DRAFT REPORT FOR THE
ADMINISTRATIVE CONFERENCE OF THE UNITED STATES

QUALITY ASSURANCE SYSTEMS IN AGENCY ADJUDICATION: EMERGING PRACTICES AND INSIGHTS

Daniel E. Ho
Stanford Law School

David Marcus
University of California, Los Angeles

&

Gerald K. Ray
Social Security Administration (retired)*

This report was prepared for the consideration of the Administrative Conference of the United States. It does not necessarily reflect the views of the Conference (including its Council, committees, or members). We thank Austin Peters for terrific research assistance, interviewees at the Board of Veterans Appeals, the Merit Systems Protection Board, the Social Security Administration, the National Labor Relations Board, the United States Citizenship and Immigration Services, the Equal Employment Opportunity Commission, the National Organization of Social Security Claimant's Representatives, and the Patent and Trademark Office for thoughtful conversations, members of the ACUS Adjudication Council for thoughtful feedback, and Danielle Schulkin, Matt Wiener, Jeremy Graboyes, and Matthew Gluth for valuable feedback and assistance in coordinating this Report. Affiliations are for identification only. Names are in alphabetical order and the authors contributed equally to this Report.

* *Gerald Ray was previously an Administrative Appeals Judge and Deputy Executive Director of the Office of Appellate Operations at the Social Security Administration. He is no longer affiliated with the Social Security Administration.*

Recommended Citation

Daniel E. Ho, David Marcus, & Gerald K. Ray, Quality Assurance Systems in Agency Adjudication (Nov. 15, 2021) (draft report to the Admin. Conf. of the U.S.).

CONTENTS

I. Introduction 2

II. Motivation 4

A. Due Process and Decisional Accuracy 4

B. Challenges for Decisional Quality 6

III. Institutional Design 9

A. Quality Assurance Personnel 9

1. Who Participates? 9

2. Length of Assignments 11

B. Whose Work Gets Reviewed? 11

C. Selecting Cases for Quality Review 12

D. Determining Quality 15

1. Standard of Review 15

2. Decisional Inconsistency 17

3. Appellate Outcomes as a Measure 18

E. Timing 18

F. Feedback Mechanisms 19

IV. Emerging Tools for Quality Assurance 22

A. Data Infrastructure 22

B. Data-Driven Quality Insights 24

C. Collaborative Learning and Peer Review 25

D. Artificial Intelligence 26

V. Conclusion 28

Appendix: Agency Interviews 29

I. Introduction

How can agencies monitor, assess, and improve the quality of adjudicatory decisions? This Report surveys and distills emerging lessons learned from work that followed ACUS Recommendation 1973-3[1] and the accompanying report[2] on quality assurance. The 1973

[1] Admin. Conf. of the U.S., Recommendation 73-3, *Quality Assurance Systems in the Adjudication of Claims of Entitlement to Benefits or Compensation*, 38 Fed. Reg. 16840 (June 27, 1973).

[2] Jerry L. Mashaw, Report in Support of Recommendation 73-3 (June 30, 1974) (report to the Admin. Conf. of the U.S.).

recommendation recommended that agencies utilize "positive caseload management systems,"[3] including "statistical quality assurance reporting systems,"[4] to ensure accurate, timely, and fair adjudication.

Our Report builds upon the 1973 recommendation in three main respects. First, we draw on the experience of several agencies to implement quality assurance efforts, designed both to measure quality and to improve decision-making. The shift toward performance management in the 1990s, in particular, provides valuable lessons on the design of such systems.[5] Second, we focus on institutional design, which we view as critical to implementing a meaningful program of quality assurance.[6] Third, we describe several promising emerging tools that build on the 1973 recommendation's notion of "statistical quality assurance reporting systems."[7] In particular, we highlight the role of data infrastructure, analysis, and machine learning to aid in improving the accuracy and consistency of decisions.

At the outset, we note several limitations to our Report. First, because a small number of agencies, most notably the Social Security Administration, have had the most extensive experience in designing quality assurance programs, we rely heavily on these cases, but aim to extract general lessons. Due to rising caseloads and attendant challenges for decisional accuracy, numerous agencies have only recently started to make institutional commitments to quality assurance. Our Report aims to highlight steps agencies can take as they continue to pilot and develop relevant programs. Second, both agency experience with and academic study of quality assurance has focused on high-volume adjudication.[8] Still, agencies with lower volumes of matters may find much of our advice helpful and we attempt to identify practices that appear more feasible to low-volume adjudication. Third, our report does not address the design of internal appeals systems, addressed in another ACUS recommendation[9] and its accompanying report.[10] We do, however, offer guidance for how quality assurance assessments can make use of decisions rendered by both appellate tribunals within an agency and reviewing courts.

---

[3] Recommendation 73-3, *supra* note 1, ¶ 1.
[4] *Id.* at ¶ 2.
[5] *See*, *e.g.*, Government Performance and Results Act of 1993, Pub. L. No. 103-62, 107 Stat. 285 (codified as amended in scattered sections of 5 U.S.C. and 31 U.S.C.); NAT'L PERFORMANCE REVIEW, OFFICE OF THE VICE PRESIDENT, FROM RED TAPE TO RESULTS: CREATING A GOVERNMENT THAT WORKS BETTERS & COSTS LESS (1993); DAVID E. OSBORNE & TED GABLER, REINVENTING GOVERNMENT (1993); BA RADIN, CHALLENGING THE PERFORMANCE MOVEMENT: ACCOUNTABILITY, COMPLEXITY, AND DEMOCRATIC VALUES (2006) (criticizing performance measure movement); Burn S. Barnow, *Exploring the Relationship between Performance Management and Program Impact: A Case Study of the Job Training Partnership Act*, 19 J. POL'Y ANALYSIS & MGMT. 118 (2000) (evaluating the impact of performance measures).
[6] *See* Daniel E. Ho & Sam Sherman, *Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement*, 13 ANN. REV. L. & SOC. SCI. 251, 253-54 (2017) (providing a typology of management techniques organized, in part, by different institutional design choices).
[7] Recommendation 73-3, *supra* note 1,¶ 2.
[8] *See*, *e.g.*, David Ames et al., *Due Process and Mass Adjudication: Crisis and Reform*, 72 STAN. L. REV. 1 (2020) (analyzing quality assurance at three high-volume agencies).
[9] Admin. Conf. of the U.S., Recommendation 2020-3, *Agency Appellate Systems*, 86 Fed. Reg. 6618 (Jan. 22, 2021).
[10] Christopher J. Walker & Matthew Lee Wiener, Agency Appellate Systems (Dec. 14, 2020) (report to the Admin. Conf. of the U.S.).

We also note some clarifications about terminology. When discussing programs in which a non-governmental party applies for a benefit, license, permit, etc. administered by an agency, we use the generic term "claimant" to refer to the non-governmental party involved in the adjudication. However, the specific term applicable varies by the agency (e.g., patent applicant, asylum seeker, and disability claimant). Similarly, we use "quality assurance," "quality review," and "quality improvement" programs interchangeably to refer to practices to improve the quality of decision making, but in some academic works, these terms can refer to somewhat distinct concepts. In addition, when discussing quality review models, we analyze *both* formalized quality assurance programs (e.g., independent units with full-time employees dedicated to quality review) and less formalized interventions and practices, such as peer review or unstructured quality feedback.[11]

Our Report proceeds as follows. Part II articulates the motivations for quality assurance, major challenges for assuring the quality of adjudicatory decision-making, and the need for a focus on systemic improvement, as opposed to case-by-case error correction. Part III addresses quality assurance from an institutional design perspective by presenting several best practices for constructing a quality assurance program. Part IV describes emerging tools in quality assurance, paying particular attention to recent innovations in data infrastructure, data analysis, and machine learning. Part V concludes with future areas for research.

II. Motivation

A. Due Process and Decisional Accuracy

When government agencies adjudicate the rights of veterans, immigrants, the disabled, and other claimants, agencies have a responsibility to ensure quality decision-making. Of course, "quality" can be hard to define, and agencies have employed a range of approaches. The SSA, for instance, defines a high-quality decision as "factually accurate, procedurally adequate, supported by the record, and policy compliant."[12] The Board of Veterans Appeals (BVA) defines accuracy as an opinion that is "deficiency free."[13] The Merit Systems Protection Board (MSPB) considers the "quality of decision" based on "proper identification of all material legal and factual issues," "[a]ppropriate . . . consideration of relevant facts, evidence, and authority," "[p]roper and thorough analysis of the issues," style, and citation formatting.[14]

---

[11] We do so with an eye toward agencies concerned about the costs of starting an independent quality assurance program.

[12] Kurt Glaze et al., *Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance*, *in* OXFORD HANDBOOK ON AI GOVERNANCE 7 (Justin Bullock et al. eds., forthcoming 2022). The Equal Employment Opportunity Commission (EEOC) has adopted a similar standard. *See Federal Sector Quality Practices for Effective Hearings, Appeals and Oversight*, U.S. EQUAL EMP. OPPORTUNITY COMM'N, https://www.eeoc.gov/federal-sector/federal-sector-quality-practices-effective-hearings-appeals-and-oversight (last visited Oct. 28, 2021).

[13] U.S. GEN. ACCOUNTING OFFICE, GAO-02-806, VETERANS' BENEFITS: QUALITY ASSURANCE FOR DISABILITY CLAIMS AND APPEALS PROCESSING CAN BE FURTHER IMPROVED 7 (2002).

[14] U.S. MERIT SYS. PROT. BD., PERFORMANCE ELEMENTS AND STANDARDS, ATTORNEY EXAMINER (ADMINISTRATIVE JUDGES), GS-905-13/14/15 (2021).

We use "accuracy" as shorthand for these considerations. In doing so, we acknowledge that decisional quality involves considerations not fully captured by this term.[15] The Equal Employment Opportunity Commission (EEOC), for instance, specifies an extensive set of determinants, including timeliness and hearing fairness, in its *Federal Sector Quality Practices for Effective Hearings, Appeals and Oversight*.[16] Various agencies' complaint procedures recognize adjudicator comportment as crucial to their mission and to claimants' experiences.[17] Without excluding attention to these other considerations, however, decisional accuracy should always remain a core emphasis for a quality assurance program for several reasons. Agencies have relatively straightforward measures for timeliness goals, for instance, while measures of decisional accuracy prove comparatively underdeveloped.[18] Likewise, while adjudicator comportment is particularly critical for decision-making processes that involve hearings, accuracy may be paramount for the millions of decisions made without claimant appearances before an ALJ or AJ.[19] Regardless of the specific definition or its emphasis, however, the practices we identify below are relevant to ensuring decisional quality.

