

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

Daniel E. Ho[†] Cassandra Handan-Nader[‡] David Ames[§] David Marcus[¶]

August 2, 2018

Abstract

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” by a team of full-time staff attorneys for nearly 15 years. This performance program had the express goals of measuring accuracy and reducing reversal rates on appeal. In cases of legal error, the quality review team wrote memoranda to judges to permit correction before opinions were issued. We use rich internal administrative data on nearly 600,000 cases from 2002-2016 to provide the first rigorous study of this review process. With precise estimates, we show that the program had no appreciable effect on reducing appeals or reversals. Based on internal records, we demonstrate that this inefficacy is likely by design, as meeting the performance measure of “accuracy” was at cross-purposes with error correction. These findings inform longstanding questions of law, organization, and bureaucracy, including performance management, standards of review, and the institutional design of mass adjudication.

*This manuscript is authors’ original version. A revised version will be published in the *Journal of Law, Economics, and Organization*. We thank Reid Whitaker for research assistance and Daryl Levinson, James Ridgway, and Bill Simon for helpful comments and conversations.

[†]Corresponding Author, William Benjamin Scott and Luna M. Scott Professor of Law, Professor (by courtesy) of Political Science; & Senior Fellow at Stanford Institute for Economic Policy Research; Stanford University, 559 Nathan Abbott Way, Stanford, CA 94305; Tel: 650-723-9560; Fax: 650-725-0253; Email: dho@law.stanford.edu

[‡]Research Fellow, Stanford Law School, 559 Nathan Abbott Way, Stanford, CA 94305; Ph.D. student, Department of Political Science, Stanford University

[§]Director, Appellate Operations at Bergmann and Moore, LLC; Former-Chief of Office of Quality Assurance, Board of Veterans’ Appeals, U.S. Department of Veterans Affairs; 7920 Norfolk Ave #700, Bethesda, MD 20814

[¶]Professor of Law, UCLA School of Law, 385 Charles E. Young Drive East, Los Angeles, CA 90095

1 Introduction

A hallmark feature of the American administrative state is mass adjudication. Each year, roughly 1,300 administrative law judges in the Social Security Administration (SSA) resolve over 600,000 disability and social security appeals, more than the caseload of all U.S. federal district courts combined. In the Executive Office for Immigration Review, roughly 250 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 1,000 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., Asimow, 2016; Braithwaite and Braithwaite, 1995; Ho, 2017; Mashaw et al., 1978; Mashaw, 1985a; Noonan et al., 2009; Ramji-Nogales et al., 2007). In landmark studies, Mashaw et al. (1978) and Mashaw (1985a) documented dramatic disparities in how SSA administrative law judges adjudicated comparable cases. Mashaw argued these failures amounted to a constitutional due process problem, requiring an internal *management* systems for quality assurance and performance management (Mashaw, 1973, 1985a). Mashaw pointed to the VA's "statistical quality assurance control" system as one positive example (Mashaw, 1973, pp. 793–96) and famously argued that agencies could and should internally develop such mechanisms for bureaucratic rationality (Mashaw, 1985a). The then-General Accounting Office (GAO), in turn, 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 (Gelbach and Marcus, 2016, 2018; Hausman, 2016; Ho, 2017; Krent and Morris, 2013; Verkuil, 2017, 1991) and much writing about law and management of the bureaucracy (e.g., Brodtkin, 2006; Chassman and Rolston, 1979; Metzger, 2014; Simon, 1983, 2006; Wilson, 1991), there exists little rigorous evi-

dence about the effectiveness of quality management systems in adjudication specifically or the public sector generally (Greiner and Matthews, 2016; Ho and Sherman, 2017; Margetts, 2011). To date, the evidence exists exclusively of useful but limited qualitative case studies (e.g., Brodtkin and Lipsky, 1983; Koch Jr 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 question 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 quality review process by an independent team of full-time attorneys. For decisions remanded by the U.S. Court of Appeals for Veterans Claims (CAVC), which hears appeals from 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 issued. Roughly 75% 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 2003-2016, 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 quality review 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 contains rich covariate information, and we show balance on over 80 dimensions, including

¹We 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.

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 the quality review 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 quality review. Cases that underwent quality review 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 simply ignored memoranda written by the quality review team. Our evidence also does not support the possibility that the results are explained by arbitrariness of CAVC decisions or the quality review team. To the contrary, we show that conditional on quality review, the presence of an error is associated with a higher risk of a remand. This shows that the quality review team was in fact able to identify low-quality *types* of opinions. But even for opinions that the quality review deemed to have *no errors*, the remand rate remained a stunning 74%. As a 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 standard 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 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.” The internal training manual later clarified that an error should only be called when “undebatable.” We demonstrate this divergence empirically by comparing the rate at which the quality review 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 quality review calls the error. We also show that more stringent quality reviewers are more likely to agree with CAVC's disposition in a case.

