10% selection rule for CAVC-remanded opinions. We use public reports, internal records secured through FOIA, and institutional knowledge to exclude cases ineligible for QR. First, we exclude decisions that were "supplemental actions" (actions taken after the Board entered a decision), reconsiderations of final decisions, and procedural actions (e.g., designations of records for appeal to CAVC). Second, we exclude any decision subsequently made by the VBA upon remand by the Board. Third, we exclude dismissals due to the death of the appellant. After these exclusions, we are

13. See Richard B. Standefer, Acting Chairman, Memorandum No. 01-98-15 (May 14, 1998) ("Each [Deputy Vice Chairman (DVC), the head of a Decision Team] is responsible for maintaining high quality ... in the performance of individual staff counsel and Board members ... The DVC shall use QR data available from within the team, from VACOLS, and from opinions of the Court as management tools to assist in the identification of areas needing improvement and the implementation of corrective action.").

left with 785,812 QR-eligible decisions, comprising about 29% of all VACOLS appeals.

The left panel of Figure 2 plots the volume of original cases and CAVC-remanded cases over time. As caseload has been increasing over time, the number of original cases has been rising over time, with an average of roughly 11,500 per quarter in 2016. CAVC remands have similarly been rising over time, with an average of about 1500 remands per quarter in 2016.

3.3 Observation Window

While the modern QR program was created in 1998, it was subject to revision and critiqued in an influential GAO Report in 2002 (US General Accounting Office 2002). Specifically, GAO pointed out that the early implementation of the program was beset by sampling irregularities. We continue to observe evidence of such irregularities until August 2003, so we limit our observation window to cases eligible for QR from August 1, 2003 to November 9, 2016, the last date appeals were selected for the same QR program. On November 15, 2016, the program was substantially revised to terminate random sampling of cases for QR.

To check our replication of the QR selection process, we calculate selection rates for original and CAVC-remanded cases, which should be around 5% and 10%, respectively. This calculation requires recreating the precise timing for QR selection, as a case was QR-eligible after the decision was signed but before the decision was dispatched to the appellant. Although we observe the dispatch date, we do not observe the signature date. As a proxy for this date, we used the date that the decision attachment was uploaded to the system, which was conducted in the majority of cases by an administrator after signature but before the dispatch of the decision.[14]

The right panel of Figure 2 displays the selection rate over time, with the lower line plotting the time series for original decisions and the upper line plotting the time series for CAVC-remanded decisions. Gray horizontal lines indicate the expected selection rates of 5% and 10%. The pre-2003 time series confirms sampling irregularities documented by GAO, as well as changes in the QR program from 1999 to 2003. (Conducting this check at the VLJ-level also led us to uncover that cases written by senior management were exempt from QR, a fact confirmed by staff, leading us to exclude these decisions from our analysis.) From 2003 to 2016, we are able to cleanly replicate the Board's publicly stated sampling rate. The

14. To ensure that the attachment date was pretreatment, we excluded cases in two scenarios where the attachment date was modified after QR selection: (1) we exclude appeals that had a decision attachment date greater than the decision dispatch date (0.16% of all cases); (2) we exclude quality reviewed cases where the user name of the reviewer matched the user name of the attachment uploader, as this indicates that the QR corrected missing documents (0.25% of all cases).

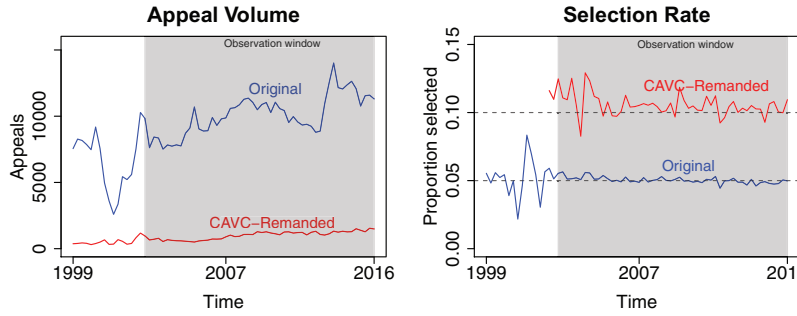

Figure 2. Quarterly Eligible Case Volumes (left) QR Selection Rates (right) Over All Available Time Periods, with the Observation Window for the Study Shaded in Gray (August 1, 2003–November 9, 2016).

Notes: Cases are split by whether they had been remanded by CAVC, resulting in the decision of interest. Within the observation window, we are able to replicate the 5% and 10% selection rates for original and CAVC-remanded cases, respectively.

sampling variability for CAVC-remanded appeals is higher, as these constitute less than one-fifth of QR-eligible decisions. For 508,801 original cases, we calculate a 5.01% selection rate, and for 47,981 CAVC-remanded cases, we calculate a 10.49% selection rate. Although the right panel shows that there is some variability in the quarter-by-quarter selection rate—driven by idiosyncratic factors such as the fiscal year, staffing, and turnover—the rates suggest we have replicated BVA's selection scheme.

3.4 Balance

Random selection should ensure that QR cases are comparable on all observable dimensions to non-QR (or control) cases. We verify this by checking balance on a wide range of preselection (or pretreatment) covariates. Table 1 displays the difference between QR and control cases for selected covariates, along with t-tests for statistical significance. All differences are small in absolute magnitude and not statistically significant. For instance, 49% of QR cases involve Vietnam War veterans, compared with 49% of control cases. The average age of the veteran is 55.6 for QR cases and 55.7 for control cases. Because the covariate set is so rich, Figure 3 summarizes balance with a quantile–quantile plot of t-statistics of all 80 covariates against a reference t-distribution. As expected, these test statistics line up on the 45° line. We test for distributional equivalence between the observed and reference distributions (using a Kolmogorov–Smirnov test), yielding p-values of 1 and 0.3 for original and CAVC-remanded cases. Across all salient dimensions—BVA appeals history, prior hearings, issue types, and a rich set of diagnostic codes—there are no substantively or statistically significant differences between QR and control cases.

Table 1. Balance on Selected Covariates between Cases Not Selected for QR (Ctrl) and Cases Randomly Selected for QR between August 1, 2003 and November 9, 2016

	Original cases				CAVC-remanded cases			
	Ctrl.	QR	Diff.	*p*-value	Ctrl.	QR	Diff.	*p*-value
Appellant age (years, at notice of disagreement)	55.62	55.70	0.08	0.91	54.59	54.43	−0.15	0.74
Appellant is male	0.94	0.94	−0.00	0.91	0.95	0.94	−0.00	0.84
Appellant service period (prop.)								
World War II (9/16/40–7/25/47)	0.08	0.08	0.00	0.91	0.08	0.08	0.00	0.88
Peacetime (7/26/47–6/26/50)	0.04	0.04	0.00	0.99	0.04	0.03	−0.00	0.59
Korean conflict (6/27/50–1/31/55)	0.09	0.09	0.00	0.91	0.11	0.10	−0.00	0.74
Post-Korea (2/1/55–8/4/64)	0.14	0.14	0.00	0.99	0.17	0.17	0.00	0.84
Vietnam Era (8/5/64–5/7/75)	0.49	0.49	−0.00	0.91	0.53	0.52	−0.01	0.74
Post-Vietnam (5/8/75–8/1/90)	0.35	0.35	0.00	0.99	0.33	0.34	0.01	0.60
Persian Gulf (8/2/90–Present)	0.25	0.25	−0.00	0.91	0.17	0.16	−0.01	0.74
Issues per appeal	2.62	2.61	−0.00	0.99	2.12	2.17	0.04	0.60
Compensation issue types (number of issues per appeal)								
Service connection								
New and material evidence	0.07	0.08	0.00	0.80	0.04	0.04	0.00	0.74
Accrued benefit	0.02	0.02	−0.00	0.91	0.02	0.03	0.01	0.40
All others	1.45	1.45	−0.00	0.99	0.99	1.02	0.03	0.60
Increased disability rating								
Schedular	0.65	0.64	−0.01	0.91	0.54	0.52	−0.01	0.74
Schedular and extraschedular	0.04	0.04	0.00	0.87	0.08	0.09	0.01	0.74
Extraschedular	0.02	0.02	−0.00	0.91	0.03	0.03	−0.00	0.57
Issue diagnosis categories (number of issues per appeal)								
Skeletal injury or motion loss	0.64	0.65	0.00	0.91	0.50	0.51	0.01	0.84
Nonpsychotic emotional illness	0.27	0.27	0.01	0.22	0.27	0.29	0.02	0.08
Hearing loss	0.16	0.16	0.00	0.99	0.07	0.07	0.00	0.84
Musculoskeletal disease	0.14	0.14	−0.01	0.74	0.11	0.12	0.00	0.78
Skin disability	0.12	0.12	−0.00	0.99	0.08	0.07	−0.01	0.40
Sense organ disability	0.10	0.10	−0.00	0.99	0.05	0.05	0.00	0.88
Peripheral nerve paralysis	0.09	0.09	−0.00	0.99	0.07	0.07	−0.00	0.88
Digestive system disease	0.10	0.11	0.00	0.95	0.08	0.08	0.00	0.88
Disease of arteries and/or veins	0.09	0.09	−0.00	0.74	0.06	0.06	0.00	0.84
Genitourinary disability	0.06	0.06	−0.00	0.74	0.04	0.05	0.00	0.74
Representation at BVA								
Disabled American Veterans	0.31	0.31	−0.00	0.99	0.22	0.22	−0.00	0.84
A State Service Organization	0.17	0.17	−0.00	0.91	0.05	0.05	0.00	0.74
American Legion	0.18	0.19	0.00	0.80	0.14	0.13	−0.00	0.74
Veterans of Foreign Wars	0.10	0.10	−0.00	0.88	0.03	0.03	−0.00	1.00
Unrepresented	0.10	0.10	0.00	0.99	0.07	0.08	0.01	0.08
Attorney	0.06	0.06	−0.00	0.91	0.44	0.43	−0.01	0.74
BVA procedural posture								
Duration (years)	4.38	4.40	0.02	0.80	7.94	7.99	0.05	0.74
Prior BVA decision (prop.)	0.37	0.37	0.00	0.99	1.00	1.00	−0.00	0.85
Number of BVA appeals (sample size)	508,801	26,821			47,981	5622		

Notes: Cases are split by whether they had been remanded by CAVC leading to the decision at issue. Tests for all issue categories (e.g., medical diagnosis) were statistically insignificant but are omitted for readability. *p*-values are adjusted for multiple testing using Benjamini and Hochberg (1995). State service organization category excludes state service organizations in Maryland and Virginia, as these have separate representative codes. Duration is measured between notice of disagreement and BVA decision.

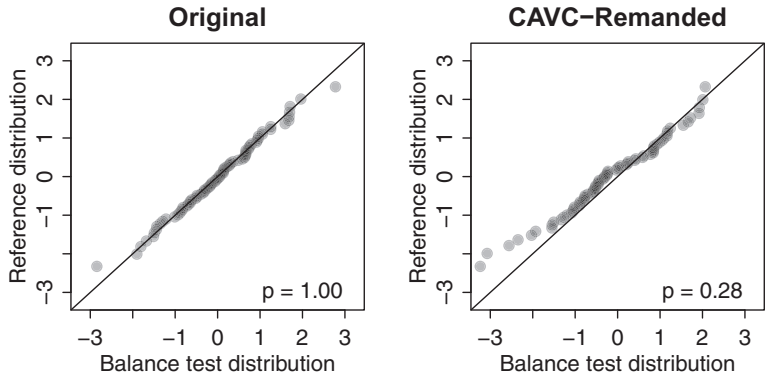

Figure 3. Quantile–Quantile Plot of *t*-Statistics from 80 Balance Tests between Cases Selected and Not Selected for QR Against Reference $t_{n_1+n_2-2}$-Distribution.
Notes: Cases are split by original and CAVC-remanded cases. *p*-values are from a Kolmogorov–Smirnov test of distribution equality between the observed *t*-statistics and reference distribution.

Our understanding of the QR process, the ability to replicate the selection rates, and the wide range of balance checks on the same internal data used to administer the system give us confidence that we have replicated the QR selection process. Random selection with such a large sample ensures balance across QR and control cases, enabling us to assess the impact of QR on case outcomes.

4. Results

4.1 Causal Effect of QR

We now test whether the QR performance program met its stated goal of reducing the number of remands/reversals at CAVC. Because of the divergent QR selection schemes for original and CAVC-remanded cases, we conduct separate analyses for each case type. First, we examine the effect on whether the BVA decision was subsequently appealed to CAVC. If the QR process reduced the number of legal errors by correcting draft decisions, we should expect claimants to be less likely to appeal the decision. The top row in the first panel of Table 2 shows that there is no appreciable reduction in the appeals rate: the rate remains at 6%, regardless of whether the case was subject to the QR process.[15] The relatively large sample sizes allow us to rule out effect sizes of appreciable magnitude: the 95% confidence interval (CI) is [−0.62%, 0.16%]. This first finding

15. We do not focus on the rate of appeals conditional on a denial of at least one issue, because random selection for QR review occurred regardless of disposition. For comprehensiveness, Appendix E examines the effect on the subset of cases with at least one issue denied.