In several respects, the justification for an agency's commitment to quality decision-making is self-evident. Adjudication often involves matters, such as disability benefits or defenses to deportation, where outcomes are of supreme importance to the individuals engaging with the agency. An agency that erroneously denies claims harms often vulnerable populations and may deepen "social malaise" with such execution of its mission.[20] An agency that erroneously grants claims would be a fiscal drain on the public coffers and weaken public confidence in the policy it administers.[21] Public confidence may also weaken if an agency's adjudicators decide similar cases inconsistently, as such patterns might suggest that outcomes can turn as much on adjudicator assignment as on the facts and the law.

But decisional accuracy is not just a matter of wise policy administration. Rather, an agency's obligation to quality decision-making emanates from due process principles,[22] in addition to rationality constraints under the Administrative Procedure Act. If some base level of accuracy—ensuring that those entitled to prevail do so, and that those not entitled do not—cannot be guaranteed, procedural due process may demand more. Under *Mathews v. Eldridge*'s well-known

---

[15] *See* Jerry L. Mashaw, *The Supreme Court's Due Process Calculus for Administrative Adjudication in Mathews v. Eldridge: Three Factors in Search of a Theory of Value*, 44 U. CHI. L. REV. 28, 48 (1976).
[16] *See* U.S. EQUAL EMP. OPPORTUNITY COMM'N, *supra* note 12.
[17] *See Complaints Regarding EOIR Adjudicators*, U.S. DEP'T OF JUSTICE, https://www.justice.gov/eoir/complaints-regarding-eoir-adjudicators (last visited Nov. 9, 2021); *How to File an Unfair Treatment Complaint Concerning an Administrative Law Judge*, SOC. SEC. ADMIN., https://www.ssa.gov/pubs/EN-05-10071.pdf (last visited Nov. 9, 2021).
[18] *See EOIR Performance Plan: Adjudicative Employees*, EXEC. OFFICE FOR IMMIGRATION REVIEW, https://www.justice.gov/eoir/page/file/1358951/download (last visited Nov. 9, 2021) (discussing case completion goals).
[19] *See generally* Christopher J. Walker & Melissa F. Wasserman, *The New World of Agency Adjudication*, 107 CAL. L. REV. 141, 153-154 (2019).
[20] Goldberg v. Kelly, 397 U.S. 254, 265 (1970).
[21] Mathews v. Eldridge, 424 U.S. 319, 348 (1976).
[22] *See* Ames et al., *supra* note 8, at 20-25 (2020) (discussing connection between decisional accuracy, quality assurance programs and due process case law).

balancing framework, due process requires weighing the individual's interest in the sought-after benefit or relief, the government's interest, and "the risk of an erroneous deprivation of [the individual's] interest through the procedures used."[23]

The *Mathews* calculus is best conceived of in terms of *systemic accuracy*, not accuracy in each lone-standing case. Procedural due process protects the accuracy of the system as a whole.[24] Recommendation 1973-3 and the accompanying report anticipated this aggregate focus, emphasizing systemic improvement as quality assurance programs' objective.[25] Compared to other forms of error correction, such as episodic appeals, quality assurance programs may be better positioned to understand the source of adjudicators' error and the most effective interventions to remedy them.[26] These programs, then, can effectively realize due process's commitment to systemic accuracy.

B. Challenges for Decisional Quality

Taking stock of best practices for quality assurance is particularly important given the challenges agencies face in the complexity, volume, and appellate review for adjudication.

First, administrative cases often involve voluminous, complex factual records, including technical or scientific considerations, as well as complex and evolving legal frameworks. Administrative judges (AJs)[27] or administrative law judges (ALJs) must assemble a record, potentially hear the testimony of claimants and witnesses, weigh the evidence, conduct legal research, and issue a written decision to justify the outcome.[28] A case may involve complex statutory provisions and regulations, which are subject to change and, frequently, open to more than one interpretation. And, even where the law is clear, difficult factual circumstances often require adjudicators to make tough judgment calls.[29]

---

[23] Mathews v. Eldridge, 424 U.S. 319, 335 (1976).
[24] *See* Richard H. Fallon, Jr., *Some Confusions About Due Process, Judicial Review, and Constitutional Remedies*, 93 COLUM. L. REV. 309, 336-37 (1993) (describing *Mathews* as requiring procedures that achieve a "tolerable average level of accuracy"); Mashaw, *supra* note 15, at 48; Gillian E. Metzger, *The Constitutional Duty to Supervise*, 124 YALE L. J. 1836, 1865 (2015) (describing *Mathews*' "decidedly systemic and managerial cast"); Robert L. Rabin, *Federal Regulation in Historical Perspective*, 38 STAN. L. REV. 1189, 1314 (1986) (describing *Mathews* as concerned with "potential error costs" to the group of potential recipients).
[25] *See* Mashaw, *supra* note 2, at 161 (stating caseload management systems are "essential" to promoting accurate, timely, and fair adjudication).
[26] *See id.* at 165-66 (highlighting potential weaknesses of using appeals as a tool for error correction).
[27] Here we use AJs as a catchall term for a host of non-ALJ hearings. *See* Kent Barnett et al., Non-ALJ Adjudicators in Federal Agencies: Status, Selection, Oversight, and Removal 1 (Sept. 24, 2018) (report to the Admin. Conf. of the U.S.) (detailing the prevalence of non-ALJs in administrative adjudication).
[28] *See*, *e.g.*, U.S. GOV'T ACCOUNTABILITY OFF., GAO-17-438, IMMIGRATION COURTS: ACTIONS NEEDED TO REDUCE CASE BACKLOG AND ADDRESS LONG-STANDING MANAGEMENT AND OPERATIONAL CHALLENGES 13-19 (2017) (summarizing steps required to conduct typical immigration court proceeding).
[29] *See*, *e.g.*, U.S GOV'T ACCOUNTABILITY OFF., GAO-05-655T, BOARD OF VETERAN'S APPEALS HAS MADE IMPROVEMENTS IN QUALITY ASSURANCE, BUT CHALLENGES REMAIN FOR VA IN ASSURING CONSISTENCY 9-10 (2005) (discussing the judgmental calls VLJs must make to determine severity of disability).

Second, agencies may also face substantial pressures from high caseloads and production expectations.[30] Between 2005 and 2014, the Executive Office for Immigration Review (EOIR), for instance, experienced a 44 percent growth in immigration court cases.[31] In 2018, approximately 1,500 SSA ALJs processed over 750,000 dispositions.[32] Staffing has not always kept up with workload. For example, 65 percent of Asylum Officers reported to the GAO in 2008 that they had insufficient time to thoroughly adjudicate cases,[33] while 82 percent of Immigration Judges reported time constraints as challenging to their capacity to adjudicate cases.[34] In recent years, U.S. Customs and Immigration Services' Office of Administrative Appeals has trimmed supervisory review of some officers' draft decisions as caseloads have risen.[35]

A rise in caseloads may increase pressure to meet production goals, which in turn may raise the risk of adjudicator error.[36] According to a GAO survey of SSA ALJs, many adjudicators are concerned that output demands reduce the accuracy of their decisions.[37] In the words of one SSA ALJ, "If productivity increases, quality decreases."[38] Mandatory performance measures, particularly those developed to keep backlogs at bay, may pressure adjudicators to focus more on quantity and timeliness than accuracy.[39] Agencies may find ways to increase output and maintain and even improve quality.[40] But agencies have also appeared to acknowledge the potential negative impact of

---

[30] *See Social Security Disability Benefits: Did a Group of Judges, Doctors, and Lawyers Abuse Programs for the Country's Most Vulnerable?*: *Hearing Before the S. Comm. on Homeland Sec. & Governmental Affairs*, 113th Cong. 129 (2013) (statement of Debra Bice, Chief ALJ, Social Security Administration) (summarizing caseload pressures facing SSA); Daniel L. Skoler, *The Many Faces of High-Volume Administrative Adjudication: Structure, Organization, and Management*, 16 J. NAT'L ASS'N ADMIN. L. JUDGES 43, 58 (1996) (providing caseload statistics across several agencies).
[31] U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 28, 20.
[32] *Annual Statistical Supplement, 2019*, SOC. SEC. ADMIN., https://www.ssa.gov/policy/docs/statcomps/supplement/2019/2f8-2f11.html (last visited Aug. 13, 2021).
[33] U.S. GOV'T ACCOUNTABILITY OFF., GAO-08-935, AGENCIES HAVE TAKEN ACTIONS TO HELP ENSURE QUALITY IN THE ASYLUM ADJUDICATION PROCESS, BUT CHALLENGES REMAIN 7 (2008).
[34] *Id.* at 64.
[35] Interview with Agency Official(s), U.S. Citizenship & Immigr. Serv. (Oct. 12, 2021) [hereinafter UCIS Interview].
[36] *See*, *e.g.*, Heckler v. Day, 467 U.S. 104, 113-15 (1984) (summarizing Congressional concerns about the effect of caseloads on decisional quality in SSA adjudication); U.S. GOV'T ACCOUNTABILITY OFF., GAO-21-242, MEANINGFUL PERFORMANCE MEASURES COULD HELP IMPROVE CASE QUALITY, ORGANIZATIONAL EXCELLENCE, AND RESOURCE MANAGEMENT 28 (2021) (noting some NLRB stakeholders believe heavy emphasis on timeliness leads staff to "cut corners"); LEWIN GRP., INC. ET AL., EVALUATION OF SSA'S DISABILITY QUALITY ASSURANCE (QA) PROCESSES AND DEVELOPMENT OF QA OPTIONS THAT WILL SUPPORT THE LONG-TERM MANAGEMENT OF THE DISABILITY PROGRAM 17 (2001) (noting "a widespread belief in a trade-off between accuracy and productivity" among SSA adjudicators); Stuart L. Lustig et al., *Inside The Judges' Chambers: Narrative Responses from the National Association of Immigration Judges Stress and Burnout Survey*, 23 GEO. IMMIGR. L.J. 57, 65-66 (2008) (reporting perceptions about quantity-quality tradeoff among IJs).
[37] U.S. GOV'T ACCOUNTABILITY OFF., GAO-09-398, ADDITIONAL PERFORMANCE MEASURES AND BETTER COST ESTIMATES COULD HELP IMPROVE SSA'S EFFORTS TO ELIMINATE ITS HEARINGS BACKLOG 26 (2009).
[38] U.S. GOV'T ACCOUNTABILITY OFF., GAO-21-341, SOCIAL SECURITY DISABILITY: PROCESS NEEDED TO REVIEW PRODUCTIVITY EXPECTATIONS FOR ADMINISTRATIVE LAW JUDGES 42 (2021).
[39] *See id.* (reporting results from surveys of agency staff). *See also* Ass'n of Admin. Law Judges v. Colvin, 777 F.3d 402, 405 (7th Cir. 2015) (discussing the relationship between quality and quantity in SSA ALJ decision-making).
[40] Gerald K. Ray & Jeffrey S. Lubbers, *A Government Success Story: How Data Analysis by the Social Security Appeals Council (with a Push from the Administrative Conference of the United States) Is Transforming Social Security Disability Adjudication*, 83 GEO. WASH. L. REV. 1575, 1605-06 (2015).