The likely reason for this functional divergence of standards of review was 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 and scrutinized in congressional oversight hearings (House Committee on Veterans' Affairs, 2007, 2008; Senate Committee on Veterans' Affairs, 2005). 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 quality review program in the administrative state. The internal data, used by BVA to run the quality review program, allows 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 effects of any substantial magnitude. Second, the BVA quality review program exemplifies the kind of program scholars and policymakers have envisioned as curing the due process problems of mass adjudication (Administrative Conference of the United States, 1978; US Government Accountability Office, 2005; Gelbach and Marcus, 2016; Mashaw, 1973). The review process was resource-intensive, involving four to six full-time staff attorneys, with the case load of the team exceeding that of most U.S. 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 rate. 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 quality review 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., Asimow, 2016; Gelbach and Marcus, 2018; Congressional Research Service, 2012; Sabel and Simon, 2017; Verkuil, 2017, 1991), few studies have empirically examined decision making in veterans adjudication.²

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 (Baker, 1992; Barnow, 2000; Bevan and Hood, 2006; Dixit, 2002; Duflo et al., 2013; Holmstrom and Milgrom, 1991). Our findings underscore the difficulty of monitoring bureaucratic and judicial quality, which is central to questions of presidential and congressional oversight of agencies (Boyd and Driscoll, 2013; Cuéllar, 2006; McCubbins and Schwartz, 1984), 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 non-random selection of appeals (see Hausman, 2016; Mashaw, 1980; Simon, 2015; but cf. Gelbach and Marcus, 2018). Our paper shows limitations to the ability to reduce reversal rates even *with* random selection of judicial decisions. 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., Miles and Sunstein, 2006; Schuck and Elliott, 1990). 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). Our evidence is consistent with Blanes i Vidal and Leaver (2015), who find that favoritism bias leads

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

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 quality review – and administrative law – from within an agency.

Our paper proceeds as follows. Section 2 provides institutional background to veterans adjudication and the quality review process. Section 3 describes our unique BVA dataset and demonstrates that we are able to replicate the randomized case selection process for quality review. Section 4 presents results and Section 5 presents limitations. Section 6 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. Claimants can appeal benefits determinations by first filing a “Notice of Disagreement.” The VA may revise the determination in response. Claimants may then further appeal to the Board of Veterans Appeals (BVA). In 2015, the BVA’s annual budget was around \$94 million (Board of Veterans’ Appeals, 2015, p. BVA-1), with most of it allocated for the personnel of roughly 90 Veterans Law Judges (VLJs) and 450 staff attorneys (Board of Veterans’ Appeals, 2016). VLJs are appointed by the President and removable only for cause (31 U.S.C. § 7101A). Staff attorneys are hired and subject to the federal civil service (General Schedule, GS) system. VLJs hear appeals, hold hearings, and issue opinions under a de novo review standard, applied to the full claims file. Claimants file for internal appeal in 11-12% of all VA claims and roughly 4-5% of all VA claims are heard by the BVA. 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. Veterans who received a resolution by the Board in 2017 waited an average of 6 years from filing the Notice of Disagreement (VA Office of Inspector General, 2018).

Until 1988, BVA decisions were final. With the Veterans’ Judicial Review Act of 1988, Congress added another layer of appeal, creating the U.S. Court of Appeals for Veterans Claims (CAVC).

Roughly 6% of 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). 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 U.S. Court of Appeals for the Federal Circuit and then the U.S. Supreme Court, but such appeals are exceedingly rare.

The creation of CAVC and its high remand rate led BVA to develop a more systematic quality review (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.³

While the program was revised in the early years, in part due to criticism by the GAO (US General Accounting Office, 2002), it remained essentially 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 Office of Quality Review, comprised of four to six attorneys, reviewed draft opinions for specific errors. 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, structure). Each QR team member coded these categories along with a more exhaustive subcategory coding.⁴ 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]”

³Richard B. Standefer, Acting Chairman, Memorandum No. 01-98-15 (May 14, 1998) (emphasis added).

⁴As we detail in Appendix C, the subcategories were refined over time.

(Board of Veterans’ Appeals, 2002, p. 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 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).⁵ VLJs are subject to regular performance reviews (38 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 potentially permitted its use in performance reviews.⁶

3 Descriptive Statistics and Balance

Data. We secure data, never before analyzed by outside researchers, on all BVA decisions from October 1, 1999, to January 31, 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,

⁵This performance goal itself changed over time. In its budget request of the 2008 fiscal year, for instance, BVA published a target of 92%.

⁶See 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.”).

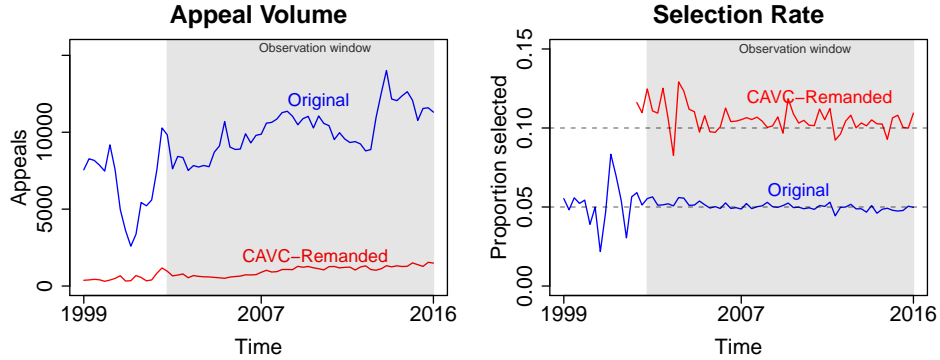


Figure 1: Quarterly eligible appeal volumes (left) QR selection rates (right) over all available time periods, with the observation window for the study shaded in gray (August 1, 2003 - November 9, 2016). Appeals are split by whether they had been remanded by CAVC, resulting in the decision of interest. Within the observation window, we are able to replicate the 5% and 10% selection rates for original and CAVC-remanded cases, respectively.