Table 2. Means and Differences-in-Means (Diff.) for Outcomes, Comparing Control (Crtl.) Cases Not Selected for QR and Treatment Cases Randomly Selected for QR between August 1, 2003 and November 9, 2016

	Original cases				CAVC-remanded cases			
	Ctrl.	QR	Diff.	*p*-value	Ctrl.	QR	Diff.	*p*-value
Of all BVA cases...								
Prop. appealed to CAVC	0.06	0.06	−0.00	0.70	0.14	0.13	−0.01	0.36
Sample sizes (cases)	508,801	26,821			47,981	5622		
Conditional on CAVC appeal...								
Case outcome by CAVC								
Vacated and remanded	0.76	0.75	−0.01	0.89	0.72	0.70	−0.02	0.88
Affirmed	0.22	0.24	0.02	0.70	0.25	0.27	0.01	0.90
Abandoned	0.17	0.17	0.01	0.89	0.11	0.12	0.01	0.90
Dismissed	0.08	0.07	−0.01	0.70	0.05	0.06	0.01	0.90
Reversed	0.01	0.01	−0.00	1.00	0.01	0.02	0.00	0.90
Sample sizes (CAVC appeals)	31,590	1604			6782	728		

Notes: The left panel presents data for original cases and the right panel presents cases for CAVC-remanded cases. Outcomes are all actions taken after dispatch of the BVA decision. "Appealed to CAVC" represents the proportion of cases appealed to CAVC. Because each appeal can involve multiple issues, "case outcome" presents the average number of cases with at least one issue subject to each disposition. *p*-values are adjusted for multiple tests using Benjamini and Hochberg (1995). For readability, we exclude disposition codes with very low case counts (i.e., vacated and dismissed, settled, and dismissed due to death), which also have no statistically significant differences between QR and control cases.

suggests that corrections are not substantial enough to change the impression by a claimant or attorney of whether a case should be appealed.

Second, we test for whether the QR program had effects on CAVC resolution, conditional on a CAVC appeal. Because CAVC dispositions occur at the issue level, but BVA recorded QR results at the case level,[16] we summarize CAVC dispositions by calculating the proportion of appeals with at least one issue in each disposition type (e.g., affirm, remand). We find no statistically or substantively significant differences between QR and control cases on CAVC dispositions. For instance, roughly 76% of non-QR appeals had at least one issue vacated and remanded, compared with 75% of reviewed appeals (95% CI for difference: [−2.97%, 1.82%]).

We conduct the same analyses for CAVC-remanded cases in the right panel of Table 2. As it may be procedurally confusing, it is worth remembering the temporal sequence of cases in the right panel of Table 2: CAVC earlier issued a remand and the cases comprise BVA decisions responding to that remand. That BVA decision may be selected for QR (at a 10% rate) and then potentially be appealed to CAVC again.

The appeals rate for CAVC-remanded cases was 14% for non-QR decisions and 13% for QR decisions. This 1% difference is not statistically

16. To be clear, BVA retains a separate internal database for QR at the issue level, but these QR data are currently merged only at the case level with the VACOLS data. This database structure was subject to criticism by GAO (US Government Accountability Office 2005).

significant, although the 95% CI of [−2.65%, 0.27%] is wider due to the smaller sample size. Conditional on a CAVC appeal, 72% of non-QR appeals had at least one issue vacated and remanded at the CAVC, compared with 70% of QR appeals (95% CI [−6.90%, 2.27%]). These findings suggest that, for both original and CAVC-remanded appeals, the QR program did little to stem the backlog of appeals sent back to the BVA for multiple rounds of decisions.

Table 3 presents logistic (fixed effects) regression results to adjust for differences over time and by VLJ hearing the case.[17] For comparability, the first model (top left) provides unadjusted regression results that are analogous to the simple difference-in-means in the top left cell of Table 2. We then add fixed effects for each unique year-quarter (second column) and for each VLJ. Effects for original cases remain statistically insignificant for both the appeal rate (top left) and the remand rate (bottom left). The right columns present comparable fixed effects models for CAVC-remanded cases. While the QR effect on CAVC disposition (conditional on appeal) is again statistically insignificant, we observe statistically significant estimates of the QR program on the probability of appeals for CAVC-remanded cases, corresponding to a 1% reduction in the appeals rate. The magnitude, however, remains small. With 5622 CAVC-remanded cases undergoing QR, the best estimate is that the QR process, staffed by four to six full-time attorneys, avoids roughly 60 appeals over a 15-year period. To put that in context, the Board received over 90,000 cases in 2017 alone, and a single VLJ has 1000 cases docketed annually. A staff attorney is expected to prepare 3.25 cases for full decision each week. These figures illustrate that the effect of QR is small relative to attorney resources committed to it.[18]

In order to isolate the effect of the memoranda drafted by the QR team, we also estimate a series of instrumental variables models. The QR effect models above can be conceived of as recovering "intention to treat" effects, when the treatment of interest may be the memorandum written by the QR team (Angrist et al. 1996). Randomized QR selection can then be used as the instrument for whether a memorandum was written to the VLJ, which occurred for all substantive errors. Because the memoranda formed the principal mode of communication between the QR team and VLJs, and because no communication occurred when no errors were called, the exclusion restriction—that QR selection affected outcomes exclusively through memoranda—is plausible. Table 4 presents results. Again, the results other than the memorandum effect for CAVC-remanded cases on the probability of an appeal are statistically insignificant.

17. We note that such adjustment is not uncontested (see, e.g., Freedman 2008).

18. Appendix E also shows that the effect on appeals for CAVC-remanded cases vanishes when focusing on denials, which are the large majority of cases appealed to CAVC.

Table 3. Logistic Regression Results of the Probability of an Appeal to CAVC (Top Panel) and the Probability of a Reversal or Remand by CAVC, Conditional on an Appeal (bottom panel) for Original Cases (Left Columns) and CAVC-Remanded Cases (Right Columns)

Outcome		Original cases			CAVC-remanded cases		
Appealed to CAVC	QR effect	−0.040	−0.040	−0.040	−0.101*	−0.100*	−0.098*
		(0.026)	(0.026)	(0.026)	(0.042)	(0.042)	(0.042)
	VLJ FEs	N	N	Y	N	N	Y
	Year-quarter FEs	N	Y	Y	N	Y	Y
	N	535,622	535,622	535,622	53,603	53,603	53,603
Reversed/ remandedby CAVC	QR effect	−0.033	−0.045	−0.039	−0.113	−0.098	−0.117
		(0.060)	(0.060)	(0.061)	(0.086)	(0.087)	(0.090)
	VLJ FEs	N	N	Y	N	N	Y
	Year-quarter FEs	N	Y	Y	N	Y	Y
	N	33,194	33,194	33,194	7510	7510	7510

Notes: The QR effect row presents the coefficient on the treatment indicator, with SEs in parentheses. FEs indicate fixed effects estimated using the pseudo-demeaning algorithm described in Stammann et al. (2016), which are not displayed for readability; *N* indicates sample size. $p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

There is hence little evidence of a stronger effect in the subset of cases receiving memoranda.[19]

We now examine whether the effects of QR are heterogeneous across VLJs. One of the recurrent criticisms of BVA adjudication is the lack of "consistency" across judges (US General Accounting Office 2002; US Government Accountability Office 2005). As in many other administrative systems, some judges are perceived as tough and others as lenient. Even if the QR program had no mean effects, it is possible that the feedback would help to reduce inter-VLJ disparities (Ho 2017). If a "tough" judge interprets the "duty to assist" claimants too narrowly, for instance, feedback from QR might increase claimant-favorable dispositions by that judge. If a judge with high allowance rates misinterprets a CAVC precedent about a certain disease category, the QR program might reduce that judge's allowance rates.

To test for such heterogeneous treatment effects, we first conduct VLJ-specific balance checks. In Appendix B, we report rejection rates of covariate balance tests for each VLJ. As expected, we find that chance imbalance is much higher with VLJs who have decided fewer opinions. Similarly, we find that the QR selection rate stabilizes only for VLJs with higher numbers of opinions. We hence focus our inquiry on 57 VLJs with at least 4000 control opinions available. While this might seem like a high threshold, the expected number of opinions that would both go through

19. Sometimes this "complier average causal effect" is also referred to as the "intention to treat" effect for the subgroup of compliers (Hirano et al. 2000).

Table 4. Instrumental Variable Linear Models for the Effect of an "Exception Memorandum" on Outcomes, Using Random Selection for QR as an Instrument

Outcome		Original cases	CAVC-remanded cases
Appealed to CAVC	Memorandum effect	−0.032	−0.210*
		(0.021)	(0.087)
	First stage R^2	0.07	0.05
	Second stage R^2	0.00	0.00
	N	535,622	53,603
Reversed/remanded by CAVC	Memorandum effect	−0.045	−0.257
		(0.082)	(0.197)
	First stage R^2	0.13	0.08
	Second stage R^2	0.00	0.00
	N	33,194	7510

***Notes*: The "memorandum effect" is the causal effect of an "exception memorandum" on the subset of cases that received such a memorandum because of the QR process, and the row presents coefficients with SEs in parentheses. $p < 0.001$, $^{**}p < 0.0$, $^{*}p < 0.05$.

QR and be appealed to CAVC under the null would be less than 12 (=4000 cases ×5% QR selection rate×6% appeal rate). Given small cell counts, we hence use Fisher's exact test for whether the odds ratio of an appeal (or CAVC remand) is higher for QR cases specific to each judge. As a measure of stringency, we calculate the baseline relief rate for each VLJ.[20] We omit results on CAVC-remanded decisions, as VLJ-specific effects are too imprecise.

Figure 4 plots VLJ stringency against the treatment effect. Each dot represents odds ratio of an outcome for each VLJ, weighted by QR sample size, with 95% CIs. If the QR program affected VLJs at the extremes of the allowance rate range, we should observe statistically significant treatment effects at the low and high end of allowance rates. Most effects, however, are centered around the origin, and there is no detectable correlation of effects with the allowance rate. In fact, we reject the null hypothesis for 8% of VLJs, which is close to expected under the null at $\alpha = 0.05$. (A correction for multiple testing using Benjamini and Hochberg (1995) yields no statistically significant VLJ effects at $\alpha = 0.05$.) In short, there is little evidence of heterogeneous treatment effects that would reduce the inter-VLJ disparities.

4.2 Mechanism

Why is the program so ineffective? One potential explanation is that CAVC outcomes are unpredictable. In 2010, the BVA Vice Chairman wrote in an internal memorandum that the "chance of prevailing before [CAVC]" was "difficult to predict." Ridgway et al. (2016) documents substantial disparities across CAVC judges. And if CAVC decisions are

20. We calculate the average allowance rate across all issues for each VLJ.

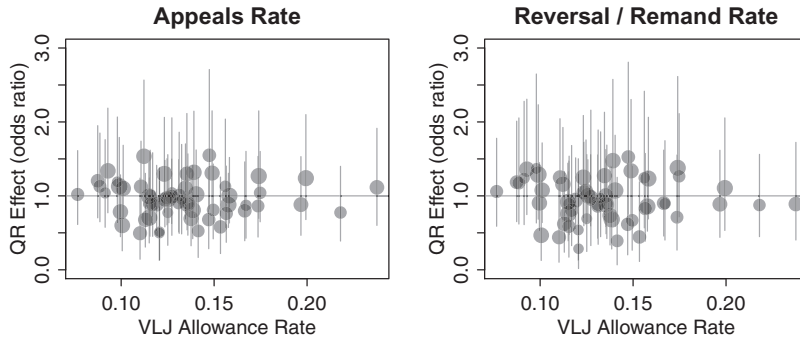

Figure 4. Treatment Effects by VLJ with 95% CIs for the Odds Ratio of an Appeal to the CAVC (left) and a Remand/Reversal by the CAVC (right) for Original Cases and VLJ Allowance Rate on the *x*-Axis.
Note: Only VLJs with at least 4000 control cases are shown.

simply arbitrary, then there would be no reason to expect the QR program to affect likelihood of success at CAVC.

We test this hypothesis by focusing on the subset of cases that underwent QR and comparing cases for which a substantive error was flagged with cases for which no error was flagged. If CAVC outcomes are indeed unpredictable, there should be no association between QR-detected errors and CAVC outcomes. Table 5, however, shows that this is not the case. The top row shows that BVA decisions with flagged errors (excluding formatting errors) had higher rates of appeal for both original and CAVC-remanded cases. The QR team, attorneys, and claimants appear capable of distinguishing higher and lower quality decisions. Conditional on an appeal to the CAVC, original appeals with any error were 12% more likely to have at least one issue vacated and remanded ($p < 0.01$). In contrast, there was no statistically significant difference in the remand rates conditional on appeal for CAVC-remanded cases ($p = 0.48$). In short, while CAVC judges may differ in propensities, the lack of effectiveness of the QR process cannot be explained by sheer randomness of appeal outcomes at CAVC.