quantity pressures on quality. As part of agency efforts to improve decisional quality, for instance, the SSA capped the number of decisions SSA ALJs could hear per year in 2013 to 960 dispositions.[41]

To be sure, ALJs have robust job protections, which make formal employment consequences for failure to meet performance expectations less likely absent severe disparities.[42] But informal motivations, including concerns for professional reputation and the desire to keep up with peers, may affect ALJ decision-making. Moreover, performance pressures may impact the decision-making of non-ALJ adjudicators and staff attorneys, who frequently lack formal job protections that ALJs enjoy.[43]

Finally, individual protections—the traditional lynchpin of due process doctrine—often fail to produce systemic accuracy and consequently fail to protect many claimants. Due process doctrine provides that, as the risk of decisional inaccuracy rises, the procedural protections individuals may choose to avail themselves of expands.[44] As the 1973 Recommendations recognized, and as subsequent research has confirmed, institutional factors may blunt the efficacy of standard procedural rights to ensure systemic accuracy.[45] The right to appeal, to name one such protection, may correct errors in individual cases. But, as discussed below, institutional determinants unrelated to decisional quality may affect the composition of appellate dockets and produce a misleading sample of decisions for reviewing courts and tribunals to evaluate. The decision to detain an immigrant, for instance, will affect the likelihood that this immigrant appeals an Immigration Judge's order, however flawed. More generally, reliance on individuals to exercise particular rights will fail when these individuals lack the resources or wherewithal to avail themselves of available protections. The right to counsel's assistance, for example, means little to an individual who lacks the means to hire a lawyer. Finally, decisions in individual appeals, without more, do not translate into systemic learning. By the early 2000s, for instance, SSA had learned that reasons for remands of ALJ decisions tended to largely remain static, and that remands had not changed ALJ behavior significantly.[46]

An agency should not rely exclusively on parties threatened or adversely affected by an erroneous decision to take action to correct it. To ensure systemic accuracy, quality assurance requires a proactive commitment by the agency to rigorous measurement and improvement.

---

[41] *See* OFFICE OF THE INSPECTOR GEN., SOC. SEC. ADMIN., A-12-15-15005, THE SOCIAL SECURITY ADMINISTRATION'S EFFORTS TO ELIMINATE THE HEARINGS BACKLOG 3-4 (2015) (discussing production caps instituted by SSA's Office of Disability Adjudication and Review).
[42] Shapiro v. Social Security Administration, 800 F.3d 1332 (Fed. Cir. 2015).
[43] *See* Letter from Douglas E. Massey, Counsel, Bd. of Veterans' Appeals, to David Shulkin, Sec'y of Veterans Affairs 1 (Sept. 18, 2017) (suggesting an increased production quota may lead to a decline in the quality of staff decision-making).
[44] Mathews v. Eldridge, 424 U.S. 319, 335 (1976).
[45] *See* David Hausman, *The Failure of Immigration Appeals*, 164 U. PA. L. REV. 1177, 1193 (2016).
[46] Ray & Lubbers, *supra* note 40, at 1591.

III. Institutional Design

We now outline the key design dimensions for quality assurance programs: the who, what, how, when, and so what of institutional design. Who should serve in the quality assurance program? What cases and whose work should they select to review? How should such a review be conducted to measure the accuracy of decisions? When should review be conducted? And what follows from such results to improve the quality of decision-making?

Agencies with large dockets and sizeable cadres of adjudicators should answer these institutional questions as they design formal quality assurance programs, with staff dedicated to quality review. But these questions deserve attention even at agencies that can devote fewer resources to standalone quality assurance programs. The specific tools deployed may vary, but a smaller-sized docket or adjudicator corps does not lessen an agency's obligation to systemic improvement.

A. Quality Assurance Personnel

1. Who Participates?

Quality assurance personnel should meet three basic criteria: (1) they must have significant experience in a decision-making domain; (2) they must command the respect of peers; and (3) they must be willing to exercise independent judgment about decisional accuracy. The rationale for the first criterion is straightforward. Without the relevant expertise, a reviewer will not be able to comment meaningfully on the quality of an adjudicator's decision. The second relates to quality assurance's ultimate goal – systemic improvement. If adjudicators do not trust or respect a reviewer's feedback, they are less likely to respond productively. Finally, independence ensures that a quality assurance program remains focused on quality, as objectively defined as possible, and does not get deployed in the service of other agency goals or concerns.[47]

These criteria create a dilemma of institutional design.[48] On the one hand, quality reviewers may be appointed from outside the division of the agency handling adjudication, ensuring greater independence at the cost of less expertise. On the other hand, appointing quality reviewers from inside the division of the agency handling adjudication poses potential conflicts of interest, exchanging independence for greater expertise.

---

[47] *Cf.* Ass'n of Admin. Law Judges, Inc. v. Heckler, 594 F. Supp. 1132, 1142 (D.D.C. 1984) (criticizing an SSA quality review program for its "unjustifiable preoccupation with allowance rates" and suggesting that the agency deployed the program to lower allowance rates and not necessarily to improve quality).

[48] This type of dilemma has been studied extensively in the private sector by research exploring the independence of corporate boards and management. *See*, *e.g.*, ADA DEMB & F.-FRIEDRICH NEUBAUER, THE CORPORATE BOARD: CONFRONTING THE PARADOXES 6-7 (1992) (describing the tension between critical and independent judgment on corporate boards); John W. Byrd & Kent A. Hickman, *Do Outside Directors Monitor Managers?: Evidence from Tender Offer Bids*, 32 J. FIN. ECON. 195 (1992) (analyzing the effect of outside directors on outside bid offers); Stuart Rosenstein & Jeffrey G. Wyatt, *Outside Directors, Board Independence, and Shareholder Wealth*, 26 J. FIN. ECON. 175 (1990) (studying the effect of outside directors on firm performance).

A related design problem lies in the relationship between reviewers and adjudicators. Quality reviewers who are otherwise *subordinates* to adjudicators may not be able to exhibit the desired independence, for fear of subsequent employment consequences.[49] A more benign version of this problem involves existing social relationships. Reviewers may also not be able to independently assess colleagues' work product,[50] although a recusal policy can safeguard against such potential conflicts.

Conversely, too great of an institutional difference between adjudicators and reviewers can breed distrust and diminish adjudicators' receptivity to feedback.[51] Quality assurance schemes centered on appellate review can illustrate these dynamics. Tension between frontline adjudicators and appellate tribunals is not uncommon. An ALJ or AJ may believe that judges on a reviewing court lack sufficient understanding of the conditions they labor under, have a skewed understanding of their docket, or have unreasonable expectations for appropriate decision-making.[52] Adjudicators may perceive appellate bodies as adopting an insufficiently deferential standard of review.

With these criteria in mind, agencies can pick from three personnel models of quality assurance. First, supervisors can review subordinates' work for quality.[53] Second, agencies can form independent quality review teams, which can be made up of adjudicators or other agency personnel, such as staff lawyers.[54] Third, peers can review their peers' work, an approach which can be institutionalized or run on a more informal basis. And these models can be used in a complementary fashion. For example, at the MSPB, supervisors conduct formal quality assessment for new AJs, while in some offices all AJs participate in an informal peer review.[55]

When deciding on quality assurance personnel, some agencies may be worried about costs and capacity. Smaller agencies or those with low case volumes may have fewer resources they can devote to an independent quality assurance team staffed with full-time personnel. For these agencies,

---

[49] U.S GOV'T ACCOUNTABILITY OFF., QUALITY ASSURANCE FOR DISABILITY CLAIMS AND APPEALS PROCESSING CAN BE FURTHER IMPROVED 7 N 7 (2002).

[50] *See*, *e.g.*, Vanessa Urch Druska & Steven B. Wolff, *Effects and Timing of Developmental Peer Appraisals in Self-Managing Groups*, 84 J. APPLIED PSYCH. 58, 58 (1999) (describing potential for "popularity contents" in peer appraisals); Maury A. Peiperl, *Conditions for the Success of Peer Evaluation*, 10 J. INT'L HUM. RES. 429, 431 (1999) (describing potential for bias in peer evaluation).

[51] *See* Greg L. Stewart, *A Meta-Analytic Review of Relationships Between Team Design Features and Team Performance*, 32 J. MGMT. 29, 33 (2006) (summarizing studies on heterogenous vs. homogenous teams).

[52] *See* Ames et al., *supra* note 8, at 61 (reporting results from interviews with adjudicators).

[53] For example, supervisors at the EEOC review every decision drafted by attorneys working on appeals, and a significant component of this review involves quality assessment. Interview with Agency Official(s), U.S. Equal Emp. Opportunity Comm'n (Oct. 25, 2021) [hereinafter EEOC Interview].