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 quality review.

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 secure 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 Veterans Benefits Administration (VBA) upon remand by the Board. Third, we exclude dismissals due to the death of the appellant. After these exclusions, we are left with 785,812 QR-eligible decisions, comprising about 29% of all VACOLS appeals.

The left panel of Figure 1 plots the volume of original appeals and CAVC-remanded appeals over time. As caseload has been increasing over time, the number of original appeals 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 1,500 remands per quarter in 2016.

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 an appeal 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 vast majority of cases by an administrator after signature but before the dispatch of the decision.⁷

The right panel of Figure 1 displays the selection rate over time, with the blue line plotting the time series for original decisions and the red 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 1999-2003. (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 2003-2016, we are able to cleanly replicate the Board's publicly stated sampling rate. The sampling variability for CAVC-remanded appeals is higher, as these constitute less than one fifth of QR-eligible decisions. For 508,801 original appeals,

⁷To 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).

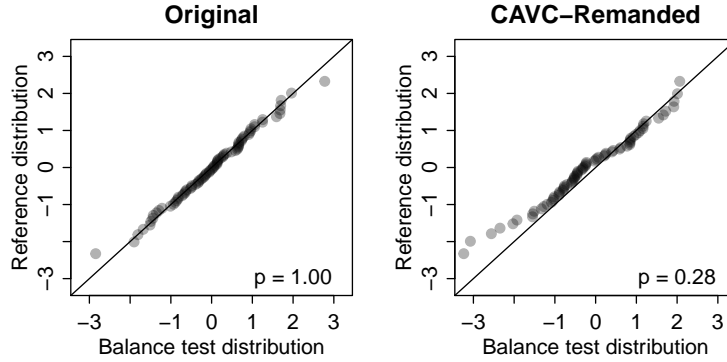


Figure 2: Quantile-quantile plot of t -statistics from 80 balance tests between appeals selected and not selected for QR against reference $t_{n_1+n_2-2}$ -distribution. Cases are split by original and CAVC-remanded cases. p -values are from a Kolmogorov-Smirnov test of distribution equality between the observed t -statistics and reference distribution.

we calculate a 5.01% selection rate, and for 47,981 CAVC-remanded appeals, 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.

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 to 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 2 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 appeals. 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.

	Original Cases				CAVC-Remanded Cases			
	Ctrl.	QR	Diff.	<i>p</i> -val.	Ctrl.	QR	Diff.	<i>p</i> -val.
Appellant age at notice of disagreement (years)	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.)								
WWII (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 (no. of issues per appeal)								
Service connection								
All others	1.45	1.45	-0.00	0.99	0.99	1.02	0.03	0.60
New and material	0.07	0.08	0.00	0.80	0.04	0.04	0.00	0.74
Accrued	0.02	0.02	-0.00	0.91	0.02	0.03	0.01	0.40
Increased rating								
Schedular	0.65	0.64	-0.01	0.91	0.54	0.52	-0.01	0.74
Schedular & 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
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
Issue diagnosis categories (no. 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
Appeal Representation								
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
Appeal Process								
Length of appeal (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 SSOCs submitted (prop. of appeals)								
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				
Number of BVA Appeals (sample size)	508,801	26,821			47,981	5,622		

Table 1: Balance on selected covariates between appeals not selected for quality review (Ctrl) and appeals randomly selected for quality review (QR) between August 1, 2003 and November 9, 2016. Appeals 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. Length of appeal is measured between notice of disagreement and final appeal decision. 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.

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 gives us confidence that we have replicated the QR selection process. Random selection with such a large sample ensures balance across and QR and control, enabling us to assess the impact of QR on case outcomes.

4 Results

4.1 Causal Effect of Quality Review

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. The relatively large sample sizes allows us to rule out effect sizes of appreciable magnitude: the 95% confidence interval (CI) is [-0.62%, 0.16%]. This first finding 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,⁸ 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 and 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

⁸To be clear, BVA retains a separate internal database for quality review at the issue level, but this QR data is 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).

	Original Cases				CAVC-Remanded Cases			
	Ctrl.	QR	Diff.	<i>p</i> -val.	Ctrl.	QR	Diff.	<i>p</i> -val.
<i>Of all BVA Cases...</i>								
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	5,622		
<i>Conditional on CAVC appeal...</i>								
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	1,604			6,782	728		

Table 2: Means and differences-in-means (Diff.) for outcomes, comparing control (Ctrl.) cases not selected for quality review and treatment cases randomly selected for quality review (QR) between August 1, 2003 and November 9, 2016. 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.

to 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, which forms one outcome of interest.

The appeals rate for CAVC-remanded appeals was 14% for non-QR appeals and 13% for QR appeals. This 1% difference is not statistically significant, although the 95% confidence interval 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 to 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.