Another potential explanation is that VLJs might simply defy memoranda written by the QR team, making no revisions to correct legal errors documented in draft decisions. This hypothesis requires one to believe that the principal work output by a full-time team of four to six staff attorneys is being ignored by VLJs. There are several reasons to doubt this. First, memoranda were routed to VLJs through supervisors, providing an incentive for VLJs to respond. Second, one of the common complaints by staff attorneys and VLJs is about the lack of time to conduct extensive legal research on all cases, given the caseload expectations at the BVA. An individualized memorandum offering advice and legal research on how to correct errors would appear to be a welcome method of improving opinions. Third, while we do not have direct evidence of revisions made to

Table 5. Means and Differences-in-Means (Diff.) for the Subset of Cases that Went through QR, Comparing Cases with No Errors Found (No Error) and Cases with Some Error Found (Error) between August 1, 2003 and November 9, 2016

	Original cases				CAVC-remanded cases			
	No error	Error	Diff.	*p*-value	No error	Error	Diff.	*p*-value
Of all BVA QR cases...								
Prop. appealed to CAVC	0.06	0.11	0.05	0.00	0.13	0.20	0.08	0.00
Sample sizes (cases)	24,895	1926			5305	317		
Conditional on CAVC appeal and QR...								
Case outcome by CAVC								
Vacated and remanded	0.74	0.86	0.12	0.00	0.69	0.77	0.07	0.48
Affirmed	0.25	0.17	−0.08	0.03	0.27	0.25	−0.02	1.00
Abandoned	0.16	0.24	0.08	0.03	0.11	0.17	0.06	0.45
Dismissed	0.07	0.06	−0.01	0.77	0.06	0.05	−0.01	1.00
Reversed	0.01	0.00	−0.00	1.00	0.01	0.03	0.02	0.48
Sample sizes (appeals)	1392	212			664	64		

Notes: Formatting errors are excluded. The left panel presents data for original cases and the right panel presents cases for CAVC-remanded cases. Outcomes are all actions taken after dispatch of the BVA decision. "Appealed to CAVC" represents the proportion of cases appealed to CAVC. Because each decision can involve multiple issues, "case outcome" presents the average number of cases with at least one issue subject to each disposition. *p*-values are adjusted for multiple tests using Benjamini and Hochberg (1995). For readability, we exclude disposition codes with very low case counts (i.e., vacated and dismissed, settled, and dismissed due to death), which also have no statistically significant differences between QR and control cases.

draft opinions, as BVA does not retain those records, through interviews with former BVA officials and institutional knowledge gained from one of the coauthors as former-Chief of the BVA's Office of Quality Review, our understanding is that VLJs commonly incorporated decision-specific feedback from the QR process. Last, while some VLJs may have paid less attention to QR memoranda, we do not detect statistically significant effects for nearly all VLJs, suggesting something more general is transpiring.

Based on internal documents describing the QR program over time, the results of Table 5 offer a more compelling explanation: the standard of review. Notwithstanding the fact that the internal standard of review was formally equated with that of CAVC, internal documents secured through FOIA show that the QR process in fact gave significantly more deference to VLJ determinations. In a Chairman's Memorandum, the QR team was instructed to ignore instances with "legitimate differences of opinion." In updating QR instructions in 2017, the Board became even clearer in its QR training manual, stating that an error must be "undebatable" to be flagged. The net effect was that while some errors were corrected through the QR process, the process was nowhere close to the stringency required to withstand scrutiny on appeal. Most stunningly, Table 5 shows that for cases to which the QR team gave a clean bill of health (i.e., with no errors identified), CAVC remanded 74% of the time when appealed.

To corroborate this explanation, the dark red bars of Figure 5 report the proportion of times that specific error categories are called by the QR team. Across the board, these call rates appear low: the QR team called an error for failure to explain the "reason or basis" of an opinion (veteran's law jargon for administrative law's demand for a reasoned explanation) in under 5% of QR cases. The light red bar indicates the error rate increase for the sample of QR cases that were also appealed to CAVC. The combined error citation rates are slightly higher, as would be expected if the QR team is able to identify lower quality cases. To compare this to the CAVC standard of review, we leverage the fact that BVA's own data code whether the reason for a CAVC remand was due process or "reasons or bases."[21] The blue bars report these remand rates for the same set of cases that both went through QR and were appealed. The column shows that CAVC remand rates are substantially higher than BVA's error rates. CAVC remands on due process grounds in 10% of appeals, an issue flagged only 4% of the time by the QR team in the same cases. Most dramatic is that CAVC remands 62% of appeals for inadequate "reasons or bases," but BVA's QR team flags these errors only 10% of the time. These data provide strong evidence that the QR process does not review cases as stringently as CAVC.

We can also probe this explanation by examining variation between the QR team members. In general, QR cases were assigned to each reviewer in the chronological order that they were drawn.[22] This allows us to measure the stringency of each reviewer by calculating the rate at which each reviewer calls errors. We find substantial variability across reviewers, with one reviewer calling errors for 17% of all cases and three reviewers calling errors in under 3% of cases. We hence test whether this internal variation in stringency is associated with agreement with CAVC on the sample of QR cases that were also appealed. We measure agreement by correspondence between (a) whether CAVC reversed or remanded and (b) whether the reviewer called an error. Figure 6 plots reviewer stringency on the x-axis against the agreement rate on the y-axis. Each dot represents 1 of 41 reviewers, weighted by the number of QR cases processed. We indeed observe that more stringent reviewers are more likely to agree with CAVC's disposition. Based on a least squares fit, a 10% increase in the error call rate is associated with a 25% increase, plus or minus 10% at a 95% level, in the CAVC agreement rate. This variation suggests that increasing stringency would align BVA's internal standard of review with that of CAVC.

21. Appendix D provides detail on the coding of remand reasons. While specific codes have changed over time, these can be largely mapped to broader categories of due process and reasons or bases.

22. For extremely complicated cases, the chief of the QR office would ensure balance of workloads across reviewers.

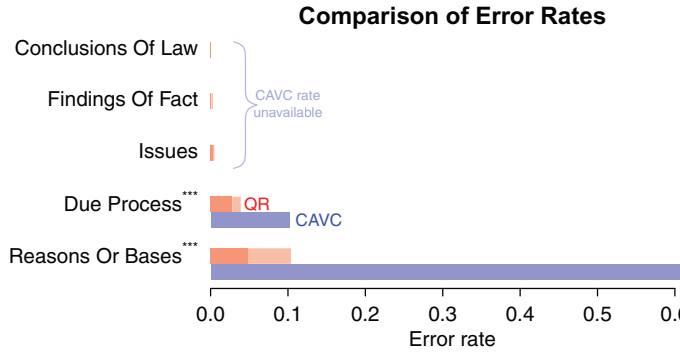

Figure 5. Comparison of Rates at which QR Team and CAVC Identify Errors by Error Type.
Notes: Red bars plot QR team error call rate. Dark red indicates rate for all QR cases and combined red bars indicate rate for QR cases also appealed to the CAVC. Blue bars indicate rates at which QR cases appealed to the CAVC were remanded "reasons or bases" and "due process." CAVC remand reasons were not available for three error rate categories. Stars report statistical significance tests on the difference between the QR call rate and the CAVC remand rate. $p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

What explains BVA's weakening of the internal standard of review? One potential explanation lies in the incentive structure for QR team members. In 2002, GAO had critiqued the initial arrangement, calling for "separation of key duties and the governmental performance audit standard calling for organizational independence for agency employees who review and evaluate program performance" (US General Accounting Office 2002; US Government Accountability Office 2005). BVA's response was to create a distinct QR team of (non-VLJ) staff attorneys to carry out the QR process. But because these QR team members might later return to writing for VLJs or seek elevation to a VLJ position,[23] these staff members may be willing to abide by a lower standard of review.

Another compelling, and not necessarily exclusive, explanation stems from the fact that the QR program had dual purposes, namely (1) to reduce errors and (2) to report a performance measure pursuant to the GPRA. Under the GPRA, BVA published its accuracy rate as the principal performance measure to support its annual budget requests. With performance targets, weakening the standard of review internally may have been the easiest method of generating the *appearance* of effectiveness. The best evidence of this dynamic comes from a memorandum by Vice Chairman Steven Keller in 2010. The VA's own Office of General Counsel (OGC) had sharply questioned the BVA's reported accuracy rate of 94%. OGC noted that in 2009, CAVC alone had reversed or remanded a higher

23. Nearly half of former QR staff members have gone on to serve as VLJs.

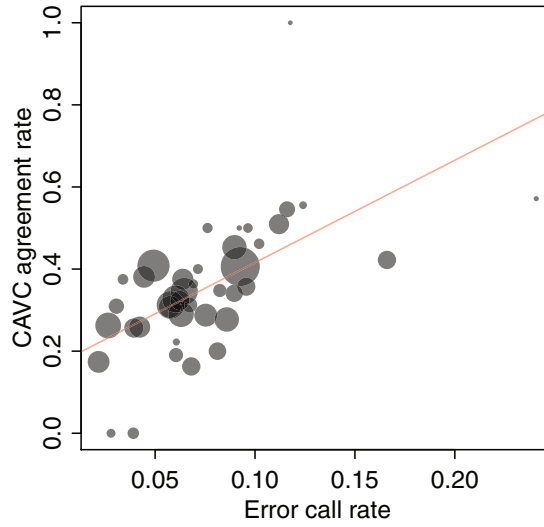

Figure 6. Correlation between Stringency of Each QR Reviewer (i.e., the error call rate) on the *x*-Axis and the Reviewer's Agreement with CAVC on the *y*-Axis.
Note: Reviewers with higher error call rates are more likely to agree with CAVC, conditional on a case being appealed.

absolute number of cases than would be mechanistically possible under a mere 6% "error rate." Vice Chairman Keller responded by stating that CAVC reversal or remand did not necessarily mean that the BVA decisions contained error. Keller argued that a remand for failure to provide an adequate explanation—textbook administrative law—should not be counted as error because the standard is "highly subjective and inconsistently applied."[24] Discounting the one error that is the predominant reason for CAVC remands is effectively an admission of decoupling BVA's standard of review from CAVCs.

Last, it is worth noting that appeals selection does not explain the discrepancy between BVA's claim of a high accuracy rate and CAVC's high remand rate. To the contrary, we have direct evidence that appeals do not perfectly sort erroneous and non-erroneous cases. The QR team calls errors in 6.5% of cases that are not appealed, compared with a baseline of 6.9% across all cases. In other words, even with a lenient standard of review, non-appealed cases appear to have significant errors.

5. Program Effect

While our study has, for the first time, identified the causal effect of QR review on case outcomes, our research design may be limited in a different

24. Steven L. Keller, Vice Chairman, Memorandum on Monthly Performance Review Submission on the Board of Veterans' Appeals' Accuracy Rate (August 3, 2010).

sense. Case-specific randomization does not directly permit evaluation of the causal effect of the QR program as a whole. It is theoretically possible that while QR review had no effect on individual cases, it affected quality overall. Knowing that 5% of (original) cases would be subject to QR might have caused VLJs to improve decision-making when the modern QR system was implemented in 1998. We note at the outset that at least in theory, the creation of an additional mechanism to detect errors may actually create *moral hazard*. For instance, several interview subjects indicated that volume required VLJs to issue decisions that were "good enough," with CAVC playing a potential backstop. As put vividly by one judge:

> [T]he regional office level is like the medic that's out with the squad in combat. They're just doing a quick triage, trying to do a very fast, rough justice. The Board of Veterans Appeals is like the MASH unit. Just given the volume, we're doing the field surgery kind of treatment. That's basically all we can realistically do and then the CAVC is like Walter Reed. They've got more time to really get it right.

The existence of QR could similarly provide a form of insurance.

While it is hard to test these hypotheses as rigorously as we can assess the causal effect of QR review on cases, we consider evidence of program effect here.

5.1 Time Series Evidence

If the implementation of the modern QR system in May 1998 shifted the BVA to a higher quality equilibrium, we should be able to observe an interruption in the appeals rate to CAVC and the CAVC remand rate. Unfortunately, the VACOLS microdata do not span to pre-1998. We hence hand collect information from VA budget requests and annual reports by the VA, the BVA, and CAVC to construct historical time series of the appeals rate to CAVC and CAVC dispositions. Statistics from congressional hearings provide information about fiscal year 1997,[25] and the CAVC annual report includes information on all appeals from 1998 onward. Due to filing delay, the majority of CAVC appeals filed in the fiscal year of 1998 should be from BVA cases decided before the QR program announcement in May 1998.[26] These sources provide us with 2 years of pre-program information. The left panel of Figure 7 shows that the appeals rate did not exhibit any difference before and after the modern QR system was implemented in 1998. Contrary to any notion of

25. See House Committee on Veterans' Affairs (2007b) ("It is clear that the Veterans Court's caseload has increased continually since it opened its doors in 1989. For example, 10 years ago, sir, in 1997, the Court received 2229 cases.").

26. Claimants have up to 120 days to file an appeal and the 1998 fiscal year runs from October 1, 1997 to September 30, 1998.