[54] Such independent quality assurance programs were the institutional forms envisioned in the 1973 Recommendation and Report. *See generally* Recommendation 73-3, *supra* note 1 (recommending agencies adopt a positive caseload management systems and statistical quality assurance reporting systems). Currently, the SSA operates a quality assurance program which is modeled roughly along these lines. *See* Felix F. Bajandas & Gerald K. Ray, Implementation and Use of Electronic Case Management Systems in Federal Agency Adjudication 44-52 (2018) (report to the Admin. Conf. of the U.S.) (describing how the SSA Appeals Council and division of quality conduct quality review).

[55] Interview with Agency Official(s), U.S. Merit Sys. Prot. Bd. (Aug. 18, 2021) [hereinafter MSPB Interview].

other institutional forms that draw on existing staff and involve less oversight, such as a peer review, may be easier to implement. Still, informal review like peer review isn't costless. To be successful, peers must dedicate time to reviewing their colleagues' work, which naturally gives them less time to decide cases.[56]

2. Length of Assignments

Next, agencies must decide *how long* reviewers will engage in quality assurance. Will reviewers take on temporary positions or engage in the work indefinitely? And, if they are temporary, will there be a rotation system cycling agency staff on and off?

Short-term positions have the advantage of providing fresh perspectives, a wider set of reference points, and a potential infusion of quality perspectives into the frontlines if employees return to the ranks.[57] Longer-term positions, on the other hand, enable reviewers to gain experience and institutional knowledge over time, can reduce challenges from turnover, and enable longer-term investments in quality initiatives. Such long-term positions, however, also run the risk of groupthink and a sense of distance from the operational constraints of frontline work. Quality review teams, for instance, should not be perceived as what some administrative judges noted about appellate courts: that the appellate court is distant and lacks an understanding of the day-to-day demands of high caseloads.[58]

Whether quality reviewers serve for a year, a decade, or in perpetuity, quality review should be recognized as a distinct job responsibility. To avoid the perception that quality review can be a kind of "unfunded mandate," agencies should ensure reviewers have sufficient bandwidth to fulfill their quality review responsibilities. In particular, productivity pressures should not be allowed to crowd out time spent on quality review assignments.[59] Relatedly, if quality review assignments are temporary, successful participation should count toward career advancement within the agency.

B. Whose Work Gets Reviewed?

A quality assurance system should review the work of all personnel who have important roles in the adjudication of cases or matters. This seemingly straightforward design principle becomes more complicated in implementation as the range of individuals involved in decision-making broadens.

---

[56] For a more in-depth discussion about peer review, see Part IV (C).

[57] Such tradeoff between experience and new perspectives has long been studied in the management literature. *See*, *e.g.*, Susan Cohen & Diane E. Bailey, *What Makes Teams Work*, 23 J. MGMT. 239, 273 (1997) (discussing tradeoff between the creative benefits of diverse perspectives with coordination costs); Dawn Harris & Constance Helfat, *Specificity of CEO Human Capital and Compensation*, 18 STRATEGIC MGMT. J. 895, 897-99 (1998) (comparing the value of "fresh perspectives" with firm specific expertise).

[58] *See* Ames et al., *supra* note 8, at 61 (providing examples from interviews with veterans law judges).

[59] *See* U.S GOV'T ACCOUNTABILITY OFF., SOCIAL SECURITY DISABILITY: ADDITIONAL MEASURES AND EVALUATION NEEDED TO ENHANCE ACCURACY AND CONSISTENCY OF HEARINGS DECISIONS 40 (2017) (noting that SSA quality assurance staff struggled to conduct reviews due to productivity pressures).

In a number of agencies, support staff play critical roles in the review of evidence and the drafting of decisions. At SSA, for instance, ALJs review evidence, hold hearings, and make decisions, but may rely on support staff to write a decision's first draft. An ALJ bears ultimate responsibility for any decision that goes out under that adjudicator's name and, in theory, should catch and correct errors in draft orders before effectuation. But caseload pressures and other challenges may complicate the rigor of this review. Relatedly, an ALJ may discount the feedback they receive on erroneous decisions if the adjudicator believes that low-quality work by support staff makes improvement futile.

A quality assurance program that only reviews ALJ or AJ work product may fail to identify where in a decision-making process errors are made and by whom. One that also evaluates work done by support staff may better focus attention on the source of error and identify more efficient, effective interventions.[60]

A quality assurance program should also subject all adjudicators' work to review and not exempt or phase out those with more seniority. Even experienced adjudicators who diligently adjust their decision-making to new laws or policies can develop decisional heuristics that can produce error. These heuristics, or analytical shortcuts that decision-makers develop to aid them as they work through complex problems, can increase an adjudicator's productivity. But an adjudicator, even an experienced one, may rely on a heuristic (a mental shortcut) that causes their decision to be non-compliant with current policy.[61]An agency can leverage a probationary period of review for new adjudicators or staff to set a cultural expectation that quality review is a consistent and expected part of agency practice.

C. Selecting Cases for Quality Review

If caseloads and available resources permit an agency to review every decision for accuracy, there may be strong reasons to do so. Otherwise, when selecting cases for review, agencies should choose a case selection method that (a) provides a representative perspective on quality issues and (b) enables focused inquiry into distinct sources of errors.

As the 1973 Report highlighted, sampling has several advantages.[62] First, it is more efficient to analyze a subset of cases.[63] Second, sampling naturally focuses quality review on system-wide

[60] If appropriate, agencies might also consider tracking compliance with their final decision as a component of quality review. For instance, the EEOC monitors whether agencies implement remedies ordered by the commission (e.g., reinstatement). *See* U.S. EQUAL EMP. OPPORTUNITY COMM'N, *supra* note 12.
[61] *See* Ray & Lubbers, *supra* note 40, at 1598-99, 1598 n. 147 (describing policy compliant "pathing").
[62] Mashaw, *supra* note 2, at 170-71.
[63] *Id.* at 170.

improvement rather than one-off error correction.[64] Third, proper sampling reveals patterns in decisions and consequently helps agencies separate random and systematic errors.[65] There are, however, some disadvantages to simple random sampling. Most importantly, reviewers may spend too much time on cases that are already known to have low error rates. Random sampling may also be inefficient at examining specific issues. And while random sampling can help estimate overall error rates, an exclusive focus on the overall error rate can fail to provide sufficiently detailed information that is actionable.

Several refinements can help agencies achieve specific goals in quality review. S*tratified random sampling*[66] is an alternative that allows agencies to sample cases in a way that is representative but over-samples cases based on chosen characteristics (e.g., claim type, issue). Such an approach can focus on specific legal issues or factual circumstances suspected to cause errors.

Thus far, quality assurance programs have used both sampling strategies. The Office of Patent Quality Assurance (OPQA) randomly selects different agency actions for review.[67] For over 15 years, the BVA's quality assurance program randomly sampled 5 percent of original decisions by the Board of Veterans Appeals.[68] SSA has used various sampling strategies for quality review since the mid-1970s.[69] For instance, its Appeals Council, the agency's internal appellate body, both randomly samples from ALJ decisions and samples through stratification based on issues and facts, but not based on the identity of the decisionmaker or the office location.[70]

The need for a representative perspective on errors explains why agencies should not rely solely on appeals for quality review.[71] As noted in the 1973 Recommendation, appeals can present a distorted picture of decisional quality.[72] Litigants choose which cases to appeal and strategically

---

[64] *Id.* at 170-71. To flesh this point out, Mashaw compares sampling a subset of cases with analyzing 100 percent of decisions. When analyzing the entire population, Mashaw argues there is less pressure for agencies to adopt system-wide improvement since they are already correcting all errors in a one-off fashion. *Id.*

[65] *Id.* at 171.

[66] Note that random sampling and stratified sampling are not necessarily mutually exclusive. When implementing a stratified sampling strategy, agencies can randomly sample within their selected strata.

[67] *Promoting the Useful Arts: How Can Congress Prevent the Issuance of Poor Quality Patents?: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 116th Cong. (2019) (responses to questions for the record of Andrew Hirshfeld, Commissioner for Patents, USPTO).

[68] Daniel E. Ho et al., *Quality Review of Mass Adjudication: A Randomized Natural Experiment at the Board of Veterans Appeals, 2003–16*, 35 J. L. ECON. & ORG., 239, 240-41 (2019).

[69] Ames et al., *supra* note 8, at 32.

[70] 20 C.F.R. § 404.969(b)(1) (2020).

[71] Besides the possibility of bias, other institutional features of appeals make them less helpful as a tool for quality review. For one, appeals (whether internal or external) can take years to complete, degrading the usefulness of any accuracy signal for systemic learning. *See, e.g.*, Ames et al., *supra* note 8, at 19 (discussing delays in SSA and VA appeals). Additional challenges present themselves when relying on external appeals. Frequently, appeals make up a small fraction of courts caseloads, and external judges often lack expertise. See, e.g., Jonah B. Gelbach & David Marcus, *Rethinking Judicial Review of High Volume Agency Adjudication*, 96 TEX. L. REV. 1097, 1158-60 (2018) (discussing the "experience critique" as applied to district court judges and SSA appeals).

[72] *See* Mashaw, *supra* note 2, at 165 (describing the "mysteriously selective" nature of appeals).

select cases that are more likely to succeed.[73] Parties often have asymmetric appeal rights, such as laws or regulations that only allow those denied benefits to appeal their claim.[74] When this happens, some outcomes—for example, a grant of benefits—are rarely reviewed. In addition, many individuals lack sufficient representation or find appeals to be too costly and therefore may not bring forward meritorious claims.[75] For instance, one study of quality assurance found the prevalence of errors in all cases was indistinguishable from non-appealed cases, suggesting non-appealed cases can have a significant rate of error.[76]

That said, appellate outcomes still provide useful information. For example, agencies can conduct anomaly detection (the detection of anomalous patters or outliers) to identify the adjudicators, issues, or fact patterns generating a disproportionate share of appeals and remands.[77]

Importantly, an agency should administer a sampling strategy attentive to its impact on different types of claimant groups. An agency that over-samples a particular trait that affects claims, for instance, may select cases that mostly involve one gender of claimants or claimants from a particular ethnic group. The agency should ensure that whatever results its quality assurance program generates, it does not lead adjudicators to treat claimants differently, and certainly not disfavorably, based on these demographic criteria. Relatedly, an agency should be attentive to the possibility that adjudicators may focus more on cases they believe subject to over-sampling, while paying less attention to other matters on their dockets. A sampling strategy is arguably counter-productive if it improves decision-making in one domain while weakening it in others.