<i>Outcome</i>		Original Cases			CAVC-Remanded Cases		
Appealed to CAVC	QR effect	-0.040 (0.026)	-0.040 (0.026)	-0.040 (0.026)	-0.101* (0.042)	-0.100* (0.042)	-0.098* (0.042)
	VLJ FEs	N	N	Y	N	N	Y
	Year-quarter FEs	N	Y	Y	N	Y	Y
	<i>N</i>	535,622	535,622	535,622	53,603	53,603	53,603
Reversed/ remanded by CAVC	QR effect	-0.033 (0.060)	-0.045 (0.060)	-0.039 (0.061)	-0.113 (0.086)	-0.098 (0.087)	-0.117 (0.090)
	VLJ FEs	N	N	Y	N	N	Y
	Year-quarter FEs	N	Y	Y	N	Y	Y
	<i>N</i>	33,194	33,194	33,194	7,510	7,510	7,510

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 column). The QR effect row presents the coefficient on the treatment indicator, with standard errors in parentheses. FEs indicate fixed effects, which are not displayed for readability; N indicates sample size. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 3 presents logistic (fixed effects) regression results to adjust for differences over time and by VLJ hearing the case.⁹ 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 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 CAVC disposition (conditional on appeal) is again statistically insignificant, we observe statistically significant estimates of the QR program on the probability of appeals, corresponding to a 1% reduction in the appeals rate. The magnitude, however, remains small. With 5,622 CAVC-remanded cases undergoing QR, the best estimate is that the QR process avoids roughly 60 appeals. To put that in context, the Board received over 90,000 cases in 2017 alone. Staffing a QR office with four to six full-time attorneys to avoid 60 appeals, when a single VLJ has 1,000 cases docketed annually, illustrates that this effect is substantively not very meaningful.¹⁰

In order to isolate the effect of the memoranda drafted by the QR team, we also estimate a series

⁹We note that such adjustment is not uncontested (see, e.g., Freedman, 2008).

¹⁰Appendix 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.

<i>Outcome</i>		Original Cases	CAVC-Remanded Cases
Appealed to CAVC	Memorandum effect	-0.032 (0.021)	-0.210* (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 (0.082)	-0.257 (0.197)
	First stage R^2	0.13	0.08
	Second stage R^2	0.00	0.00
	N	33,194	7,510

Table 4: Instrumental variable linear models for the effect of an “exception memorandum” on outcomes, using random selection for QR as an instrument. 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 standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

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. There is hence little evidence of a stronger effect in the subset of cases receiving memoranda.¹¹

We now examine whether the effects of quality review 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 of 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

¹¹Sometimes this “complier average causal effect” is also referred to as the “intention to treat” effect for the subgroup of compliers (Hirano et al., 2000).

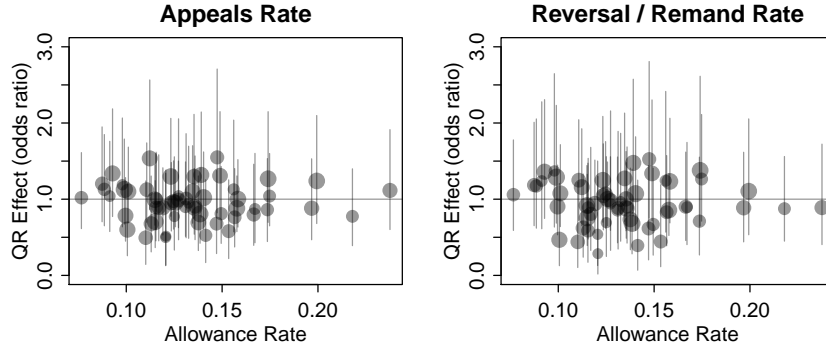


Figure 3: Treatment effects by VLJ with 95% confidence intervals for the odds ratio of an appeal to the CAVC (left) and a remand/reversal by the CAVC (right) for original cases. Only VLJs with at least 4000 control cases are shown.

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 4,000 control opinions available. While this might seem like a high threshold, the expected number of opinions that would both go through QR and be appealed to CAVC under the null would be less than 12 ($= 4000 \text{ cases} \times 5\% \text{ QR selection rate} \times 6\% \text{ appeal rate}$). 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.¹² We omit results on CAVC-remanded decisions, as VLJ-specific effects are too imprecise.

Figure 3 plots VLJ stringency against the treatment effect. Each dot represents odds ratio for an outcome for each VLJ, weighted by QR sample size, with 95% confidence intervals. If the QR

¹²We calculate the average allowance rate across all issues for each VLJ.

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 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 appeals. 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 sheer randomness of case outcomes at CAVC.

	Original Appeals				CAVC-Remanded Appeals			
	No Error	Error	Diff.	<i>p</i> -val.	No Error	Error	Diff.	<i>p</i> -val.
<i>Of all BVA QR Cases...</i>								
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	1,926			5,305	317		
<i>Conditional on CAVC appeal and QR...</i>								
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)	1,392	212			664	64		

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. 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 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.