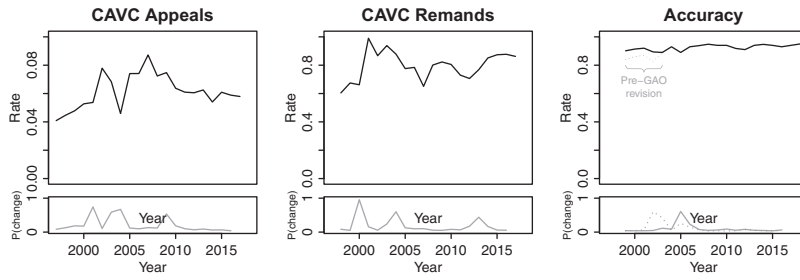

Figure 7. Time Series Evidence about Program Effect.
Notes: The top panels plot the time series of the CAVC appeal, the probability of a CAVC remand conditional on appeal, and the accuracy rate derived from the QR program. The CAVC remand rate is calculated as the proportion of decisions by CAVC on the merits of appeals from the BVA in which at least one issue was reversed/vacated and remanded. CAVC's decisions on the procedural grounds of appeals from the BVA and on petitions for extraordinary relief are excluded from the remand rate calculation. The bottom panels plot the posterior probability of a structural break in the time series using the Bayesian change point model of Barry and Hartigan (1993). Years represent fiscal years, so that the 1998 fiscal year, for instance, represents October 1, 1997 to September 30, 1998. The accuracy before 2003 employed a different sampling scheme and accuracy calculation and was changed due to the report by the US General Accounting Office (2002). The pre-2003 rate is reported unadjusted in gray dashed lines and adjusted based on the proportion of non-substantive errors in id., p. 8 (finding that 5/13 errors were non-substantive). The accuracy rate for 2017 is interpolated due to the abolition of the program in 2017. Change point probabilities for the accuracy time series are presented for the adjusted and unadjusted series in solid and dashed lines, respectively.

improvement in 1998, the lowest appeals rates were for 1997 and 1998. To formalize this, the bottom panel presents posterior probabilities of a structural break, using the Bayesian change point model of Barry and Hartigan (1993).[27] We find no evidence of any structural break.

The CAVC annual reports also provide us with information about the disposition of appeals. While this information is reported only starting in fiscal year 1998, we downloaded docket sheets for all 1998 CAVC decisions to estimate the dates of the underlying BVA decision using the Notice of Appeal filed at CAVC. We find that between 81% and 88% of CAVC decisions involved BVA decisions pre-dating 1998.[28] If the implementation of QR in May 1998 improved quality, we would expect to observe a decrease in the CAVC remand rate relative to this baseline year.

27. We use the implementation by Erdman and Emerson (2007), with default settings of a prior probability of 0.2, 50 burn-in iterations, and 500 Markov Chain Monte Carlo draws.

28. This estimate is for filings between 1 and 3 months after a BVA decision is issued, when there is a 120-day filing deadline. If all appeals were filed at 90 days, roughly 49%, 30%, and 9% were decided by BVA in 1997, in 1996, and by 1995, respectively.

The middle panel of Figure 7 shows that the remand rate is, if anything, the lowest in fiscal year 1998. While there are complications due to statutory changes (e.g., the Veterans Claims Assistance Act of 2000), the time series provides no evidence of reductions in the CAVC remand rate around 1998. The bottom panel confirms that the (posterior) probability of structural breakpoints is negligible.

The above time series evidence is inconsistent with the notion that the 1998 QR implementation had any substantial effects on the quality of decision-making. That said, it is possible that the lack of longer pre-program information does not enable us to detect structural shifts. Yet congressional hearings suggest that the CAVC remand rate has remained "remarkably consistent" from 1995 to 2006, with CAVC remanding in a "whopping 77.7 percent" of appeals (House Committee on Veterans' Affairs 2007a).

The time series evidence also does not suggest that the QR program has gradually improved accuracy over time. The right panel of Figure 7 plots the accuracy rate over time. For nearly the entire observation period, the accuracy rate has hovered above 90%. There is weak evidence that accuracy was lower in the earlier years. But given the evidence above that is most plausibly explained by the changing standard of review. [The lower accuracy rates in 1999–2002, plotted in dashed gray, reflect a change in how errors were calculated, due to US General Accounting Office (2002).]

To be sure, this time series evidence is limited, as other factors (e.g., the changing veterans bar, the evolving role of CAVC) may have changed around 1998. That said, we are not aware of any major shift around 1998 that would mask a real program effect.

5.2 Institutional Reasons

Based on extensive interviews conducted with current or former BVA officials, we also believe there are strong institutional reasons to doubt that the implementation of the QR program in 1998 had any effect. We articulate three principal reasons.

First, our interviews revealed that QR results were not considered in VLJ or staff attorney performance reviews, notwithstanding the formal position articulated in 1998 that such data *could* be used in such reviews. One interview subject stated that it "always bugged me as both a quality review member and as a VLJ that [QR data] didn't seem to factor into anything." Another emphasized that using QR data for performance reviews "would be a very difficult thing to execute. I mean the errors would have to be pretty blatant and, I would say, rise to an undebatable level to be able to actually use that in a judge's performance evaluation." Another emphasized that performance evaluations largely focused on case production: "It's just numbers. It's really just numbers. I've never had anyone come to me . . . or even heard of somebody being talked [to] about quality or outcomes." Formally ratifying that practice, the most recent Case Review Manual instructs the QR team to send the memorandum only

to the VLJ and *not* to the respective Chief VLJ or Deputy Vice Chairman—that is, the individuals bearing significant responsibility for performance monitoring.

Second, VLJs were only informed of QR selection when a substantive error was identified. As our analysis above documents, the standard of review was exceptionally lenient, meaning that the probability of receiving an exception memorandum was very low. For instance, a VLJ who decides 700 cases annually can expect to receive a memorandum in only two cases. Indeed, several interview subjects indicated that they did not recall the last time they received an exception memorandum. In the compliance literature, the received wisdom is that a higher probability of error detection increases compliance (e.g., Andreoni et al. 1998: 841–3). Given that the error rate was low during the entire observation period, the salience of QR review may simply be too low to have a program effect.

Third, one last reason to doubt that the QR program was particularly effective in its early years is that GAO found so. In its 2002 report, GAO highlighted considerable limitations in BVA's QR efforts, contradicting the notion that the 1998 establishment of QR shifted the agency into a higher accuracy equilibrium. For instance, GAO pointed out that the Board's QR database did not capture the specific issue on which an error was flagged, making it hard to assess which medical conditions were causing problems. The Board also did not collect information when quality reviewers disagreed with a VLJ, but not enough to call an error. Such information, GAO noted, would help "identify opportunities to improve the quality of decision making by improving training." GAO reported variation in decisions across Board members, but noted that "[n]o systematic study has been done to explain the variance in remand rates" between Board decision teams. These limitations undercut the idea that the QR process may have improved training, which is corroborated by our interviews that revealed that training efforts were limited. Indeed, it was the perceived inefficacy of the QR program, based on an internal (non-public) report, that led BVA to revise its QR system in 2016 to shift to more systematic, rather than individualized, detection of errors.

5.3 Exempt Officials

We also conduct a VLJ analysis to assess the evidence that simply being subjected to QR review may have had an effect. We carry out two tests. First, we leverage the fact that some senior officials were *de facto* exempted from QR. To account for differential effort by staff attorneys based on seniority, we construct a comparison group of senior officials (e.g., Chief VLJs) who were subjected to QR. Because BVA did not provide precise dates for promotion, we distinguish staff attorneys serving as Acting VLJs from VLJs based on caseload as detailed in Appendix F. If the QR program's existence had substantial effects on performance, we would expect exempted officials to have higher appeals rates than non-exempt officials.

We find the opposite: decisions written by exempt officials have lower appeals rates (p-value $=0.002$).

Our second test leverages temporal variation for one official, who was exempted from QR during the first 2 years as a senior official, but then was subjected to QR. If the application of QR to a VLJ has an effect, we should observe a discontinuous improvement after that official is subjected to QR. We find, however, no statistically significant difference before or after that official was subject to QR.[29]

While the evidence presented above does not meet the rigor of the design around random selection for QR, we believe it leads to the conclusion that the mere creation of the QR program did not have a substantial effect. If this were not true, it would suggest that quality assurance could be had at little cost: a very low probability of a correction, coupled with little sanction, could have considerable effects on performance. It is worth contrasting that with the Massachusetts welfare quality control program described by Brodkin and Lipsky (1983). Although only qualitative, Brodkin and Lipsky (1983) argue that the program led to changes in policy and practice, due to a "blitzkrieg" of interventions, including the threat of demotion and placing 17 office directors on probation when error rates were too high.

6. Limitations

We now discuss several other limitations of our study.

First, while our finding of the effect of QR is a well-identified estimate of the average causal effect on the population of BVA cases, the causal inference about the effect of the standard of review is an in-sample effect. We only observe the same standard of review being applied to a (nonrandom) *subset* of BVA decisions that are appealed, which may magnify the difference. On the other hand, BVA decisions may have been partially corrected in response to the QR team memoranda, therefore muting the difference. Nonetheless, the fact that three quarters of BVA opinions, which are deemed error-free under BVA's standard of review, are remanded by CAVC shows that the standard of review matters for a subset of cases.

Second, while many scholars have viewed immigration courts, social security adjudication, and the BVA as close institutional cousins (Verkuil 1991, 2017; Congressional Research Service 2012; Asimow 2016; Sabel and Simon 2017; Gelbach and Marcus 2018), our evidence may have limited external validity for other quality improvement programs. SSA's programs, for instance, have rapidly evolved, making much greater use of technology (Ray and Lubbers 2015). The peer review program for immigration courts appeared to be much more of a training program. We believe the political and institutional tension in

29. For details, see Appendix F.

performance measurement, however, is likely common to many of these systems. Most importantly, without opening the black box of the agency, it is simply not possible to know. As Merrill (2017: 59) notes, whether the "internal law of administration . . . work[s] well in administrative schemes" is "a serious objection and can be answered only by undertaking further empirical investigations." Many surface descriptions of QR programs may appear compelling. Indeed, none other than Mashaw pointed to the VA's system of statistical quality assurance as an exemplar for internal management (Mashaw 1973a).[30] Moreover, our findings speak directly to current efforts at the BVA. In 2017, as part of a push for a renewed focus on reducing the backlog, BVA abandoned its 2016 reforms, returning to the system we studied here, but reducing the sampling rate and QR staff. The fact that a more intensive review process yielded few benefits suggests that the prospective reform is unlikely to address the longstanding quality problems in BVA adjudication.

7. Conclusion

Our study is the first to leverage randomization of QR to credibly assess its effects on case outcomes and contributes to central questions of administrative justice. We conclude with several implications.

First, the divergence in the BVA and CAVC standards of review highlights tensions in the role of judicial review in mass adjudicatory systems. The Veterans' Judicial Review Act of 1988 imported a model of adversarialism that posed a tension with VA's historical model of paternalistic charity (Cragin 1994; Ridgway 2010). BVA's internal rejection of CAVC's demand for reasoned explanation—notwithstanding a 75% remand rate when no errors are called—illustrates the continuing internal conflict around these models. Our findings raise important questions about the costs of judicialization of mass adjudication and its impact on veterans (Mashaw 1985). Procedure has substantive impact. It now takes an average of 7 years from filing a notice of disagreement to Board resolution (Department of Veterans Affairs 2018). VA's Inspector General estimated that 7% of VBA appeals were deemed "resolved" in the first quarter of 2016 because the veteran died while waiting for a decision (VA Office of Inspector General 2018).

Second, if judicial review has not solved these problems, our results also paint a sobering picture about the ability for an agency to internally develop such quality assurance initiatives. The degradation of BVA's QR challenges more optimistic accounts of bureaucratic rationality (Mashaw 1983) and internal administrative law (Metzger and Stack 2016). Many have suggested random audits as the cure for mismanagement (Bevan and Hood 2006; Cuéllar 2006), but our evidence shows that random audits may be insufficient when agency supervisors have the discretion to adjust

30. To be clear, Mashaw did not discuss any system for quality assurance or review at the BVA.

audit criteria and performance metrics. The lenient internal standard of review and exclusion of cases by senior managers underscore the importance of separation of functions and institutional independence of QR (Mashaw 1973b).

Third, our results suggest that case-specific QR cannot remedy structural challenges stemming from the volume of cases. Errors stemming from caseload cannot easily be addressed by adding to caseload. BVA's efforts in 2016 to reform the QR program to focus less on case-specific review and more on feedback at the systemic level may be more promising (Ho 2017; Gelbach and Marcus 2018).

Fourth, our findings illustrate the difficulty of performance measurement in the public sector (Dixit 2002), popularized by Osborne and Gaebler (1993) that inspired the GPRA.[31] The BVA's inflated accuracy rate can be conceived of as a form of supervisor–agent collusion (Tirole 1986) or as an example of biased peer review given the connections between QR staff attorneys and VLJs (Blanes i Vidal and Leaver 2015). Similar strategizing around performance measures in labor training programs led Barnow (2000) to find only weak evidence of a correlation between performance measures and program impact based on randomized controlled trials. Our evidence demonstrates that performance measurement is not just uncorrelated with, but can actually undermine, program impact.