Also, case selection should not improperly infringe upon adjudicators' qualified *decisional independence*.[78] Quality assurance programs perceived to erode decisional independence and push outcomes in political directions have attracted significant legislative and judicial pushback.[79] A quality assurance program should not be used as a tool of political control to drive specific adjudicatory outcomes but should be employed to assess whether outcomes align with agency

---

[73] *See generally* Steven Shavell, *The Appeals Process as a Means of Error Correction*, 24 J. LEGAL STUD. 379 (1995) (describing the strategic incentives for appeals and comparing possible institutional rules for selecting appeals).
[74] *See* Gelbach & Marcus, *supra* note 71, at 1109 (describing the "baseline problem" in appeals to external courts); *see also* Bice, *supra* note 30 (discussing how ALJ decisions granting SSA disability benefits are not subject to appeal).
[75] *See* Hausman, *supra* note 45, at 1193 (analyzing connection between immigrants' decision to appeal and whether they have a lawyer).
[76] Ho et al., *supra* note 68, at 261.
[77] *See*, *e.g.*, Gelbach & Marcus, *supra* note 71, at 1142-44 (illustrating the use of data heat maps).
[78] ALJs are afforded a high degree of decisional independence. *See* 5 U.S.C. § 4301(2)(D) (2018) (exempting ALJs from the definition of "employee" subject to performance appraisal); *see also* Nash v. Califano, 613 F.2d 10, 15 (2d Cir. 1980) (recognizing that ALJs have "a qualified right of decisional independence"). That said, agencies can—and, in practice, do—review ALJ decisions for quality. As a general standard, ALJ decisions can be measured against existing legal standards, provided that the agency sets the standards out clearly and there is a general consensus as to their meaning. And, moreover, ALJs are not the only adjudicators entitled to decisional independence. Other adjudicators (such as AJs) are granted significant independence through statutory law and agency policy.
[79] *See* Nash v. Bowen, 869 F. 2d 675, 680-81(2d Cir. 1989) (describing the potential for quality assurance systems to infringe upon ALJ decisional independence).

policy, guidance, and regulations, as well as inform areas for policy development.[80] Importantly, the Administrative Procedure Act does not prohibit the review of ALJ decisions for these purposes.[81]

D. Determining Quality

A central question in the design of quality assurance programs—often glossed over—is the standard by which reviewers assess quality. This is the internal analogue to administrative law's core question about the standard of judicial review. How much deference should reviewers give to adjudicators on findings of fact and the application of law to facts? An appropriate quality measure may also implicate concerns about decisional consistency, and it requires agencies to think through how best to account for appellate outcomes.

1. Standard of Review

No single standard of review may be uniformly appropriate for all agencies, as the costs of errors can differ dramatically across contexts. The less deference the reviewer affords, the more rigorous a threshold the quality assurance program will set for an error-free decision. But such review also increases the workload for the quality assurance team and either increases decisional costs for frontline adjudicators or can set unachievable standards. At an extreme, a strict standard may effectively substitute the quality assurance team's judgment for the adjudicator's when reasonable minds might differ.

One important reference point is the ultimate judicial review standard. Some agencies may expressly indicate that quality review should anticipate whether an appeals court would reverse or remand the decision.[82] This "predictive" notion of the standard of the review has several advantages, namely (a) vesting the standard of review with some external source of legitimacy, so adjudicators understand it and respect it, (b) reducing the overall number of appeals and thus avoiding the costs and burdens of appellate review, (c) addressing the asymmetric access to appellate review problem, and (d) avoiding inter-branch conflict over proper policy administration.

While such a predictive notion is sensible as a starting point, it may not be sufficient to meet other objectives of quality assurance. For instance, if an appeals court reviews findings of fact more deferentially, a less deferential standard of review (e.g., de novo review) may make sense at the

---

[80] For example, the Merit Systems Protection Board (MSPB) formally includes the quality of decisions as an element in their performance evaluation of their administrative judges (AJs). In particular, the agency analyzes how successfully the AJ identifies material legal and factual issues, recognizes relevant facts, evidence or authority, and the overall quality of their analysis. *See* U.S. MERIT SYS. PROT. BD., *supra* note 14.

[81] *See* U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 38, 8 (explaining that agencies can review ALJ decisions for compliance with policy and procedures).

[82] *See* BD. OF VETERANS' APPEALS, U.S. DEP'T VETERANS AFF., ANNUAL REPORT FISCAL YEAR (FY) 2002 7 (2002) (describing how BVA's quality review overturns decisions when an error would be "outcome determinative" at appellate body).

quality assurance level to ensure that all individuals with matters before the agency get treated equally. In that sense, quality review can serve in part as a complement and in part as a substitute to appellate review. In addition, a predictive standard may be difficult to administer if an agency's decisions get appealed to courts in different circuits, and if the stringency of judicial review differs geographically.[83] Quality reviewers may struggle with the perceived inconsistency of appellate reviewers as well.

The predictive standard can also unnecessarily restrict the focus of quality review on outcomes. While agencies should care about *outcomes*, quality control of the decisional *reasoning* may be equally important to understand policy compliance. Adjudicators can reach the right decision for the wrong reasons, especially if the law is complex and rapidly evolving, leading to potential errors in other decisions.[84] Moreover, quality assurance should include an assessment of antecedent steps in the process leading to a decision, including, for instance, the thoroughness with which the agency develops the record.

Whatever standard of review the agency employs, it should be transparent about how strictly its quality review team evaluates decisions and whether and how the team uses a predictive measure. Some agencies, including, for instance, the Board of Veterans' Appeals, include quality review measures in reports available to the public.[85] This sort of reporting is a helpful accountability measure. But policymakers and the public may have a hard time discerning the meaning of an accuracy measure that the agency reports unless the agency reports the precise criteria it uses for quality review. Such lack of transparency fundamentally impedes oversight efforts of adjudicatory agencies. Similarly, agencies should prioritize consistently reporting the results of quality review programs when they disclose these results in reports to Congress or other publicly accessible documents. Inconsistent reporting makes it difficult for quality measures to sufficiently incentivize institutional change and learning within the agency.

Agencies should also maintain the stringency of their selected standard. Although quality assessment may have one *de-facto* standard, pressure from internal or external stakeholders to boost accuracy measures may cause reviewers to relax their evaluation.[86] For instance, reviewers might adopt a predictive standard based on appellate outcome but, in reality, evaluate decisions on a much more lenient curve.[87]

Finally, agencies should take steps to validate their selected standard in at least two ways. They should take steps to ensure that different reviewers are administering the standard in roughly

---

[83] *See* David Marcus & Jonah Gelbach, A Study of Social Security Litigation in the Federal Courts 74-75 (July 28, 2016) (report to the Admin. Conf. of the U.S.) (describing practice adopted by SSA hearing office).
[84] For instance, an adjudicator might arrive at the correct conclusion while applying a five-factor balancing test but still misapply two factors in their decision.
[85] *E.g.,* BD. OF VETERANS' APPEALS, U.S. DEP'T VETERANS AFF., ANNUAL REPORT FISCAL YEAR (FY) 2019 15 (2019).
[86] *See* Ames et al., *supra* note 8, at 52-53 (explaining how one agency's standard of review for quality assurance weakened without being formally changed).
[87] *Id.*

consistent ways to ensure inter-rater reliability. Quality reviewers themselves could be reviewed, or quality reviewers could review each other's work. The perception that different reviewers score quality differently can be a source of frustration for adjudicators. In addition, the agency should attempt to determine whether its standard of review conforms with other possible measures of agency performance. If, for instance, decisions sampled, reviewed, and pronounced error-free are appealed and remanded as often as decisions that go unreviewed altogether, the agency may need to reconsider its standard.[88]

2. Decisional Inconsistency

An agency may consider using consistency among decision-makers, or whether adjudicators with similar dockets generate significantly different results, as a quality measure. One ALJ deciding disability benefits appeals, for instance, granted claims 89 percent of the time, while another ALJ in the same hearing office, dealing with the same claimant population, granted claims 19 percent of the time.[89]

Decisional inconsistency is not an indicator of poor-quality decision-making *per se*. Adjudicator independence and the discretion that fact-intensive adjudication requires make variation an irreducible reality. But significant disparities could indicate a cohort of adjudicators with fundamentally mistaken understandings of agency policy, or they could indicate policy ambiguity causing widespread confusion. While an agency should avoid naively equating inconsistency (at low levels) with poor quality decision-making, decisional inconsistency may still invite targeted quality review, whereby the agency focuses on a particular issue or set of adjudicators for evaluation. Targeted review can raise concerns about intrusions on decisional independence and fairness to claimants, thereby fueling ALJ and public opposition to the program. SSA's "Bellmon Review" program in the early 1980s, for instance, initially used a targeted sampling protocol that selected for ALJs particularly inclined to allow benefits. The program prompted litigation and significant political blowback that ultimately ended it.[90] However, agencies can target individuals or sets of adjudicators in a post-adjudicatory review designed to gather information about the overall quality of the work. SSA describes this type of review as a focused review.[91]

---

[88] *See* Ames et al., *supra* note 8, at 48 (2020) (summarizing empirical analysis of BVA's Quality Review Program).
[89] Harold J. Krent & Scott Morris, *Inconsistency and Angst in District Court Resolution of Social Security Disability Appeals*, 67 HASTINGS L.J. 367, 378 (2016).
[90] Ho et al., *supra* note 68, at 33-35.
[91] *See* U.S. GOV'T ACCOUNTABILITY OFFICE, *supra* note 59, 31 (discussing SSA's "focused reviews").

3. Appellate Outcomes as a Measure

Another quality measure used by agencies is the rate at which an adjudicator's decisions are overturned on administrative appeal or judicial review.[92] For example, agencies may analyze how often an internal appeals body overturns decisions, or how often cases are reversed by courts.[93] While such decisions are observable and relevant to the concerns of adjudicators (who presumably do not like to be overturned), they should not be the sole measure of quality. For the reasons discussed above, they present a biased picture of decision-making. And, even if appeals were not a biased quality signal, agencies should not rely on any single quality measure because quality is a multi-dimensional concept.[94]

E. Timing

When should quality review be conducted? Existing practices have centered on two models: (1) review after a decision has been drafted, but prior to when it has taken effect and been issued to the claimant, and (2) review after a decision has taken effect and been issued to the claimant.