Another potential explanation is that VLJs might simply ignore 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 draft opinions, 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 Assurance, 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

Error Type	All QR Cases	QR Cases Appealed to CAVC	
	QR call rate	QR call rate	CAVC remand rate
Reasons or bases	0.048	0.104	0.624***
Due process	0.028	0.039	0.102***
Other issues	0.003	0.005	
Findings of fact	0.001	0.003	
Conclusions of law	0.001	0.001	
N	26,821	1,604	

Table 6: Rates at which QR team calls errors by error type for all (a) QR cases (left) or (b) QR cases appealed to the CAVC (middle), and CAVC remand rates for cases that underwent QR and were appealed to CAVC by available error types (right). Only original cases that were selected for QR are included. Stars report statistical significance tests on the difference between the QR call rate and the CAVC remand rate for the same error code for the same sample of cases. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

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 show that the QR process in fact gave significantly more deference to VLJ determinations. In a Chairman’s Memorandum, the QR was 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, they were nowhere close to the stringency required to withstand scrutiny on appeal. Most stunningly, Table 5 shows that for cases given a clean bill of health (i.e., with no errors identified) by the QR team, CAVC remanded 74% of the time when appealed.

To corroborate this explanation, the first column of Table 6 reports 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 second column reports error call rates for the sample of QR cases that were also appealed to CAVC. The 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.”¹³ The right column calculates 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.¹⁴ 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 4 plots reviewer stringency on the x -axis against the agreement rate on the y -axis. Each dot represents one 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.

It is worth noting that appeals selection does not explain the discrepancy between BVA’s claim

¹³Appendix 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.

¹⁴For extremely complicated cases, the chief of the QR office would ensure balance of workloads across reviewers.

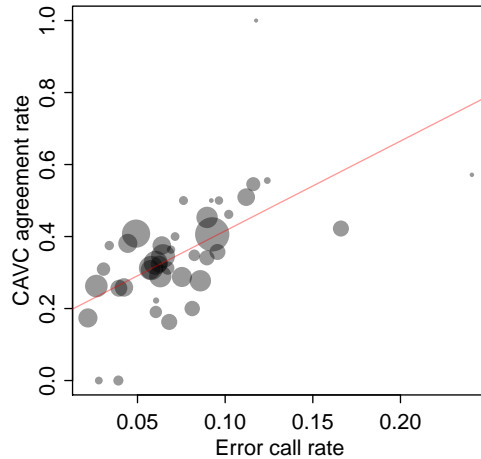


Figure 4: Correlation between stringency of each QR reviewer (i.e., the error call rate) on the x -axis and the agreement with CAVC on the y -axis. Reviewers with higher error call rates are more likely to agree with CAVC, conditional on a case being appealed.

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 to a baseline of 6.9% across all cases. In other words, even with a lenient standard of review, unappealed cases appear to have significant errors.

What explains BVA’s weakening of the internal standard of review? Recall that the QR program had dual purposes, namely (1) to reduce errors, and (2) to report a performance measure pursuant to the Government Performance and Results Act (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 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 absolute number of cases than would be mechanically 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.”¹⁵ Discounting the one error that is the predominant reason for CAVC remands is effectively an admission of decoupling BVA’s standard of review from CAVC’s.

5 Limitations

We now discuss limitations of our study. First, while our study has, for the first time, identified the causal effect of QR review on case outcomes, our research design does not provide leverage over the impact of the QR program as a whole. It is theoretically possible that the QR program had no effect on individual cases, but affected quality overall. For instance, if the QR team identified systematic errors and developed a training program for all VLJs, the benefit might manifest itself in all cases. There are reasons to doubt this account, however. The accuracy rate not in fact been uniformly increasing over time. And the fact that memoranda written specifically for draft decisions did not have an impact does not bode well for the effectiveness of training derived from those memoranda. 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 on more systematic, rather than individualized, detection of errors.

Second, 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 BVA shows that the standard of review matters for a subset of cases.

Third, while many scholars have viewed immigration courts, social security adjudication, and

¹⁵Steven L. Keller, Vice Chairman, Memorandum on Monthly Performance Review Submission on the Board of Veterans’ Appeals’ Accuracy Rate (Aug. 3, 2010).

the BVA as close institutional cousins (Asimow, 2016; Gelbach and Marcus, 2018; Congressional Research Service, 2012; Sabel and Simon, 2017; Verkuil, 2017, 1991), 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, 2014). The peer review program for immigration courts appeared to be much more of a training program. We believe the political and institutional tension in 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, p. 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, 1973). 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 suggest that the prospective reform is unlikely to address the longstanding quality problems in BVA adjudication.

6 Conclusion

Our study is the first to leverage randomization of quality review 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 highlight the tenuous 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 (Ridgway, 2010). BVA’s internal rejection of CAVC’s demand for reasoned explanation – notwith-

standing a 75% remand rate when no errors are called – illustrates the continuing internal conflict around these models. Procedure, after all, adds delay. As Justice Black noted in his dissent in *Goldberg v. Kelly*, procedures meant to protect claimants from termination of benefits can just as well delay eligibility.¹⁶ It now takes an average of 6 years from filing a notice of disagreement to Board resolution (VA Office of Inspector General, 2018). The VA’s Inspector General estimated that 7% of appeals were counted as “resolved” because the veteran died while waiting for a decision (id.). These facts raise basic questions of whether judicialization of mass adjudication in fact serves the veteran population (Mashaw, 1985b).

Second, if judicial review won’t solve 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 quality review challenges more optimistic accounts of bureaucratic rationality Mashaw (1985a) 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 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 in the institutional design of quality review. In critiquing the pre-2002 arrangement, where VLJs were simultaneously conducting QR and deciding appeals, GAO rightly called 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 Government Accountability Office, 2005). Our findings highlight the difficulty and importance of institutional independence of quality review.