Most generally, the changing standard for accuracy exemplifies the "quantity–quality" tradeoff that is the subject of much public administration scholarship (e.g., Bevan and Hood 2006). While accuracy was for years the first performance measure featured in budget requests and continues to be reported in the Board's annual reports, accuracy rates were removed from BVA's budget requests starting in 2010. In 2017, some 100 Board attorneys signed a loss of confidence statement, sent to House and Senate Veterans Affairs committees. The statement argued that the production quota, mismanagement, and inadequate training would effectively render the Board's *de novo* standard—meant to "ensure[] accuracy"—"meaningless." In contrast to the Acting Chairman's 1998 declaration that quality was BVA's "single most important goal," the agency's own performance measures now reflect a fixation on that which is easily measured: caseload.

Appendix

A. Additional Balance Checks

A.1 Balance Statistics for Additional Covariates

Because of space constraints, Table 1 presented balance on only the most salient covariates. Table A1 presents balance on additional covariates not shown in the main balance table.

31. For a critique of performance measurement under the GPRA and its implications on disability adjudication, see Mashaw (1996).

Table A1. Additional Balance Checks on Selected Covariates between Cases Not Selected for QR (Ctrl) and Cases Randomly Selected for QR between August 1, 2003 and November 9, 2016

	Original cases				CAVC-remanded cases			
	Ctrl.	QR	Diff.	*p*-value	Ctrl.	QR	Diff.	*p*-value
Compensation issue types (number of issues per case)								
TDIU—entitlement	0.09	0.09	0.00	0.91	0.14	0.15	0.01	0.74
DIC—service connection cause of death	0.03	0.03	−0.00	0.91	0.04	0.04	−0.00	0.74
Effective date: service connection grant or severance	0.03	0.03	−0.00	0.91	0.05	0.04	−0.00	0.74
Compensation: increased rating/other	0.01	0.01	−0.00	0.99				
Number of SSOCs submitted (prop. of cases)								
First	0.52	0.53	0.00	0.91	0.39	0.39	0.00	0.84
Second	0.20	0.20	0.00	0.95	0.07	0.07	−0.00	0.88
Third	0.07	0.07	0.00	0.74	0.02	0.02	0.00	0.84
Fourth	0.03	0.03	0.00	0.91	0.01	0.01	−0.00	0.74
Fifth	0.01	0.01	0.00	0.95				
Case documents (count per case)								
Physical claims folders	1.60	1.61	0.01	0.88	2.28	2.35	0.07	0.35
Physical medical folders	0.02	0.02	−0.00	0.99	0.02	0.01	−0.01	0.40
Service department records envelopes	0.72	0.72	−0.00	0.99	0.57	0.59	0.02	0.74
eFolder documents	27.34	26.34	−0.99	0.74	52.61	50.73	−1.87	0.74
Hearings (count per case)								
Total	0.27	0.26	−0.00	0.99	0.03	0.03	0.00	0.74
Central office	0.01	0.01	−0.00	0.91				
Travel board	0.15	0.15	0.00	0.99	0.01	0.01	0.00	0.88
Videoconference	0.10	0.10	−0.00	0.91	0.01	0.02	0.00	0.74

(continued)

Table A1. Continued

	Original cases				CAVC-remanded cases			
	Ctrl.	QR	Diff.	*p*-value	Ctrl.	QR	Diff.	*p*-value
DRO hearings (prop. cases)								
Review	0.37	0.37	0.00	0.99	0.02	0.02	0.00	0.40
Formal hearing	0.05	0.05	0.00	0.91				
Informal hearing	0.01	0.01	0.00	0.74				
Medical opinions (prop. cases)								
Medical opinion from Veterans Health Administration (prop.)	0.01	0.01	0.00	0.91	0.02	0.03	0.00	0.84
Independent Medical Expert Opinion (IME) [20.901(d)] only					0.00	0.01	0.00	0.74
Hearings (prop. cases)								
Travel board	0.24	0.24	0.00	0.80	0.24	0.23	−0.00	0.74
Informal	0.13	0.14	0.00	0.74	0.12	0.13	0.00	0.99
Videoconference	0.15	0.14	−0.00	0.80	0.13	0.13	0.00	0.74
Field office	0.06	0.06	0.00	0.88	0.09	0.10	0.01	0.26
Central office	0.02	0.02	−0.00	0.99	0.03	0.03	−0.00	0.88
Issue diagnosis categories (number of issues per case)								
Nontuberculous disease of lungs and/or pleura	0.05	0.05	−0.00	0.99	0.03	0.03	0.00	0.88
Endocrine system disability	0.05	0.05	−0.00	0.91	0.04	0.04	−0.00	0.98
Heart disease	0.05	0.05	0.00	0.99	0.04	0.04	−0.00	0.60
Misc. neurological diseases	0.04	0.04	−0.00	0.92	0.04	0.04	0.00	0.74
Eye disability	0.05	0.04	−0.00	0.74	0.03	0.03	0.00	0.84

(continued)

Table A1. Continued

	Original cases				CAVC-remanded cases			
	Ctrl.	QR	Diff.	*p*-value	Ctrl.	QR	Diff.	*p*-value
Disease of trachea and/or bronchi	0.04	0.04	−0.00	0.88	0.03	0.03	0.00	0.88
Disease of nose or throat	0.04	0.04	0.00	0.99	0.02	0.02	−0.00	0.40
Central nervous system disease	0.03	0.03	0.00	0.99	0.02	0.02	0.00	0.74
Muscle injury	0.02	0.02	0.00	0.91	0.02	0.02	−0.00	0.74
Infectious disease, immune disorder, or nutritional deficiency	0.01	0.01	−0.00	0.95	0.01	0.01	−0.00	0.84
Dental or oral condition	0.01	0.01	−0.00	0.91	0.01	0.01	0.00	0.84
Peripheral nerve neuritis	0.01	0.01	−0.00	0.22	0.01	0.01	−0.00	0.40
Psychotic disorder	0.01	0.01	0.00	0.99	0.02	0.03	0.00	0.74
Hemic or lymphatic system disability	0.01	0.01	−0.00	0.91	0.01	0.01	0.00	0.84
Gynecological or breast disability	0.01	0.01	−0.00	0.99				
Undiagnosed condition	0.01	0.01	0.00	0.99	0.01	0.01	−0.00	0.99
Organic mental disorder	0.01	0.01	0.00	0.91	0.01	0.01	0.00	0.88
Peripheral nerve neuralgia	0.01	0.01	−0.00	0.99				
Epilepsy	0.01	0.01	−0.00	0.91	0.01	0.00	−0.00	0.74
Number of BVA cases (sample size)	508,801	26,821			47,981	5622		

Notes: Cases are split by whether they had been remanded by CAVC leading to the decision at issue. DIC stands for dependency and indemnity compensation. TDIU stands for total disability rating due to individual unemployability. SSOC stands for supplemental statement of the case. *p*-values are adjusted for multiple testing using Benjamini and Hochberg (1995).

A.2 Equivalence Tests

Recent work has highlighted the importance of considering equivalence regions when establishing the validity of a research design (Hartman and Hidalgo 2018). Unlike traditional tests of balance, equivalence tests propose the null hypothesis to be a difference between treatment and control groups, while the alternative is equivalence. As an additional robustness check to ensure proper replication of the QR random sampling process, we perform equivalence tests using the balance covariates reported in Tables 1 and A7. Due to the difficulty of determining substantively informed equivalence regions for so many covariates, we follow the recommendations of Hartman and Hidalgo (2018) and set an equivalence region of $\pm 0.36 \times$ pooled SD. We are able to the reject the null hypothesis of a difference between control and QR cases for each covariate. Figure A1 plots the 95% CIs (in units of pooled SDs) obtained by inverting the equivalence tests, along with the benchmark of ± 0.36 SDs for reference. The CIs for all covariates lie within ± 0.08 pooled SDs from zero, well within the balance benchmark.

B. VLJ-Specific Balance

To assess VLJ-specific treatment effects, we present balance diagnostics at the VLJ level. The left panel of Figure A2 plots number of control opinions by a VLJ on the *x*-axis against the QR selection rate for that VLJ on the *y*-axis for original decisions. As expected, the selection rate is centered around 5%, with VLJs with fewer opinions exhibiting higher sampling variability. We also discovered through this balance check that there was a cluster of individuals whose cases had unexpectedly low selection rates, as indicated by the cluster in the lower left corner of the left panel. These individuals are all part of the senior management team (e.g., Chairman of the Board or Chief Counsel for Policy and Procedure). Upon verifying with staff, it appears that these senior managers were excluded from having their cases undergo QR because of the potential conflict. We hence exclude these individuals from our QR-eligible cases.

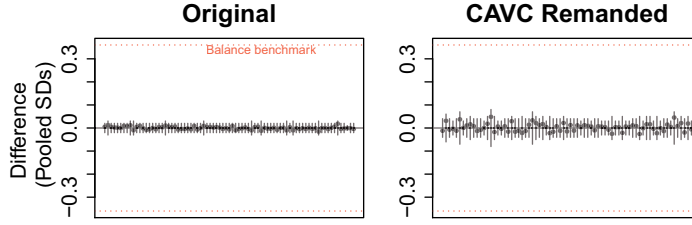

Figure A1. Ninety-Five Percent CIs for Equivalence Tests (in units of pooled SDs) Conducted on Each Balance Covariate, Relative to the Benchmark for Balance Suggested by Hartman and Hidalgo (2018) (dashed line).
Note: Cases are split by original (left) and CAVC-remanded cases (right).

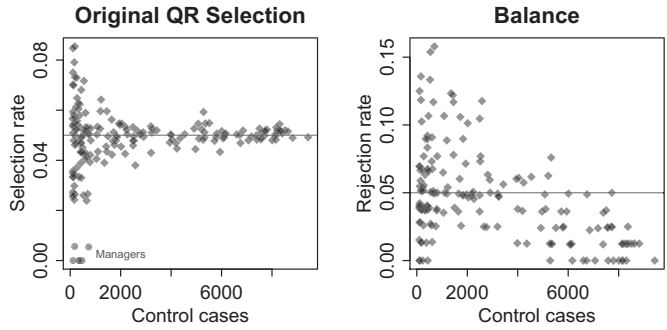

Figure A2. QR Selection Rate within VLJs for Original Cases by Control Case Volume (left); Balance Tests within VLJs by Control Case Volume (right).

Notes: Each dot represents one VLJ. The rejection rate indicates the fraction of balance tests that reject the null hypothesis, with the horizontal line indicating $\alpha = 0.05$ and sampling variability generating higher rejection rates with VLJs deciding few cases. VLJs represented by the round dots in the left panel were excluded from QR because of their managerial duties, as evidenced by the low selection rate.

Table A2. Error Codes in the QR Program Effective between June 1, 1999 and August 1, 2015 with the Frequency Counts by Case Type in the QR-Eligible Sample

Error description (1999–2015)	Original cases	CAVC cases
Conclusions of law		
Erroneous conclusion	3	3
Failure to address every relevant theory of entitlement	25	3
Due process		
Fair process violation (Bernard, Colvin, Thurber)	43	9
Board jurisdictional error (e.g., Marsh, Barnett)	29	6
Inadequate development	211	33
Failure of duty to notify	266	25
Procedural deficiency: hearing	28	4
Procedural deficiency: representation	83	32
Procedural deficiency: 38 C.F.R 20.1304	26	2
Procedural deficiency: other	58	21
Findings of fact		
Inaccurate finding	23	2
Necessary elements of claim not addressed	15	0
Issues		
Raised but undeveloped issue(s) omitted	19	4
Developed issue(s) omitted	33	19
Inaccurate issues(s) set forth on title page	16	5
Reasons or bases		
Legal authority misapplied: Case Law	93	7
Inadequate explanation: conclusionary discussion	235	34
Inadequate explanation: relevant theory not addressed	110	18
Legal authority misapplied: law or regulation	106	11
Legal authority misapplied: precedent opinion	25	5
Legal authority not applied: case law	289	37
Legal authority not applied: law or regulation	213	25
Legal authority not applied: precedent opinion	35	3
Incorrect standard of proof	13	5
Inadequate explanation: material evidence omitted	273	40
Inadequate explanation: deficient credibility determ	98	11
Total errors called	2368	364
Total cases sampled for QR	24,019	4868

Table A3. Error Codes in the QR Program Effective between August 1, 2015 and October 1, 2017 with the Frequency Counts by Case Type in the QR-Eligible Sample

Error description (2015–2017)	Original cases	CAVC cases
Due process		
Fair process violation	2	3
Duty to notify	0	0
Duty to assist	14	6
Procedural deficiency: hearing	1	0
Procedural deficiency: representation	20	5
Procedural deficiency: 38 CFR 20.1304	4	0
Other	10	1
Issues		
Failure to address developed claim/issue	9	2
Failure to address reasonably raised claim/issue	8	1
Board jurisdictional problem	6	4
Board policy	0	0
Single/separate decisions	2	2
Other	0	0
Reasons and bases		
Failure to consider (FtC): theory/contentions	20	7
ID: OGC precedential opinion/administrative procedure	0	0
Misapply: OGC precedential opinion/administrative procedure	0	0
FtC: service records	2	1
FtC: VA records	11	3
FtC: non-VA federal records	2	0
FtC: private records	9	0
FtC: lay evidence	18	4
Lay evidence: competency	6	1
Lay evidence: credibility	2	0
Conclusory discussion	5	0
Inadequate discussion (ID): theory/contentions	4	0
Other	18	2
FtC: case law	33	6
ID: case law	1	1
Misapply: case law	6	2
FtC: statute and/or regulation	10	2
ID: statute and/or regulation	3	0
Misapply: statute and/or regulation	6	3
FtC: OGC precedential opinion/administrative procedure	0	0
Remands		
Improper development on remand	4	1
Unnecessary development on remand	9	4
Total errors called	245	61
Total cases sampled for QR	2802	754

The right panel of Figure A2 plots the number of control opinions by a VLJ on the x-axis against the rejection rate of balance across 80 covariates on the y-axis. As expected, chance imbalances are much higher for VLJs with fewer opinions. We hence use a cutoff of 4000 cases to estimate VLJ-specific treatment effects in Section 4.1.