Pre-issuance review, sometimes called in-line review, has several advantages. Deficiencies can be corrected before a decision is issued. Less time passes between the adjudicator's work on the case and quality review, and thus the adjudicator is more likely to understand and respond to feedback. But agencies must avoid interfering with any decisional independence granted to adjudicators by law or agency policy. Quality review feedback that comes weeks, months, or even years after a decision's issuance risks being ignored, especially if an adjudicator with a large docket has rendered numerous decisions in the interim.

The SSA runs a well-established "in-line quality review program." Staff in SSA's regional offices randomly sample cases at various stages before completion. They review these cases for compliance with policy and whether the evidence supports the adjudicator's drafted decision.[95] Since this quality review process occurs before the decision is finalized, the "in-line" review helps the SSA

---

[92] Beyond the measures discussed in this section, there likely exist a wide range of factors agency's use to evaluate quality. For instance, at least one agency has turned to the public to gather information about the quality of their adjudication process. *See* BD. OF VETERANS' APPEALS, *supra* note 85, 20 (discussing surveys conducted after customer "touches" with agency). This survey asks questions about the claimant's experience, such as whether they received a timely hearing and if their adjudicator explained issues clearly. *See*, *e.g.*, *Veterans Signals (VSignals) Survey*, U.S DEP'T VETERANS AFF., https://www.jackson.va.gov/features/Veterans_Signals_VSignals_Survey.asp (last visited Sept. 26, 2021). Although agencies should seek to understand the public's experience, they should be cautious about using "customer" surveys to make conclusions about decisional quality. Among other potential pitfalls, such surveys may tilt adjudicators toward outcomes favorable to claimants. Moreover, there is no inconsistency between a claimant's positive experience with an adjudicator, on the one hand, and decisional error, on the other.

[93] *See*, *e.g.*, U.S. MERIT SERV. PROT. BD., ANNUAL PERFORMANCE REPORT FOR FY 2020 AND ANNUAL PERFORMANCE PLAN FOR FY 2021 (FINAL) & FY 2022 (PROPOSED) 7 (2020) (reporting both measures).

[94] *See* Pierre J. Richard, *Measuring Organizational Performance: Towards Methodological Best Practice*, 35 J. MGMT. 718, 722 (2009) (recommending firms avoid relying on a single performance measure since performance is a multidimensional construct).

[95] SOC. SEC. ADMIN., PROGRAM OPERATIONS MANUAL SYSTEM, GN 04440.008 *Quality Review Process* (2019).

analyze instances when support staff, such as decision writers, may struggle with correct policy administration.

Post-decisional review, by contrast, will not infringe on an adjudicator's pace of decision-making by requiring the adjudicator to redo work, and thus it will do less to delay decisions. For SSA disability benefits adjudication, post-decisional reviews are known as "pre" or "post" effectuation reviews, depending upon whether the decision has been implemented by initiation of benefits. The pre-effectuation reviews allow the agency to take corrective action within 60 days of issuance of the decision.

In general, evidence from performance-management literature suggests that pre-effectuation review or ongoing assessment -- that is assessment that occurs while the decision process is still underway -- is more likely to be effective.[96] By providing feedback closer in time to when an adjudicator decides a matter, pre-effectuation review is more likely to reflect agency-wide commitment to quality, thereby cultivating a culture of learning and improvement.[97] But pre-effectuation review can take place late in the process, usually after the ALJ has made their initial decision. Also, as discussed below, agencies can develop AI tools that can help flag possible errors for adjudicators as they craft decisions.

One such ongoing management technique is peer review. Agencies can assign peers to shadow adjudicators on the same case to offer peer feedback, potentially before the final decision is issued. The ultimate decision may still rest with the primary adjudicator. Since peer review practice can be about deliberation and not necessarily performance evaluation, it can foster collaborative learning. Several regions of the MSPB, for instance, have relied extensively on a peer review system, wherein judges review each other's draft decisions.[98] In Part V(C), we discuss experimental evidence that suggests that ongoing quality initiatives, live simulation trainings and peer review may help review adjudicatory processes prior to a draft decision.

F. Feedback Mechanisms

How should the quality review team provide feedback to adjudicators? Current practices vary widely. One agency, for example, issues memoranda to individual adjudicators in reviewed cases, but only when the quality assurance team identifies a deficiency.[99] Other agencies simply tabulate aggregate performance measures, indicating the overall accuracy rate for the adjudicator corps as a whole.[100]

---

[96] *See* Ho & Sherman, *supra* note 6, at 265 (concluding existing literature suggests ex-ante and ongoing assessment techniques are more effective than ex-post tools).
[97] *See id.* (emphasizing that management intervention and commitment are crucial to efforts to improve quality).
[98] MSPB Interview, *supra* note 55.
[99] Ho et al., *supra* note 68, at 247.
[100] *See* BD. OF VETERANS' APPEALS, *supra* note 85, 15.

Several best practices have emerged.[101] First, feedback should (a) be *issued expeditiously*, (b) include both *positive* and *negative feedback*, and (c) be provided to all agency personnel who participate in the decision-making process, including key support staff. Feedback gleaned from the review of a decision issued months or even years earlier may have little value if the adjudicator has no memory of the case. Or it might impose significant costs, if the adjudicator needs to review the case's record anew to understand what went wrong and thereby take time the adjudicator might otherwise devote to new cases. Consequently, performance reviews should be one of several mechanisms for delivering quality-related feedback.[102] Although performance reviews for non-ALJ adjudicators occur regularly, more frequent feedback may be even more impactful.[103] For instance, SSA's How MI Doing tool is a web-based program that provides individualized information to ALJs about their decision-making, including the number of decisions they have rendered and whether any of their decisions have been remanded, on a continuous basis. It uses information from the ALJ's prior decisions and automatically flags relevant training modules based on particular issues after a remand is issued.[104]

Given random sampling, an adjudicator should also learn of those instances when a quality review team finds no error in a case selected for review. Such positive signals enable adjudicators to distinguish higher-quality decisions[105] from those not selected for review that may contain errors. Feedback should also go to support staff involved in key aspects of decision-making. One agency provides an error memorandum only to the AJ, and not staff attorneys, depriving the latter of important learning opportunities.

Second, while case-level feedback is helpful, agencies should also design quality review to synthesize information from a group of decisions, identify commonly occurring errors, and deliver focused training materials or group feedback.[106] By providing information about recurring problems with decisional quality, quality assurance moves from assessment to system-wide improvement. Recurring errors may demonstrate widespread misunderstandings of governing policy that targeted

---

[101] *See*, *e.g.*, MANUEL LONDON, JOB FEEDBACK: GIVING, SEEKING AND USING FEEDBACK FOR PERFORMANCE IMPROVEMENT 16 (2d ed. 2003) (recommending feedback be immediate and contain both positive and constructive elements).

[102] We view quality assurance as related to performance evaluation but emphasize that it should be considered a distinct enterprise. While quality review can provide helpful information about where individuals can improve performance, the larger goal of quality assurance is identifying systematic problems.

[103] ALJs are exempt from performance appraisal requirements that apply to other adjudicators. *E.g.*, Kent Barnett, *Against Administrative Judges*, 49 U.C. DAVIS L. REV. 1643, 1655 (2016). The possibility of annual performance reviews may make quality feedback easier to implement for agencies with non-ALJ adjudicators. Still, the use of information generated through quality assurance processes should not lessen an agency's obligation to conduct these reviews in a manner sensitive to principles of adjudicator independence. *See generally* Kent Barnett et al., Non-ALJ Adjudicators in Federal Agencies: Status, Selection, Oversight, and Removal (Feb. 14, 2018) (report to the Admin. Conf. of the United States).

[104] U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 59, 33.

[105] We say "higher quality" because the standard of review for quality assurance purposes does not mean that the underlying decision is necessarily error-free.

[106] As one recent GAO report pointed out, regularly distributing aggregate feedback also helps ensure the agency is tracking and addressing quality issues. *See* U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 36, 23-24 (recommending NLRB develop quantifiable quality metrics and follow agency policy for distributing quality findings).

training could quickly and efficiently fix. They might also highlight policy ambiguity, enabling the agency to develop guidance and policy clarification that improves accurate and consistent decision-making by frontline adjudicators. Such syntheses may be particularly valuable when quality signals are delayed. Consider, for example, remands from external courts. To make this information more useful, agencies can (1) aggregate remand decisions, (2) detect remand patterns, and (3) issue reports to staff and adjudicators detailing their findings on a *regular basis*.[107] At least one SSA hearing office issues a report along these lines summarizing district court decisions in social security cases.[108] Internal surveys of ALJs at the office suggest many support the practice.[109] Numerous staff members indicated that the reports improved the quality of their decisions.[110]

Third, agencies can also incorporate appeals information to improve internal review. For example, agencies can assess whether specific adjudicators are triggering a high rate of remands and intervene to provide "focused review" of the specific legal issues or fact patterns triggering such remands.

Finally, the results of quality assurance should not only be shared with adjudicators, but also with agency leadership, as results may suggest areas for policy improvement. This feedback loop may be particularly important to create a learning organization.[111] The results of quality review led SSA to determine that certain issues that produced persistent adjudicator error revealed ambiguity in agency policy. These findings prompted SSA to provide policy clarification and guidance that enabled adjudicators to decide cases more consistently and accurately.[112] Other agencies can use the results of quality review similarly. USCIS Administrative Appeals Office supervisors meet regularly and can discuss recurring issues with decision-making that they encounter. These discussions could inform instances when the agency issues an "adopted decision," or a decision on an appeal that clarifies policy internally and illustrates for officers how this policy applies.[113] EEOC uses information gleaned from appeals to draft reports for and provide technical assistance to other agencies whose personnel are the first-instance adjudicators when federal employees make complaints.[114]

---

[107] Along these lines, the EEOC has produced reports that detail common problems causing them to either remand or reverse agencies' handling of employment discrimination claims. *See* U.S. EQUAL EMP. OPPORTUNITY COMM'N, PRESERVING ACCESS TO THE LEGAL SYSTEM: COMMON ERRORS BY FEDERAL AGENCIES IN DISMISSING COMPLAINTS OF DISCRIMINATION ON PROCEDURAL GROUNDS 8 (2014) (identifying failure to state a claim and time limit rules as common sources of errors).