Third, our results suggest that case-specific quality review cannot remedy structural challenges stemming from the volume of cases. Errors stemming from caseload cannot easily be addressed by

¹⁶397 U.S. 254, 279 (1970) (Black, J., dissenting) (noting that the difficulty of terminating benefits may induce more screening ex ante, so that “many will never get on the rolls, or at least that they will remain destitute during the lengthy proceedings followed to determine initial eligibility.”).

adding to caseload. BVA's efforts in 2016 to reform the QR program to focus less on case-specific review, but 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 David and Gaebler (1993) that inspired the GPRA. The 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 request starting in 2010. In 2017, one hundred VLJs and staff 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.

References

- Administrative Conference of the United States. 1978. "Recommendation 78-2: Procedures for Determining Social Security Disability Claims." *Federal Register* .
- Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91(434):444–455.
- Asimow, Michael. 2016. "Inquisitorial Adjudication and Mass Justice in American Administrative Law." In *The Nature of Inquisitorial Processes in Administrative Regimes*, Routledge, pp. 107–126.
- Baker, George P. 1992. "Incentive contracts and performance measurement." *Journal of political Economy* 100(3):598–614.
- Banerjee, Abhijit, Raghavendra 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." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 19(1):118–141.
- Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the royal statistical society. Series B (Methodological)* pp. 289–300.
- Bevan, Gwyn, and Christopher Hood. 2006. "What's measured is what matters: targets and gaming in the English public health care system." *Public administration* 84(3):517–538.
- 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." *American Politics Research* 41(4):569–598.
- Braithwaite, John, and Valerie Braithwaite. 1995. "The politics of legalism: Rules versus standards in nursing-home regulation." *Social & Legal Studies* 4(3):307–341.
- Brennan, Troyen A. 1998. "The role of regulation in quality improvement." *The Milbank Quarterly* 76(4):709–731.
- 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*. Wolters Kluwer.
- Brodkin, Evelyn, and Michael Lipsky. 1983. "Quality Control in AFDC as an Administrative Strategy." *Social Service Review* 57(1):1–34.
- Brodkin, Evelyn Z. 2006. "Bureaucracy redux: Management reformism and the welfare state." *Journal of Public Administration Research and Theory* 17(1):1–17.

- Cable, G. 2001. “Enhancing causal interpretations of quality improvement interventions.” *BMJ Quality & Safety* 10(3):179–186.
- Chassman, Deborah A, and Howard Rolston. 1979. “Social Security Disability Hearings: A Case Study in Quality Assurance and Due Process.” *Cornell Law Review* 65:801–22.
- Congressional Research Service. 2012. “Disability Benefits Available Under the Social Security Disability Insurance (SSDI) and Veterans Disability Compensation (VDC) Programs.” *CRS Report for Congress* .
- Cuéllar, Mariano-Florentino. 2006. “Auditing Executive Discretion.” *Notre Dame Law Review* 82:227–312.
- David, Osborne, and Ted Gaebler. 1993. *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector*. Plume.
- Dixit, Avinash. 2002. “Incentives and organizations in the public sector: An interpretative review.” *Journal of human resources* pp. 696–727.
- Duffo, 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.” *The Quarterly Journal of Economics* 128(4):1499–1545.
- Freedman, David A. 2008. “Randomization does not justify logistic regression.” *Statistical Science* pp. 237–249.
- 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* .
- Gelbach, Jonah B, and David Marcus. 2018. “Rethinking Judicial Review of High Volume Agency Adjudication.” *Texas Law Review* 96:1097–1162.
- Greiner, D James, and Andrea Matthews. 2016. “Randomized control trials in the United States legal profession.” *Annual Review of Law and Social Science* 12:295–312.
- Hausman, David. 2016. “The Failure of Immigration Appeals.” *University of Pennsylvania Law Review* 164(5):1177–1238.
- Hirano, Keisuke, Guido W Imbens, Donald B Rubin, and Xiao-Hua Zhou. 2000. “Assessing the effect of an influenza vaccine in an encouragement design.” *Biostatistics* 1(1):69–88.
- Ho, Daniel E. 2017. “Does Peer Review Work: An Experiment of Experimentalism.” *Stanford Law Review* 69:1–119.
- Ho, Daniel E., and Sam Sherman. 2017. “Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement.” *Annual Review of Law and Social Science* 13(1):251–272.
- Holmstrom, Bengt, and Paul Milgrom. 1991. “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.” *Journal of Law, Economics, and Organization* 7:24–52.
- House Committee on Veterans’ Affairs. 2007. “House Hearing on the Board of Veterans’ Appeals Adjudication Process and the Appeals Management Center.” In *110th Congress, 1st session (Sep. 25, 2007)*.