C. QR Coding of Case Errors

For reference, this Appendix presents a more fine-grained error coding used by BVA's QR team. Figure 5 presented error rates at the highest level of aggregation (e.g., reasons and bases). BVA used one set of QR error codes from 1999 to 2015, presented in Table A2, and another set of QR error codes from 2015 to 2017, presented in Table A3.

D. CAVC Remand Reasons

This Appendix presents the BVA's coding of remands from CAVC. While BVA carried out more fine-grained issue codes, because these were switched in 2013, we present in Tables A4 and A5 each of the subcodes only to clarify how we aggregated the available remand reasons into (a) due process or (b) reasons and bases for the analysis in Figure 5.

Table A4. CAVC Remand Reasons Pre-July 2013, Divided by Error Type to Align with the QR Process

CAVC remand reason (pre-July 2013)	Cases
Due process	
Apply new caselaw	1697
Apply new legislation/regulation (Karnas)	1530
Failure to comply with prior Remand (Stegall)	1371
Other due process violation	1087
Consider new arguments (Maggitt)	187
Need AOJ consideration (prejudice under Bernard)	175
Hearing required	94
Reasons and bases	
Inadequate discussion	11,940
Failure to address credibility/evidence	5726
Laws/regulations	3751
Existing caselaw	3178
Incomplete findings, conclusions, etc. (Hensley)	801
Colvin violation	358
GC opinions	171
Administrative issue	49
Other	
Medical exam required	4138
Duty to notify	2997
VA medical records	750
Service department records	697
Other duty to assist violation	678
Private medical records	491
Social security admin records	420
Center for research of unit records	253

Table A5. CAVC Remand Reasons Post-July 2013, Divided by Error Type to Align with the QR Process

CAVC Remand Reason (Post-July 2013)	Cases
Due process	
Due process Failure to comply with prior remand (Stegall) from Board	810
Due process Inextricably intertwined	547
Due process Failure to comply with prior remand (Stegall) from Court	210
Due process Other due process violation	202
Due process Failure to adjudicate claim/issue	152
Due process Apply new law/regulation/case law	75
Due process Need AOJ consideration (prejudice under Bernard)	38
Due process Hearing required	36
Due process Offer hearing or request clarification	17
Due process Foreign language translation required	7
Reasons and bases	
Other R&B deficiency existing case law	3300
Reasons and bases failure to consider existing case law	1864
R&B failure to adequately address—duty to assist inadequate medical opinion	1819
Reasons and bases failure to consider VA medical evidence	1382
Reasons and bases failure to consider Lay evidence	1251
Other R&B deficiency statute or regulation	1068
R&B failure to adequately address—duty to assist inadequate medical exam	988
Reasons and bases failure to consider theory of entitlement/ contentions	836
Other R&B deficiency lay evidence credibility	753
Reasons and bases failure to consider statute or regulation	705
Other R&B deficiency diagnostic code	689
Other R&B deficiency VA medical evidence	649
Reasons and bases misapplication of law/regulation/caselaw/GC opinion	581
Other R&B deficiency weighing of conflicting evidence	568
Reasons and bases failure to consider private medical evidence	522
Reasons and bases failure to consider other	516
Other R&B deficiency other	498
Reasons and bases mischaracterization of evidence	429
Reasons and bases inconsistent/contradictory findings	396
Other R&B deficiency theory of entitlement/contentions	347
Reasons and bases failure to consider diagnostic code	336
Other R&B deficiency lay evidence competency	330
Reasons and bases colvin violation	211
R&B due process issue apply new law/regulation/case law	207
Reasons and bases failure to consider service treatment/personnel records	187
Other R&B deficiency private medical evidence	182
R&B due process issue failure to comply with prior remand (Stegall) from board	173
R&B due process issue inextricably intertwined	129

(continued)

Table A5. Continued

CAVC Remand Reason (Post-July 2013)	Cases
Reasons and bases failure to consider administrative issue/procedure	123
R&B failure to adequately address—duty to assist VA medical records	106
R&B failure to adequately address—duty to assist other records	83
Other R&B deficiency service treatment/personnel records	80
R&B failure to adequately address—duty to assist private medical records	71
R&B failure to adequately address—duty to assist service treatment records	69
R&B due process issue failure to adjudicate claim/issue	66
R&B due process issue other due process violation	65
Reasons and bases failure to consider social security/other federal records	59
Other R&B deficiency administrative issue/procedure	51
R&B due process issue failure to comply with prior remand (Stegall) from court	40
Reasons and bases failure to consider GC precedent opinion	32
R&B failure to adequately address—duty to notify at hearing (Bryant)	29
R&B failure to adequately address—duty to assist service personnel records	29
Other R&B deficiency social security/other federal records	28
R&B failure to adequately address—duty to assist JSRRC	22
Other R&B deficiency GC precedent opinion	21
R&B failure to adequately address—duty to assist SSA records	17
R&B failure to adequately address—duty to notify incorrect/legally inadequate notice sent	15
R&B due process issue need AOJ consideration (prejudice under Bernard)	12
R&B due process issue hearing required	7
R&B failure to adequately address—duty to assist vocational rehabilitation records	5
R&B due process issue offer hearing or request clarification	5
R&B failure to adequately address—duty to notify no notice sent	4
R&B failure to adequately address—duty to notify no notice of inability to obtain federal records	2
R&B failure to adequately address—duty to notify no notice of inability to obtain non-federal records	2
R&B due process issue foreign language translation required	1
Other	
Duty to assist medical examination/opinion required	2037
Duty to assist VA medical records	419
Duty to assist private medical records	300
Duty to assist service treatment records	131
Duty to assist service personnel records	102
Duty to assist SSA records	65
Duty to assist JSRRC	62
Duty to notify at hearing (Bryant)	59
Duty to notify incorrect/legally inadequate notice sent	28

(continued)

Table A5. Continued

CAVC Remand Reason (Post-July 2013)	Cases
Duty to notify no notice of inability to obtain federal records	19
Duty to assist vocational rehabilitation records	12
Duty to notify no notice of inability to obtain non-federal records	11
Duty to notify no notice sent	8
Duty to assist workers compensation records	5

E. Robustness

E.1 Issue Outcomes

Our principal outcomes for CAVC dispositions in Table 2 focus on whether relief is granted for at least one issue in the case. Table A6 presents results based on counts of *issues*. The effects are substantively identical. The QR treatment, for instance, has no distinguishable effects on the number of issues that are vacated and remanded.

Table A7 carries out the analysis of CAVC disposition of Table 5 with counts of *issues*. Again, the results are comparable. When the QR team calls an error, CAVC is more likely to vacate and remand original cases.

E.2 Denials

Because the Board reviews all issues, but CAVC largely reviews denials, we subset our QR-eligible sample of cases to those with at least one denial and re-run the models in Table A8. We find that the QR effect on both outcomes for original cases remains null, and the QR effect on appeals attenuates for CAVC-remanded cases to statistical insignificance. This suggests that the pooled effect on appeals for CAVC-remanded cases in Table 3 is being driven by QR review of decisions with allowances. This might be plausible if in the first CAVC decision, CAVC remanded on a denial for reconsideration, the initial VLJ decision provided an insufficient explanation for a continued denial, and the QR process converted such a denial into an allowance. There are, however, reasons to doubt whether this is a meaningful effect, largely because the QR process tends to focus on denials, as those are the likely cases to be appealed to CAVC.

E.3 CAVC Judge-Fixed Effects

As an additional robustness check, we subset the QR-eligible sample to appeals with available CAVC judge data and add CAVC judge-fixed effects to the models in Table 3, which estimate the probability of a reversal or remand by CAVC, conditional on an appeal. The subset for this robustness check represents only 9% of CAVC appeals in the QR-eligible sample, because the Board did not record CAVC judges in VACOLS for any decision for the majority of the observation window and recorded CAVC judge data inconsistently from mid-2013 onward. We further subset the sample to include only CAVC appeals decided by a single

Table A6. Means and Differences-in-Means (Diff.) for Outcomes, Comparing Control (Crtl.) Cases Not Selected for QR and Treatment Cases Randomly Selected for QR between August 1, 2003 and November 9, 2016

Issue outcome at CAVC	Original cases				CAVC-remanded cases			
	Ctrl.	QR	Diff.	*p*-value	Ctrl.	QR	Diff.	*p*-value
Vacated and remanded	1.15	1.12	−0.03	0.89	1.03	0.98	−0.05	0.88
Affirmed	0.36	0.36	0.00	0.97	0.39	0.41	0.02	0.90
Abandoned	0.36	0.34	−0.02	0.89	0.22	0.21	−0.02	0.90
Dismissed	0.17	0.16	−0.01	0.89	0.10	0.12	0.02	0.90
Reversed	0.01	0.01	−0.00	0.89	0.02	0.02	−0.00	0.98
Total issues	2.07	2.01	−0.06	0.70	1.79	1.75	−0.04	0.90
Sample sizes (CAVC appeals)	31,590	1604			6782	728		

Notes: The left panel presents data for original cases and the right panel presents cases for CAVC-remanded cases. Outcomes are all actions taken after dispatch of the BVA decision. Because each decision can involve multiple issues, "issue outcome" presents the average number of issues subject to each disposition. *p*-values are adjusted for multiple tests using Benjamini and Hochberg (1995). For readability, we exclude disposition codes with very low case counts (i.e., vacated and dismissed, settled, and dismissed due to death), which also have no statistically significant differences between QR and control cases.

Table A7. Means and Differences-in-Means (Diff.) for the Subset of Cases that went through QR, Comparing Cases with No Errors Found (No Error) and Cases with Some Error Found (Error) between August 1, 2003 and November 9, 2016

Issue outcome at CAVC	Original cases				CAVC-remanded cases			
	No error	Error	Diff.	*p*-value	No error	Error	Diff.	*p*-value
Vacated and remanded	1.09	1.32	0.23	0.02	0.95	1.25	0.30	0.20
Affirmed	0.38	0.24	−0.14	0.02	0.41	0.42	0.02	1.00
Abandoned	0.31	0.52	0.20	0.08	0.20	0.34	0.15	0.46
Dismissed	0.15	0.21	0.06	0.70	0.11	0.19	0.08	0.77
Reversed	0.01	0.00	−0.00	0.70	0.02	0.03	0.02	0.73
Total issues	1.97	2.31	0.34	0.06	1.71	2.25	0.54	0.20
Sample sizes (appeals)	1392	212			664	64		

Notes: Formatting errors are excluded. The left panel presents data for original cases and the right panel presents cases for CAVC-remanded cases. Outcomes are all actions taken after dispatch of the BVA decision. Because each decision can involve multiple issues, issue outcome presents the average number of issues subject to each disposition. *p*-values are adjusted for multiple tests using Benjamini and Hochberg (1995). For readability, we exclude disposition codes with very low case counts (i.e., vacated and dismissed, settled, and dismissed due to death), which also have no statistically significant differences between QR and control cases.