[108] *See* Marcus & Gelbach, *supra* note 83, at 119-20 (describing practice adopted by SSA hearing office).

[109] *Id.*

[110] *Id.*

[111] *See* David A. Garvin, *Building a Learning Organization*, 71 HARV. BUS. REV. 78 (1993).

[112] Ray & Lubbers, *supra* note 40, at 1601.

[113] UCIS Interview, *supra* note 35.

[114] EEOC Interview, *supra* note 53.

IV. Emerging Tools for Quality Assurance

Since ACUS's 1973 recommendations, and with increasingly in recent years, agencies have engaged in significant experimentation in methods and tools for "statistical quality assurance."[115] Here, we highlight four of those emerging tools: data infrastructure, data-driven quality insights, collaborative learning and peer review, and artificial intelligence.

A. Data Infrastructure

A successful quality assurance program requires that the agency capture sufficient data about decisions in its case management system. Historically, adjudicatory agencies have kept only minimal information about cases, largely for case handling and reporting purposes. As the 1973 Report noted, capturing salient information about decisions is critical to enabling analysis and actionable insights.[116] At a minimum, a case management system should record: (a) *who* made the decision and which support staff assisted, (b) the procedural history of the case, (c) basic information about the issues presented and decision outcomes, and (d) subsequent outcomes (e.g., whether an appeal was taken and the outcome of the appeal). If possible, the case management system should collect this data directly from the adjudicator as part of the case management system. Relying on third parties to hand code cases after the fact introduces the possibility for error and delay.

In some instances, adjustments to existing practices can generate important data to inform quality assurance practices. Supervisors at the USCIS Administrative Appeals Office flag issues as part of their regular review of officers' draft decisions.[117] These edits and corrections, if coded for and captured in a case management system, may provide a rich source of information to help inform a systemic inquiry into recurring sources of error. A coding sheet gets attached to an attorney's draft decision at the EEOC's Office of Federal Operations, capturing information about the issues on appeal.[118] These sheets could invite input from supervisors reviewing draft decisions, enabling the agency to capture data on the sorts of errors that attorneys most commonly make.

We emphasize that this is merely a starting point. The richer this "metadata," the greater the opportunities for building in quality monitoring and quality checks. In designing case management systems, agencies should thus include members of the quality review team to understand what fields, functionality, and reporting are useful to build in at the outset. Ideally, case management systems can be designed to support adjudicators by prompting for structured input (e.g., the type of benefit sought) and then leverage this information to recommend decision tools (e.g., training materials and

[115] Mashaw, *supra* note 2, at 170.
[116] *Id.* at 170-171.
[117] USCIS Interview, *supra* note 35.
[118] EEOC Interview, *supra* note 53.

decision templates). Some systems may even incorporate case analysis tools to guide adjudicators through the necessary steps for a policy-compliant decision.[119]

Data infrastructure has proven essential to SSA's recent quality assurance efforts. SSA has implemented a state-of-the-art electronic case management system (eCMS).[120] The system stores records and documents related to each claim electronically, captures metadata about case processing, and identifies information about the witnesses, representatives, and adjudicators involved.[121] But it includes more. Embedded within eCMS is a case management tool that captures medata about whether an ALJ followed policy. The Appeals Council developed a logic tree mapping the issues that must be considered when making 2,000 different types of decisions.[122] Appeals Council members then compiled a list of reasons why cases are remanded. From this list, they identified 156 different specific types of errors, each of which they then organized into one of ten broad categories that follow the sequential evaluation process SSA uses to adjudicate disability benefits claims.[123] This effort enabled consistency in information storage. The Appeals Council assigned each error a numeric code, enabling Appeals Council staff to code errors consistently and readily.[124]

SSA's "focused quality reviews" illustrate the benefits of a supportive data infrastructure.[125] Indicators like remand patterns or other data anomalies trigger an "early monitoring system," to identify an ALJ who may be struggling with the correct application of some policies. A team of Appeals Council personnel then conducts an in-depth examination of that ALJ's decisions, to identify areas of concern and develop an appropriate intervention. Focused quality reviews also have helped SSA to identify areas where policy ambiguity creates challenges for consistent, correct decision-making and thus to help guide policy clarification.[126]

We also recommend that agencies consider publishing metadata and decisions with identifying information removed. This disclosure should include information about decisions selected for quality review and that review's outcome. Currently, many agencies only publish highly aggregated reports, disclosing, for instance, the number of decisions reviewed and the overall accuracy rate. Publishing this data in de-identified form would enable the kind of research that provided the basis for ACUS's 1973 Recommendation, as well as many of the underlying studies relied upon here.

---

[119] *See* Glaze et al., *supra* note 12, at 7 (discussing "pathing" in case management tools); Ray & Lubbers, *supra* note40, at 1593-94 (describing policy compliant "pathing").
[120] *See* Bice, *supra* note 30.
[121] *See* Ray & Lubbers, *supra* note 40.
[122] *Id.*
[123] *Id.*
[124] Glaze et al., *supra* note 12.
[125] Ames et al., *supra* note 8, at 39.
[126] *See generally* Nicole Maestas et al., *The Effect of Economic Conditions on the Disability Insurance Program: Evidence from the Great Recession*, 199 J. PUB. ECON. 104410 (2021).

Finally, disclosure would facilitate independent oversight to ensure the integrity of the information generated by the agency's quality assurance program. Rather than simply take the agency's word that its adjudicators are correct in 95 percent of instances, for example, a congressional committee, inspector general, or outside researcher could ascertain the basis for the agency's accuracy rate. Independent oversight of an agency's quality review program has a number of benefits. Most importantly, adjudicators may accept critique and feedback more readily if they believe that independent evaluators have reviewed and approved of a quality assurance program. They will less likely view the program as motivated by management priorities unrelated to decisional accuracy and thus be more willing to act on the program's recommendations. Effective outside oversight may resolve some of the dilemmas an agency must address when deciding who participates in quality review. If independent evaluators can examine the program's adequacy with requisite data, the agency may be better positioned to appoint insiders, with deeper policy expertise and experience, to a quality review team notwithstanding those experts' relationships with adjudicators.

Good data infrastructure and information disclosure, then, may be true lynchpins for a successful quality assurance program.

B. Data-Driven Quality Insights

With adequate data infrastructure in place, agencies have developed key quality assessment measures. First, agencies have conducted forms of anomaly detection (i.e., identifying unexplained discrepancies in cases), using quality review output or remand orders to look for certain issues, jurisdictions, or adjudicators that exhibit high propensity for errors. SSA, for instance, developed a series of "heat maps" to understand which issues and field offices were subject to high error or remand rates. The agency then responded with targeted interventions, including focused trainings.[127] Such analyses might be particularly helpful in identifying areas of law where a precedent has shifted, but adjudicators have not fully adjusted, or where policy ambiguity, inconsistency, or gaps lead to recurring problems The BVA, for instance, conducted an in-depth analysis of one precedent, whose citation was associated with very high remand rates, and issued training materials for how to handle that precedent appropriately.[128] SSA heat maps helped the agency prioritize areas for regulatory change, informing a series of proposals for policy clarification.[129]Agencies have also used data indicating decisional inconsistency to identify problems. SSA heat maps group data from Appeals Council review of ALJ decisions into a column-and-row format. This presentation enables personnel to analyze errors at the hearing office or the individual ALJ level, to determine whether inconsistencies result from adjudicator idiosyncrasy or something more systematic.

As mentioned, high variation in similar cases might reveal a systemic misunderstanding or misapplication of policy. The goal is of course not to reduce variation to zero, but consistency can

---

[127] *See* U.S. GOV'T ACCOUNTABILITY OFFICE, *supra* note 59, 31 (describing SSA's "focused reviews").
[128] BD. OF VETERANS' APPEALS, TARGETED QUALITY ANALYSIS: APPEALS OF EXTRASCHEDULAR RATINGS UNDER 38 C.F.R. § 3.321(B)(1) 2 (2002) (analyzing application of *Thun* precedent).
[129] Ray & Lubbers, *supra* note 40, at 1607.

be an important diagnostic tool for identifying areas with quality challenges. In a non-adjudicatory context, for instance, one agency tracked which specific issues were generating high variances between decision makers to develop training materials, when it was not clear which direction was objectively correct.[130]

C. Collaborative Learning and Peer Review

In some ways, quality review in some agencies resembles a conventional law school exam: the review is conducted near the end of the process, with virtually no intermediate feedback or assessment. And just as the practice of a single end-of-the-term exam runs against much of the evidence of effective pedagogy,[131] review of near-final decisions may be ineffective for purposes of quality improvement. We hence note two emerging practices that may enable quality review to benefit decision-making from early stages of the adjudicatory process on.

The first are collaborative learning models. One SSA field office, for instance, synthesized remand orders on a semiannual basis to extract general lessons for quality improvement.[132] An online platform also enables adjudicators to reference training materials as they are working cases. Simulation exercises—rather than passive lecture-based learning—may be particularly valuable to transform knowledge into practice.[133]

The second is a more bottom-up approach to organizational learning, suggested by much of the literature on organizational learning and "democratic experimentalism,"[134] which focuses on forms of "peer review."[135] The Patent and Trademark Office, for instance, adopted peer review to enable patent examiners to share knowledge about prior art searches.[136] MSPB AJs often review each

---

130 *See* Daniel E. Ho, *Does Peer Review Work? An Experiment of Experimentalism*, 69 STAN. L. REV. 1, 56-57 (2017) (describing huddle training process).