- House Committee on Veterans' Affairs. 2008. "Examining the Effectiveness of the Veterans Benefits Administration's Training, Performance Management and Accountability." In *110th Congress, 2nd session (Sep. 18, 2008)*.
- Koch Jr, Charles H, and David A Koplou. 1990. "The Fourth Bite at the Apple: A Study of the Operation and Utility of the Social Security Administration's Appeals Council." *Florida State University Law Review* 17:199–324.
- 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." *Administrative Law Journal* 7:589–628.
- Margetts, Helen Z. 2011. "Experiments for Public Management Research." *Public Management Review* 13(2):189–208.
- Mashaw, Jerry L. 1973. "Management side of due process: some theoretical and litigation notes on the assurance of accuracy fairness and timeliness in the adjudication of social welfare claims." *Cornell Law Review* 59:772–824.
- Mashaw, Jerry L. 1980. "How Much of What Quality? A Comment on Conscientious Procedural Design." *Cornell L. Rev.* 65:823–35.
- Mashaw, Jerry L. 1985a. *Bureaucratic justice: Managing social security disability claims*. Yale University Press.
- Mashaw, Jerry L. 1985b. *Due process in the administrative state*. Yale university press.
- 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 Books.
- McCubbins, Mathew D., and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28(1):165–179.
- Merrill, Thomas W. 2017. "Jerry L. Mashaw, The Due Process Revolution, and the Limits of Judicial Power." In *Administrative Law from the Inside Out* (Nicholas R. Parrillo, editor), Cambridge: Cambridge University Press, chapter 1, pp. 39–62.
- Metzger, Gillian B. 2014. "The Constitutional Duty to Supervise." *Yale Law Journal* 124:1836–1933.
- Metzger, Gillian E, and Kevin M Stack. 2016. "Internal Administrative Law." *Michigan Law Review* 115:1239–1308.
- Miles, Thomas J, and Cass R Sunstein. 2006. "Do judges make regulatory policy? An empirical investigation of Chevron." *The University of Chicago Law Review* pp. 823–881.
- Noonan, Kathleen G, Charles F Sabel, and William H Simon. 2009. "Legal accountability in the service-based welfare state: Lessons from child welfare reform." *Law & Social Inquiry* 34(3):523–568.
- Parrillo, Nicholas R. 2017. "Jerry L. Mashaw's Creative Tension with the Field of Administrative Law." In *Administrative Law from the Inside Out* (Nicholas R. Parrillo, editor), Cambridge: Cambridge University Press, pp. 1–35.

- Ramji-Nogales, Jaya, Andrew I Schoenholtz, and Philip G Schrag. 2007. “Refugee roulette: Disparities in asylum adjudication.” *Stanford Law Review* 60:295–412.
- Ray, Gerald K, and Jeffrey S Lubbers. 2014. “A government success story: how data analysis by the Social Security Appeals Council (with a push from the Administrative Conference of the United States) is transforming social security disability adjudication.” *George Washington Law Review* 83:1575–1608.
- Ridgway, James D. 2009. “Why So Many Remands: A Comparative Analysis of Appellate Review by the United States Court of Appeals for Veterans Claims.” *Veterans Law Review* 1:113–166.
- Ridgway, James D. 2010. “The Veterans’ Judicial Review Act Twenty Years Later: Confronting the New Complexities of the Veterans Benefits System.” *NYU Annual Survey of American Law* 66:251–298.
- Ridgway, James D, and David S Ames. 2018. “Misunderstanding Chenery and the Problem of Reasons-or-Bases Review.” *Syracuse Law Review* 68:303–45.
- 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.” *Stanford Law and Policy Review* 27:1–56.
- Sabel, Charles F, and William Simon. 2017. “The management side of due process in the service-based welfare state.” In *Administrative Law from the Inside Out* (Nicholas R. Parrillo, editor), Cambridge: Cambridge University Press, chapter 2, pp. 63–86.
- Schuck, Peter H, and E Donald Elliott. 1990. “To the Chevron station: An empirical study of federal administrative law.” *Duke Lj* pp. 984–1077.
- Senate Committee on Veterans’ Affairs. 2005. “Senate Hearing on Battling the Backlog: Challenges Facing the VA Claims Adjudication and Appeal Process.” In *109th Congress, 1st session (May 26, 2005)*.
- Shavell, Steven. 1995. “The Appeals Process as a Means of Error Correction.” *The Journal of Legal Studies* 24(2):379–426.
- Simon, William H. 1983. “Legality, Bureaucracy, and Class in the Welfare System.” *The Yale Law Journal* 92(7):1198–1269.
- Simon, William H. 2006. “Toyota jurisprudence: legal theory and rolling rule regimes.” In *Law and new governance in the EU and the US* (Gráinne de Búrca, and Joanne Scott, editors), Hart Publishing, pp. 37–64.
- Simon, William H. 2012. “Where Is the Quality Movement in Law Practice.” *Wisconsin Law Review* pp. 387–406.
- Simon, William H. 2015. “The organizational premises of administrative law.” *Law and Contemporary Problems* 78:61–.
- Tirole, Jean. 1986. “Hierarchies and bureaucracies: On the role of collusion in organizations.” *Journal of Law, Economics, and Organization* 2:181–214.
- 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* .
- Verkuil, Paul. 2017. “Meeting the Mashaw Test for Consistency in Administrative Decision-Making.” In *Administrative Law from the Inside Out* (Nicholas R. Parrillo, editor), Cambridge: Cambridge University Press, chapter 9, pp. 239–46.
- Verkuil, Paul R. 1991. “Reflections Upon the Federal Administrative Judiciary.” *UCLA Law Review* 39:1341–1363.
- Blanes i Vidal, Jordi, and Clare Leaver. 2015. “Bias in open peer-review: evidence from the English superior courts.” *The Journal of Law, Economics, and Organization* 31(3):431–471.
- Wilson, James. 1991. *Bureaucracy: What Government Agencies Do And Why They Do It*. Basic Books.