Table A8. Robustness Check Including Only Cases with at Least One Issue Denied

Outcome		Original cases			CAVC-remanded cases		
Appealed to CAVC	QR effect	−0.023	−0.021	−0.021	−0.059	−0.055	−0.054
		(0.028)	(0.028)	(0.028)	(0.051)	(0.051)	(0.052)
	VLJ FEs	N	N	Y	N	N	Y
	Year-quarter FEs	N	Y	Y	N	Y	Y
	N	255,334	255,334	255,334	18,094	18,094	18,094
Reversed/ remandedby CAVC	QR effect	−0.034	−0.044	−0.036	−0.115	−0.096	−0.115
		(0.062)	(0.062)	(0.063)	(0.089)	(0.090)	(0.093)
	VLJ FEs	N	N	Y	N	N	Y
	Year-quarter FEs	N	Y	Y	N	Y	Y
	N	31,638	31,638	31,638	7052	7052	7052

***Notes*: Logistic regression results of the probability of an appeal to CAVC (top panel) and the probability of a reversal or remand by CAVC, conditional on an appeal (bottom panel) for original cases (left columns) and CAVC-remanded cases (right column). The QR effect row presents the coefficient on the treatment indicator, with SEs in parentheses. FEs indicate fixed effects estimated using the pseudo-demeaning algorithm described in Stammann et al. (2016), which are not displayed for readability; *N* indicates sample size. $p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table A9. Robustness Check Including Fixed Effects for CAVC Judges in Logistic Regressions Estimating the Effect of QR on the Probability of Reversal or Remand by CAVC, Conditional on an Appeal

	Original cases				CAVC-remanded cases			
QR effect	−0.109	−0.107	−0.112	−0.029	−0.082	−0.082	−0.091	−0.092
	(0.170)	(0.171)	(0.178)	(0.177)	(0.252)	(0.257)	(0.294)	(0.281)
VLJ FEs	N	N	Y	Y	N	N	Y	Y
Year-quarter FEs	N	Y	Y	Y	N	Y	Y	Y
CAVC judge FEs	N	N	N	Y	N	N	N	Y
N	2916	2916	2916	2916	792	792	792	792

***Notes*: The data include only the subset of QR-eligible cases with CAVC judge data. The left panel presents results for original cases, and the right panel presents results for CAVC-remanded cases. The QR effect row presents the coefficient on the treatment indicator, with SEs in parentheses. FEs indicate fixed effects, which are not displayed for readability; *N* indicates sample size. $p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

judge, excluding the 1.7% of QR-eligible appeals decided by a panel of three CAVC judges.

In Table A9, we replicate the results presented in Table 3 with the subset of appeals with CAVC judge data and include fixed effects for CAVC judges. We find that the inclusion of CAVC judge-fixed effects does not substantively alter the results.

F. Exempt Officials

We here detail how we tested for differences between senior officials exempted and non-exempted from the QR program, so as to assess whether simply being subjected to QR improved decision quality.

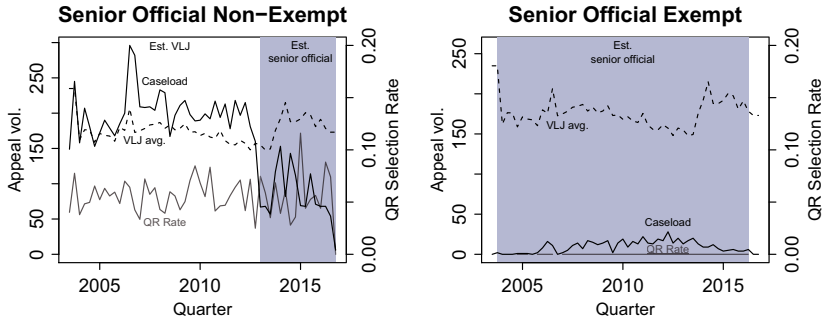

Figure A3. Classification of Senior Position Periods for a Senior Official Subjected to QR (left) and a Senior Official Exempt from QR (right).
Notes: The solid black line indicates the official's caseload and the solid grey line represents the proportion of their cases that were selected for QR. The dashed line indicates the average case volume for non-senior, line VLJs. The gray shaded area indicates the estimated time period during which the official held a senior position.

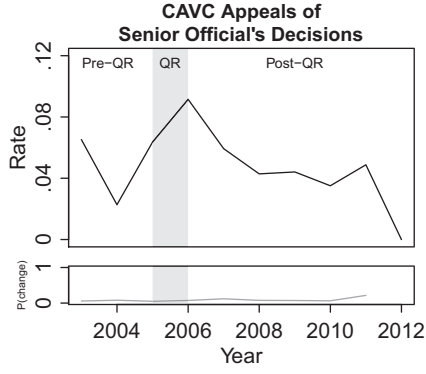

Figure A4. Change in QR Exemption Status on the Appeals Rate of One Senior Official.
Notes: The top panel depicts the proportion of the official's decisions that were appealed to CAVC each year. The gray shaded region represents the period during which the official's decisions were selected for QR. The bottom panel plots the posterior probability of a structural break in the official's appeals rate, using the Bayesian change point model of Barry and Hartigan (1993).

F.1 Measurement

Senior positions include: Chief VLJs, Deputy Vice Chairman, Principal Deputy Vice Chairman, Chairman of the Board, Chief Counsel for Policy and Procedure, Chief Counsel for Operations, and Director of the Office of Management, Planning and Analysis. While public sources provide the fiscal years during which employees held senior positions, the Board did not disclose the precise dates of when promotions or demotions occurred.

As further detailed in this Appendix, we identify start and end dates of senior positions through changes in caseload volume, because senior officials are responsible for fewer cases than non-senior employees. We combine this information with the QR selection rate and information from interviews to identify senior positions that were exempt and not exempt from QR.

First, using VACOLS data, we calculated quarterly caseload volumes for officials who held senior positions at any point during the observation window. Second, we calculated the average caseload for line VLJs for each quarter, excluding periods during which employees were serving as Acting VLJs. Third, service periods were classified as acting VLJs if the individuals decided a low caseload (i.e., fewer than 100 decisions per quarter) prior to assuming a more conventional VLJ caseload. (100 cases are a natural breakpoint in the distribution of quarterly case volumes.) Fourth, to identify precise senior staff periods, we classified senior periods as those when employees' quarterly case volume dropped below the baseline line VLJ caseload for at least two quarters. Last, to identify which senior officials were exempt from QR, we drew upon the QR selection rate and insights from interviews, as previously discussed in Appendix B. We exclude from this group individuals with direct oversight of the QR program, leaving us with three senior exempt officials.

Figure A3 depicts examples for how we classified service periods. The left panel plots case production time series of a VLJ who assumed a senior official position around 2013, but whose opinions continued to be subjected to QR. The right panel plots the case production time series for a senior official who was categorically exempt from QR.

F.2 Analysis

Using only cases decided by senior officials, we fit a logistic regression of the probability of an appeal to CAVC, with an indicator for whether the official was exempt and year-quarter fixed effects to adjust for time differences. (Due to the small sample size, we do not analyze CAVC-remanded cases.) Contrary to the expectation that QR improved outcomes, exempted senior officials had lower appeals rates than non-exempted officials (p-value $=0.002$).

We then analyze the decisions of one senior official whose exposure to QR varied during the observation window. The official was exempt from the program's inception until 2005, but between late 2005 and early 2006, the official's decisions were selected for QR. By late 2006, the official's decisions were no longer subjected to QR until his retirement from the Board in 2012.

We test for differences in decision quality, as measured by CAVC appeals rates, before and after the official's decisions were subjected to QR. Figure A4 presents the time series of the appeals rate in the top panel, and posterior probabilities of a structural break in the bottom panel (Barry and Hartigan 1993). The official's highest appeals rates occurred during

the QR period, but there is no evidence for structural changes coinciding with the application of QR.

G. CAVC Appeals Selection

To provide more context for understanding the CAVC appeals rate, we present differences between cases that are not appealed and cases that are appealed to CAVC. At the outset, it is important to note that this

Table A10. Descriptive Statistics for BVA Cases Not Appealed to CAVC and BVA Cases Appealed to CAVC

	Non-appealed cases		Appealed cases			
	Mean	*SE*	Mean	*SE*	Diff.	*p*-value
Appellant age (years, at notice of disagreement)	55.49	0.02	56.02	0.06	−0.53	0.00***
Appellant service period (prop.)						
WWII (9/16/40–7/25/47)	0.08	0.00	0.07	0.00	0.01	0.16
Peacetime (7/26/47–6/26/50)	0.04	0.00	0.03	0.00	0.01	0.00***
Korean conflict (6/27/50–1/31/55)	0.09	0.00	0.09	0.00	0.00	0.05
Post-Korea (2/1/55–8/4/64)	0.14	0.00	0.15	0.00	−0.01	0.00***
Vietnam Era (8/5/64–5/7/75)	0.49	0.00	0.52	0.00	−0.03	0.00***
Post-Vietnam (5/8/75–8/1/90)	0.34	0.00	0.34	0.00	0.00	0.00***
Persian Gulf (8/2/90–Present)	0.25	0.00	0.18	0.00	0.07	0.00***
Issues per case	2.55	0.00	2.84	0.01	−0.29	0.00***
Compensation issue types (number of issues per case)						
Service connection						
All others	1.41	0.00	1.41	0.01	0.00	1.00
New and material	0.07	0.00	0.07	0.00	0.00	0.00***
Accrued benefit	0.02	0.00	0.03	0.00	−0.01	0.00***
Increased disability rating						
Schedular	0.63	0.00	0.73	0.01	−0.10	0.00***
Schedular and extraschedular	0.04	0.00	0.06	0.00	−0.02	0.00***
Extraschedular	0.02	0.00	0.03	0.00	−0.01	0.00***
Compensation/increased rating/other	0.01	0.00	0.01	0.00	0.00	1.00
TDIU—entitlement	0.10	0.00	0.14	0.00	−0.04	0.00***
Effective date—service connection grant or severance	0.03	0.00	0.06	0.00	−0.03	0.00***
DIC—service connection cause of death	0.04	0.00	0.05	0.00	−0.01	0.00***
Representation at BVA						
Disabled American Veterans	0.30	0.00	0.34	0.00	−0.04	0.00***
A State Service Organization	0.17	0.00	0.08	0.00	0.09	0.00***
American Legion	0.18	0.00	0.18	0.00	0.00	0.78
Veterans of Foreign Wars	0.10	0.00	0.05	0.00	0.05	0.00***
Unrepresented	0.10	0.00	0.07	0.00	0.03	0.00***
Attorney	0.08	0.00	0.21	0.00	−0.13	0.00***
Prior BVA decision in case history (prop.)	0.42	0.00	0.56	0.00	−0.14	0.00***
Sample size	548,521		40,704			

Notes: BVA cases include only those eligible for QR between August 1, 2003 and November 9, 2016. DIC stands for dependency and indemnity compensation. TDIU stands for total disability rating due to individual unemployability. *p*-values are adjusted for multiple testing using Benjamini and Hochberg (1995).

comparison is *not* informative about the causal effect of QR. But it does allow us to assess substantive differences between cases that CAVC decides and the population of BVA decisions.

Table A10 displays means, standard errors (SEs), and the *p*-value from a *t*-test for differences of selected covariates. Because of the sample size, many differences are statistically significant, even if the substantive magnitude is not large. The most prominent difference is in whether an attorney represented the claimant at the BVA stage. Typically, a non-attorney official from a veterans service organization (e.g., American Legion, Disabled American Veterans) represents the claimant at the BVA stage. In 8% of unappealed cases, an attorney represents the claimant, compared with 21% of appealed cases. This sharp difference is consistent with the notion that the 6% overall CAVC appeals rate (and 14% appeals rate of decisions with at least one denial) stems from the shift from an administrative adjudication to a more formal adversarial hearing—typically seen as requiring an attorney—at CAVC. In that sense, the CAVC appeals rate is quite consistent with the appeals rate from the SSA to district court (see footnote 8).

Conflict of interest statement. None declared.

References

Administrative Conference of the United States. 1978. "Recommendation 78-2: Procedures for Determining Social Security Disability Claims," *Federal Register*.

Ames, David, Cassandra Handan-Nader, Daniel E. Ho, and David Marcus. 2020. "Due Process and Mass Adjudication: Crisis and Reform", 72 *Stanford Law Review* (forthcoming).

Andreoni, James, Brian Erard, and Jonathan Feinstein. 1998. "Tax Compliance," 36 *Journal of Economic Literature* 818–60.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables," 91 *Journal of the American statistical Association* 444–55.

Asimow, Michael. 2016. "Inquisitorial Adjudication and Mass Justice in American Administrative Law," in *The Nature of Inquisitorial Processes in Administrative Regimes*, 107–26. Washington, DC: Routledge. https://www.acus.gov/research-projects/federal-administrative-adjudication-outside-administrative-procedure-act

———. 2018. *Sourcebook on Federal Administrative Adjudication Outside the APA*. Administrative Conference of the United States.

Baker, George P. 1992. "Incentive Contracts and Performance Measurement," 100 *Journal of Political Economy* 598–614.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2012. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training." Technical Report, National Bureau of Economic Research.

Barnow, Burt S. 2000. "Exploring the Relationship between Performance Management and Program Impact: A Case Study of the Job Training Partnership Act," 19 *Journal of Policy Analysis and Management* 118–41.

Barry, Daniel, and John A. Hartigan. 1993. "A Bayesian Analysis for Change Point Problems," 88 *Journal of the American Statistical Association* 309–19.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," 57 *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Bevan, Gwyn, and Christopher Hood. 2006. "What's Measured Is What Matters: Targets and Gaming in the English Public Health Care System," 84 *Public Administration* 517–38.