131 *See* K. Anders Ericsson, *The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance*, *in* THE CAMBRIDGE HANDBOOK OF EXPERTISE AND EXPERT PERFORMANCE (K. Anders Ericsson et al. eds., 2006) (describing how repetitive feedback improves expertise); Linda Darling–Hammond et al., *Implications for Educational Practice of The Science of Learning And Development*, 24 APPLIED DEV. SCI. 90, 120 (2020) (summarizing literature which finds detailed and repetitive feedback helps students learn); Daniel Schwarcz & Dion Farganis, *The Impact of Individualized Feedback on Law Student Performance*, 67 J. LEGAL. EDUC. 139, 162 (2017) (finding individualized feedback from professors improves student test-taking performance).

132 Marcus & Gelbach, *supra* note 83, at 119-20.

133 *See generally* J. Patrick McCarthy & Liam Anderson, *Active Learning Techniques Versus Traditional Teaching Styles: Two Experiments from History and Political Science*, 24 INNOVATIVE HIGHER EDUC. 279 (2000) (conducting experiments comparing student-based group-exercises with traditional teacher-based lectures).

134 Michael C. Dorf & Charles F. Sabel, *A Constitution of Democratic Experimentalism*, 98 COLUM. L. REV. 267, 289 (1998).

135 *See* John Braithwaite & Valerie Braithwaite, *The Politics of Legalism: Rules Versus Standards in Nursing-Home Regulation*, 4 SOC. & LEGAL. STUD. 307, 322 (1995) ("The beauty of a small number of broad standards is therefore that one can design a regulatory process to ensure that the ticking of a met rating means that a proper process of information-gathering and team deliberation has occurred on that standard."); Dorf & Sabel, *supra* note 134, at 356; Charles F. Sabel & William H. Simon, *Minimalism and Experimentalism in the Administrative State*, 100 GEO L.J. 53, 93 (2011) ("[E]xperimentalist regimes . . . strive for accountability less through simple rules than through peer review of local discretion.").

136 Daniel E. Ho & Lisa Larrimore Ouellette, *Improving Scientific Judgments in Law and Government: A Field Experiment of Patent Peer Review*, 17 J. EMPIRICAL LEGAL STUD. 190, 193 (2020).

other's decisions before effectuation, building on the supervisory review that MSPB conducts of AJ decision-making during the adjudicator's first year on the job.[137] One field study for health code inspections suggested that peer review, where inspectors jointly visited an establishment and reviewed each other's work, both improved the accuracy and consistency of inspections.[138] Peer review has the advantage of being seen as less driven by top-down mandates, and can hence foster a more open attitude to learning.[139] Peer review also presents an opportunity for adjudicators to actively collaborate with colleagues while holding jobs which may feel isolating.[140] One of the downsides, however, is that instituting such a system necessarily reduces the speed of case processing of cases. Also, peer review requires some degree of oversight, as a group of frontline adjudicators might coalesce around an incorrect understanding of policy and approve each other's erroneous decisions.

One interesting type of peer review conducted in SSA is known as the Appeals Council Feedback Initiative. This initiative provides an opportunity for ALJs to raise concerns about cases remanded back to them, in essence providing an opportunity for line ALJs to provide feedback to their appellate reviewers. The initiative incentivizes ALJs to participate because raised concerns that are upheld are reflected in dispositional quality statistics maintained for each ALJ. The concerns raised by the ALJs are reviewed on a post-adjudicative basis, by the Administrative Appeals Judge (AAJ) who initiated the remand order. That AAJ provides an analysis of the concerns and offers an opinion about those concerns. If disagreement remains between the ALJ and AAJ, a panel of two management ALJs and two management AAJs meet to further discuss and resolve the issues. This panel provides feedback to the ALJ and AAJs involved in the case. Additionally, the information the panel gleans is more broadly shared in Appeals Council meetings, and is seen by participants as a way of improving relations between line judges and their appellate reviewers, while closing gaps between the way issues are adjudicated at these two adjudicative levels.

D. Artificial Intelligence

One of the frontier tools of quality assurance lies in the use of "artificial intelligence" (AI), namely the use of machine learning to help identify quality issues in decisions. The most innovative of these tools is SSA's Insight system, which permits adjudicators to upload draft decisions and returns over 30 "quality flags" for potential errors.[141] Quality flags can range from simple uses of natural language processing (e.g., does the opinion cite a provision of the C.F.R. that does not exist?) to more ambitious uses of machine learning that identify internal inconsistencies between portions of the decision (e.g., does the decision find that the claimant can be gainfully employed in an

[137] MSPB Interview, *supra* note 55.
[138] Ho, *supra* note 130, at 60-69.
[139] *Id.* at 84.
[140] *Id.* at 13. Peer review can build a sense of common cause, potentially boosting employee satisfaction. As one agency official told us about their peer review program, it's important to feel agency decision are "not just yours" but a "office decision." MSPB Interview, *supra* note 55. Agencies can also structure peer review so peers collaborate across offices, a design which may improve group cohesion across the agency. *See* Ho, *supra* note 130, at 71.
[141] Glaze et al., *supra* note 12, at 14-15.

occupation despite finding a functional impairment that would disqualify the claimant from such position?). Although there was no rigorous evaluation,[142] SSA found that the adoption of the Insight system was associated with an improvement in processing time and a reduction in errors.[143]

As this is a rapidly moving territory, we make several observations about the use of AI in quality assurance initiatives.[144] First, groundwork needs to be laid for agencies to consider deploying such tools responsibly. In particular, the data infrastructure noted above must be in place to develop and test such tools;[145] staff with both technical knowledge and expertise in the agency's adjudicatory responsibilities must be available;[146] and leadership should commit to an iterative development process, rather than a one-time off-the-shelf tool, in order to pilot, evaluate, and assess AI's use.[147]

Second, we note the potential for AI to impact the institutional design choices discussed in Part III. SSA's experience with Insight demonstrates how AI can help agencies shift from a purely *ex-post* model of quality review and toward *continuous assessment*.[148] Relatedly, Insight reflects the potential for AI to promote *continuous learning* for actors across the system. By using text analysis, Insight flags issues to both adjudicators as they write their decisions and appeals council staff as they review pending cases.[149]

Third, while Insight is based solely on the text of draft decisions, major quality improvements may be possible by incorporating the record that forms the basis of the decision (e.g., the claims folder). Natural language processing tools, for instance, may pinpoint the most salient records specific to an issue in a lengthy folder, similar to automated searches used in e-discovery. Recommendation engines may highlight relevant lines of precedent as well as factually similar cases and model decisions.[150] The explosion of research in natural language processing makes this a particularly promising area for improving the speed, accuracy, and consistency of claims adjudication.

Last, we note that governance and oversight of these systems is critical. SSA's case study is particularly valuable on this front. Experts in disability benefits law, policy, and adjudication developed SSA's AI system; it is used specifically to determine policy-compliant decision-making;

---

142 *See* OFFICE OF THE INSPECTORS GEN., U.S. SOC. SEC. ADMIN., A-12-18-503353, THE SOCIAL SECURITY ADMINISTRATION'S USE OF INSIGHT SOFTWARE TO IDENTIFY POTENTIAL ANOMALIES IN HEARING DECISIONS, at 11 (2019) (recommending SSA develop metrics to evaluate Insight's impact).
143 Glaze et al., *supra* note 12, at 20.
144 *See generally* David Freeman Engstrom et al., Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies (Feb. 2020) (report to the Admin. Conf. of the U.S.).
145 Glaze et al., *supra* note 12, at 23.
146 *See id.* at 16 (highlighting the role mixed expertise played in SSA's adoption of AI tools).
147 *Id.* at 19.
148 *See* Ho & Sherman, *supra* note 6, at 253-54 (describing typology of management techniques).
149 *See* Glaze et al., *supra* note 12, at 14 (describing how SSA uses Insight technology).
150 *See generally* Huang et al., *Context-Aware Legal Citation Recommendation Using Deep Learning*, *in* PROCEEDINGS OF THE 18TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 79 (2021) (showing machine learning can help improve citation prediction in mass adjudication).

and it serves as a decision aid, with ultimate discretion residing in adjudicators. AI systems may aid, but should not displace, the requisite human adjudicatory judgment.

V. Conclusion

As reflected in this Report, our understanding of quality assurance programs has made significant progress since ACUS's 1973 Report. Still, we have much to learn.

Most importantly, we have extremely little insight into how agencies have designed quality assurance efforts. We hence recommend that each agency—through its IG or QA unit—should collect systematic information detailing the agency's current QA program. Second, agencies should document the implementation of and changes to quality review programs. Third, agencies should publish regular reports on their implementation, administration, and development of QA initiatives. With this information in hand, future researchers can assess the impact of our Report and update the best practices discussed above. Agencies can also learn from each other, and oversight bodies like congressional committees can come to expect quality reports as regular and vital inputs for evaluating an agency's activities.

As the focus by scholars has turned toward the internal norms, practices, and policies by which agencies govern themselves,[151] quality assurance programs are central in our understanding of how adjudicatory agencies govern themselves. Future scholarship should push beyond our scope, paying particular attention to how quality review best operates in a small adjudicative setting and in non-adjudicatory settings. Many of the institutional design features mentioned also still need rigorous comparative study. What are the most effective feedback mechanisms for quality improvement? How do different case-selection methods impact quality assessment? Should quality reviews be incorporated into performance reviews?

While there are many open questions, we have also learned a tremendous amount about quality assurance and due process in the nearly 50 years since ACUS first issued its recommendation. We hope that less time will pass until we have further lessons to synthesize.

[151] *See* JERRY L. MASHAW, BUREAUCRATIC JUSTICE: MANAGING SOCIAL SECURITY DISABILITY CLAIMS 15, 149 (1983) (arguing "internal administrative law" reforms are more promising than those imposed from outside agency); Gillian E. Metzger & Kevin M. Stack, *Internal Administrative Law*, 115 MICH. L. REV. 1239, 1243 (2017) (describing how both administrative law scholarship and reality have "gone internal").

Appendix: Agency Interviews

Agency	Date
Merit Systems Protection Board	8/17/21
National Labor Relations Board	8/19/21
Board of Veterans' Appeals	9/20/21
National Labor Relations Board	9/30/21
United States Citizenship and Immigration Services	10/12/21
Social Security Administration	10/22/21
National Organization of Social Security Claimants' Representatives	10/25/21
Equal Employment Opportunity Commission	10/25/21
National Organization of Social Security Claimants' Representatives	10/25/21
Patent and Trademark Office	11/4/21