A Additional Balance Statistics

Because of space constraints, Table 2 presented balance on only the most salient covariates. Tables 7 presents balance on additional covariates not shown in the main balance table.

	Original Cases				CAVC-Remanded Cases			
	Ctrl.	QR	Diff.	<i>p</i> -val.	Ctrl.	QR	Diff.	<i>p</i> -val.
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
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
Compensation issue types (no. of issues per appeal)								
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				
Issue diagnosis categories (no. of issues per appeal)								
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
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 Appeals (sample size)	508,801	26,821			47,981	5,622		

Table 7: Additional balance checks on selected covariates between appeals not selected for quality review (Ctrl) and appeals randomly selected for quality review (QR) between August 1, 2003 and November 9, 2016. Appeals are split by whether they had been remanded by CAVC leading to the decision at issue. *p*-values are adjusted for multiple testing using Benjamini and Hochberg (1995).

B VLJ-Specific Balance

To assess VLJ-specific treatment effects, we present balance diagnostics at the VLJ level. The left panel 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 red 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.

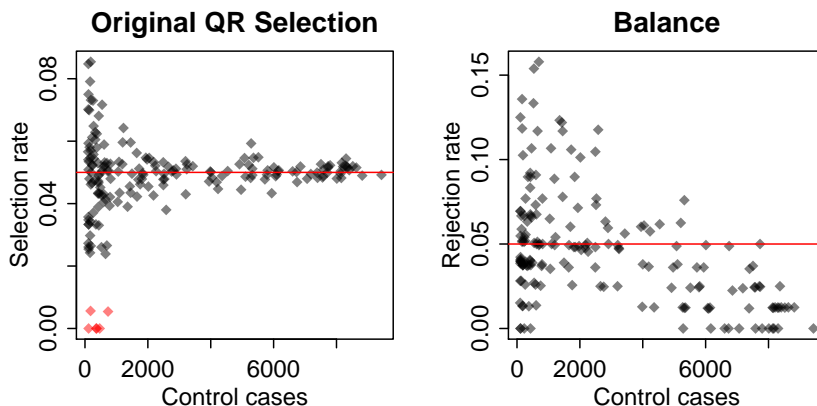


Figure 5: QR selection rate within VLJs for original cases by control case volume (left); balance tests within VLJs by control case volume (right). Each dot represents one VLJ. The rejection rate indicates the fraction of balance tests that reject the null hypothesis, with the red line indicating $\alpha = 0.05$ and sampling variability generating higher rejection rates with VLJs deciding few cases. VLJs represented by the red dots in the left panel were excluded from QR because of their managerial duties, as evidenced by the low selection rate.

The right panel of Figure 5 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 4,000 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. Table 6 presented error rates at the highest level of aggregation (e.g., reasons and bases). BVA used one set of QR error codes from 1999-2015, presented in Table 8, and another set of QR error codes from 2015-2017, presented in Table 9.

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	2,368	364
Total Cases Sampled for Quality Review	24,019	4,868

Table 8: Error codes in the quality review program effective between June 1, 1999 and August 1, 2015 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 Quality Review	2802	754

Table 9: Error codes in the quality review program effective between August 1, 2015 and October 1, 2017 with the frequency counts by case type in the QR-eligible sample..

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 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 Table 6.

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	11940
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 10: CAVC remand reasons pre-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
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 Social Security Administra-	17
tion records	
R&B Failure to adequately address - Duty to notify Incorrect/legally inadequate	15
notice sent	
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	5
records	
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	2
obtain Federal Records	
R&B Failure to adequately address - Duty to notify No notice of inability to	2
obtain Non-Federal Records	
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 Social Security Administration records	65
Duty to assist JSRRC	62
Duty to Notify At Hearing (Bryant)	59
Duty to Notify Incorrect/legally inadequate notice sent	28
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

Table 11: CAVC remand reasons post-July 2013, divided by error type to align with the QR process.

E Robustness

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 12. 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.

<i>Outcome</i>		Original Cases			CAVC-Remanded Cases		
Appealed to CAVC	QR effect	-0.023 (0.028)	-0.021 (0.028)	-0.021 (0.028)	-0.059 (0.051)	-0.055 (0.051)	-0.054 (0.052)
	VLJ FEs	N	N	Y	N	N	Y
	Year-quarter FEs	N	Y	Y	N	Y	Y
	<i>N</i>	255,334	255,334	255,334	18,094	18,094	18,094
Reversed/ remanded by CAVC	QR effect	-0.034 (0.062)	-0.044 (0.062)	-0.036 (0.063)	-0.115 (0.089)	-0.096 (0.090)	-0.115 (0.093)
	VLJ FEs	N	N	Y	N	N	Y
	Year-quarter FEs	N	Y	Y	N	Y	Y
	<i>N</i>	31,638	31,638	31,638	7,052	7,052	7,052

Table 12: Robustness check including only cases with at least one issue denied. 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 standard errors in parentheses. FEs indicate fixed effects, which are not displayed for readability; *N* indicates sample size. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.