Blanes i Vidal, Jordi, and Clare Leaver. 2015. "Bias in Open Peer-Review: Evidence from the English Superior Courts," 31 *Journal of Law, Economics, and Organization* 431–71.

Board of Veterans' Appeals. 2002. "Report of the Chairman." Department of Veterans Affairs.

Board of Veterans' Appeals. 2014. "Annual Report." Department of Veterans Affairs.

Board of Veterans' Appeals. 2015. "Congressional Budget Submission." Department of Veterans Affairs.

Board of Veterans' Appeals. 2016. "Annual Report." Department of Veterans Affairs.

Boyd, Christina L., and Amanda Driscoll. 2013. "Adjudicatory Oversight and Judicial Decision Making in Executive Branch Agencies," 41 *American Politics Research* 569–98.

Braithwaite, John, and Valerie Braithwaite. 1995. "The Politics of Legalism: Rules versus Standards in Nursing-Home Regulation," 4 *Social & Legal Studies* 307–41.

Brennan, Troyen A. 1998. "The Role of Regulation in Quality Improvement," 76 *The Milbank Quarterly* 709–31.

Breyer, Stephen G., Richard B. Stewart, Cass R. Sunstein, Adrian Vermeule, and Michael E. Herz. 2011. *Administrative Law and Regulatory Policy: Problems, Text, and Cases*. New York, NY: Wolters Kluwer.

Brignall, Stan, and Sven Modell. 2000. "An Institutional Perspective on Performance Measurement and Management in the 'New Public Sector'," 11 *Management Accounting Research* 281–306.

Brodkin, Evelyn, and Michael Lipsky. 1983. "Quality Control in AFDC as an Administrative Strategy," 57 *Social Service Review* 1–34.

Brodkin, Evelyn Z. 2006. "Bureaucracy Redux: Management Reformism and the Welfare State," 17 *Journal of Public Administration Research and Theory* 1–17.

Cable, G. 2001. "Enhancing Causal Interpretations of Quality Improvement Interventions," 10 *BMJ Quality & Safety* 179–86.

Chassman, Deborah A., and Howard Rolston. 1979. "Social Security Disability Hearings: A Case Study in Quality Assurance and Due Process," 65 *Cornell Law Review* 801–22.

Congressional Research Service. 2012. "Disability Benefits Available under the Social Security Disability Insurance (SSDI) and Veterans Disability Compensation (VDC) Programs." CRS Report for Congress.

Cragin, Charles L. 1994. "The Impact of Judicial Review on the Department of Veterans Affairs' Claims Adjudication Process: The Changing Role of the Board of Veterans' Appeals," 46 *Maine Law Review* 23–41.

Cuéllar, Mariano-Florentino. 2006. "Auditing Executive Discretion," 82 *Notre Dame Law Review* 227–312.

Department of Veterans Affairs. 2018. "Comprehensive Plan for Processing Legacy Appeals and Implementing the Modernized Appeals System Public Law 115-55, Section 3 (Feb. Update)," *Department of Veterans Affairs*.

Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review," 37 *Journal of Human Resources* 696–727.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2013. "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India," 128 *The Quarterly Journal of Economics* 1499–1545.

Erdman, Chandra, and John W. Emerson. 2007. "BCP: An R Package for Performing a Bayesian Analysis of Change Point Problems," 23 *Journal of Statistical Software* 1–13.

Freedman, David A. 2008. "Randomization Does Not Justify Logistic Regression," 23 *Statistical Science* 237–49.

Gelbach, Jonah B., and David Marcus. 2016. "A Study of Social Security Disability Litigation in the Federal Courts." Report for the Administrative Conference of the United States.

———. 2018. "Rethinking Judicial Review of High Volume Agency Adjudication," 96 *Texas Law Review* 1097–162.

Gilmour, John B., and David E. Lewis. 2006. "Does Performance Budgeting Work? An Examination of the Office of Management and Budget's PART Scores," 66 *Public Administration Review* 742–52.

Greiner, D. James, and Andrea Matthews. 2016. "Randomized Control Trials in the United States Legal Profession," 12 *Annual Review of Law and Social Science* 295–312.

Hartman, Erin, and F. Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests," 62 *American Journal of Political Science* 1000–13.

Hausman, David. 2016. "The Failure of Immigration Appeals," 164 *University of Pennsylvania Law Review* 1177–238.

Hennings, Bradley W., David E. Boelzner, and Jennifer Rickman White. 2016. "Now Is the Time: Experts vs. the Uninitiated as Future Nominees to the U.S. Court of Appeals for Veterans Claims," 25 *Federal Circuit Bar Journal* 371–400.

Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin, and Xiao-Hua Zhou. 2000. "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," 1 *Biostatistics* 69–88.

Ho, Daniel E. 2017. "Does Peer Review Work: An Experiment of Experimentalism," 69 *Stanford Law Review* 1–119.

Ho, Daniel E., and Sam Sherman. 2017. "Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement," 13 *Annual Review of Law and Social Science* 251–72.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," 7 *Journal of Law, Economics, and Organization* 24–52.

House Committee on Veterans' Affairs. 2007a. "House Hearing on the Board of Veterans' Appeals Adjudication Process and the Appeals Management Center." 110th Congress, 1st session (September 25, 2007).

House Committee on Veterans' Affairs. 2007b. "House Hearing on the Challenges Facing the U.S. Court of Appeals for Veterans Claims." 110th Congress, 1st session (May 22, 2007).

House Committee on Veterans' Affairs. 2008. "Examining the Effectiveness of the Veterans Benefits Administration's Training, Performance Management and Accountability." 110th Congress, 2nd session (September 18, 2008).

Koch Charles H. Jr., and David A. Koplow. 1990. "The Fourth Bite at the Apple: A Study of the Operation and Utility of the Social Security Administration's Appeals Council," 17 *Florida State University Law Review* 199–324.

Krent, Harold J., and Scott Morris. 2013. "Achieving Greater Consistency in Social Security Disability Adjudication: An Empirical Study and Suggested Reforms." Administrative Conference of the United States.

Lubbers, Jeffrey S. 1993. "The Federal Administrative Judiciary: Establishing an Appropriate System of Performance Evaluation for ALJ's," 7 *Administrate Law Journal* 589–628.

Margetts, Helen Z. 2011. "Experiments for Public Management Research," 13 *Public Management Review* 189–208.

Mashaw, Jerry L. 1973a. "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," 59 *Cornell Law Review* 772–824.

———. 1973b. "Report in Support of Recommendation 73-3: Quality Assurance Systems in the Adjudication of Claims Entitlement to Benefits or Compensation." Administrative Conference of the United States.

———. 1980. "How Much of What Quality? A Comment on Conscientious Procedural Design," 65 *Cornell Law Review* 823–35.

———. 1983. *Bureaucratic Justice: Managing Social Security Disability Claims*. New Haven, CT: Yale University Press.

———. 1985. *Due Process in the Administrative State*. New Haven, CT: Yale University Press.

———. 1996. "Reinventing Government and Regulatory Reform: Studies in the Neglect and Abuse of Administrative Law," 57 *University of Pittsburgh Law Review* 405–41.

Mashaw, Jerry L., Charles J. Goetz, Frank I. Goodman, Warren F. Schwartz, Paul R. Verkuil, and Milton M. Carrow. 1978. *Social Security Hearings and Appeals: A Study of the Social Security Administration Hearing System*. Lexington, MA: Lexington Books.

McCubbins, Mathew D., and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms," 28 *American Journal of Political Science* 165–79.

Merrill, Thomas W. 2017. "Jerry L. Mashaw, the Due Process Revolution, and the Limits of Judicial Power," in Nicholas R. Parrillo , ed., Administrative Law from the Inside Out, 39–62, Chapter 1. Cambridge: Cambridge University Press.

Metzger, Gillian B. 2014. "The Constitutional Duty to Supervise," 124 *Yale Law Journal* 1836–933.

Metzger, Gillian E., and Kevin M. Stack. 2016. "Internal Administrative Law," 115 *Michigan Law Review* 1239–308.

Miles, Thomas J., and Cass R. Sunstein. 2006. "Do Judges Make Regulatory Policy? An Empirical Investigation of Chevron," 73 *University of Chicago Law Review* 823–81.

Noonan, Kathleen G., Charles F. Sabel, and William H. Simon. 2009. "Legal Accountability in the Service-Based Welfare State: Lessons from Child Welfare Reform," 34 *Law & Social Inquiry* 523–68.

Osborne, David, and Ted Gaebler. 1993. *Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector*. New York, NY: Plume.

Parrillo, Nicholas R. 2017. "Jerry L. Mashaw's Creative Tension with the Field of Administrative Law," in Nicholas R. Parrillo, ed., Administrative Law from the Inside Out, 1–35. Cambridge: Cambridge University Press.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz and Philip G. Schrag. 2007. "Refugee Roulette: Disparities in Asylum Adjudication," 60 *Stanford Law Review* 295–412.

Ray, Gerald K., and Jeffrey S. Lubbers. 2015. "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," 83 *George Washington Law Review* 1575–608.

Ridgway, James D. 2009. "Why so Many Remands: A Comparative Analysis of Appellate Review by the United States Court of Appeals for Veterans Claims," 1 *Veterans Law Review* 113–66.

———. 2010. "The Veterans' Judicial Review Act Twenty Years Later: Confronting the New Complexities of the Veterans Benefits System," 66 *NYU Annual Survey of American Law* 251–98.

Ridgway, James D., and David S. Ames. 2018. "Misunderstanding Chenery and the Problem of Reasons-or-Bases Review," 68 *Syracuse Law Review* 303–45.

Ridgway, James D., Barton F. Stichman, and Rory E. Riley. 2016. "Not Reasonably Debatable: The Problems with Single-Judge Decisions by the Court of Appeals for Veterans Claims," 27 *Stanford Law and Policy Review* 1–56.

Sabel, Charles F., and William Simon. 2017. "The Management Side of Due Process in the Service-Based Welfare State," in R. Parrillo Nicholas , ed., Administrative Law from the Inside Out, Chapter 2, 63–86. Cambridge: Cambridge University Press.

Schuck, Peter H., and E. Donald Elliott. 1990. "To the Chevron Station: An Empirical Study of Federal Administrative Law," 1990 *Duke Law Journal* 984–1077.

Senate Committee on Veterans' Affairs. 2005. "Senate Hearing on Battling the Backlog: Challenges Facing the VA Claims Adjudication and Appeal Process." 109th Congress, 1st session (May 26, 2005).

Shavell, Steven. 1995. "The Appeals Process as a Means of Error Correction," 24 *The Journal of Legal Studies* 379–426.

Simon, William H. 1983. "Legality, Bureaucracy, and Class in the Welfare System," 92 *The Yale Law Journal* 1198–269.

———. 2006. "Toyota Jurisprudence: Legal Theory and Rolling Rule Regimes," in G. de Búrca and J. Scott, eds., Law and New Governance in the EU and the US, 37–64. Oxford: Hart Publishing.

———. 2012. "Where Is the Quality Movement in Law Practice," 2012 *Wisconsin Law Review* 387–406.

———. 2015. "The Organizational Premises of Administrative Law," 78 *Law and Contemporary Problems* 61.

Social Security Administration. 2018. *Appeals to Court as a Percentage of Appealable AC Dispositions*. Available at: https://www.ssa.gov/appeals/DataSets/AC04_NCC_Filed_Appealable.html.

Stammann, Amrei, Florian Heiß, and Daniel McFadden. 2016. "Estimating Fixed Effects Logit Models with Large Panel Data," in *Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel—Session: Microeconometrics, No. G01-V3*. ZBW—Deutsche Zentralbibliothek für Wirtschaftswissenschaften. https://www.econstor.eu/bitstream/10419/145837/1/VfS_2016_pid_6909.pdf

Tirole, Jean. 1986. "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations," 2 *Journal of Law, Economics, and Organization* 181–214.

US General Accounting Office. 1978. "Administrative Law Process: Better Management Is Needed." Report to the Congress of the United States.

US General Accounting Office. 2002. "Quality Assurance for Disability Claims and Appeals Processing Can Be Further Improved." Report to the Ranking Democratic Member, Committee on Veterans' Affairs, House of Representatives.

US Government Accountability Office. 2005. "Board of Veterans' Appeals Has Made Improvements in Quality Assurance, but Challenges Remain for VA in Assuring Consistency." Testimony before the Subcommittee on Disability Assistance and Memorial Affairs, Committee on Veterans' Affairs, House of Representatives.

VA Office of Inspector General. 2018. "Veterans Benefits Administration: Review of Timeliness of the Appeals Process." Report No. 16-01750-79.

Verkuil, Paul. 2017. "Meeting the Mashaw Test for Consistency in Administrative Decision-Making," in Nicholas R. Parrillo , ed., Administrative Law from the Inside Out, Chapter 9, 239–46. Cambridge: Cambridge University Press.

Verkuil, Paul R. 1991. "Reflections upon the Federal Administrative Judiciary," 39 *UCLA Law Review* 1341–63.

Wilson, James. 1991. *Bureaucracy: What Government Agencies Do and Why They Do It*. Basic Books.