Received: 11 September 2022 | Revised: 28 June 2023 | Accepted: 30 June 2023

DOI: 10.1111/puar.13705

RESEARCH ARTICLE

PAR Public Administration Review ASPA

Institutional design and biases in evaluation reports by international organizations

Steffen Eckhard[1] | **Vytautas Jankauskas**[1] | **Elena Leuschner**[2]

[1]Zeppelin University, Friedrichshafen, Germany

[2]Department of Political Science, University of Gothenburg, Gothenburg, Sweden

Correspondence
Steffen Eckhard, Zeppelin University, Am Seemooser Horn 20, 88045 Friedrichshafen, Germany.
Email: steffen.eckhard@zu.de

Funding information
The German Research Foundation, Grant/Award Number: EC 506/2-1

Abstract

Governments spend hundreds of millions on evaluations to assess the performance of public organizations. In this article, we scrutinize whether variation in the institutional design of evaluation systems leads to biases in evaluation findings. Biases may emerge because influence over evaluation processes could enable the bureaucracy to present its work in a more positive way. We study evaluation reports published by nine international organizations (IOs) of the United Nations system. We use deep learning to measure the share of positive assessments at the sentence level per evaluation report as a proxy for the positivity of evaluation results. Analyzing 1082 evaluation reports, we find that reports commissioned by operative units, as compared to central evaluation units, systematically contain more positive assessments. Theoretically, this link between institutional design choices and evaluation outcomes may explain why policymakers perceive similar tools for evidence-based policymaking as functional in some organizations, and politicized in others.

Evidence for practice

- We show that the institutional design of evaluation systems in international organizations is associated with differences in evaluation findings (i.e., biases). Across more than 1000 evaluation reports, those commissioned by operative administrative units are on average systematically more positive than those commissioned by central evaluation units.
- Whether the senior management of the administration itself, or the organization's governing body (here, member states) controls evaluation system resources (staff, budget, agenda) does not systematically affect how positive evaluation findings are.
- Our measurement relies on computational text analysis of evaluation reports, which enables automatic extraction of performance measures from evaluation reports, often hundreds of pages long.
- Publications such as this will likely increase the pressure on evaluators, because it is now clear that text-based performance statements can be quantified, which enables benchmarking. When designing evaluation systems, decision-makers should therefore focus on enlarging the institutional distance between evaluators and those evaluated to ensure independent evaluation and objective results.

INTRODUCTION

Over the past decades, evaluation has evolved to become the main tool to assess the performance and impact of public sector organizations. Following-up on the features of modern public management, many national governments around the globe now commit to evidence-based policy-making and have installed novel units or entire agencies for the conduct of evaluation.[1] A common definition describes evaluation as "an assessment, conducted

as systematically and impartially as possible, of an activity, project, program, strategy, policy, topic, theme, sector, operational area or institutional performance" (UNEG, 2016). As data-driven assessment, evaluation resembles empirical academic research, but on a pre-defined research question, along a standardized set of criteria such as effectiveness or coherence (the OECD-DAC criteria[2]), and with a systematic process of stakeholder inclusion.

The ongoing proliferation of evaluation in public administration begs questions concerning its utility in the policy process. Can evaluation results, which are submitted to policymakers as lengthy and well-crafted reports, actually be seen as an objective source to inform policy decisions? Answering this question is crucial, given that evaluation results are used to enable policy learning, ensure accountability, and to trace the transformation pathways of major policies such as the international fight against climate change.

Two perspectives prevail in public administration literature: The one—we call it the *functional* perspective—depicts evaluation as an independent, value-free, and research-based assessment that contributes to learning and evidence-based policymaking in public administration (Lee, 2006; Rossi et al., 2004). The other approach—we call it the *political* perspective—stresses that evaluation in itself is a political activity that influences the allocation of resources and power in organizations. Questioning the independence of evaluation, scholars problematize that evaluators are rarely fully independent from the organization they asses and that evaluated bureaucracies seek for overly positive portrayals of their activities (Eckhard & Jankauskas, 2020; Jankauskas & Eckhard, 2023; Malik & Stone, 2018; O'Brien et al., 2010; Pleger et al., 2017; Raimondo, 2018; Taylor & Balloch, 2005; Wildavsky, 1972).

We argue that this dissent in the literature is due to a so far overlooked confounder of evaluation processes. From an institutionalist perspective, the conduct of evaluation should be dependent on the organizational context in which it takes place. More specifically, the extent to which the organizational design enables evaluated administrative units to influence the evaluation process should influence whether the resulting findings are biased (political perspective), or not (functional perspective).

In assessing this argument, we study evaluation in the context of international organizations (IOs), which offer a comparable class of cases. The secretariats of these organizations consist of an international civil service, similar to the domestic bureaucracy, which is why they are also called the international public administration (IPA) (Bauer et al., 2017; Eckhard & Ege, 2016; Stone & Moloney, 2019; Thorvaldsdottir et al., 2021). We focus on the United Nations (UN) system, comprised of over 20 organizations.[3] They all subscribe to the same system-wide evaluation guidelines, using the same evaluation definition, principles and quality assurances, and are therefore suitable for comparison. Nevertheless, we do acknowledge certain differences between national and international administrations and address their implications on generalization in the discussion section.

In UN organizations, the institutional design of evaluation systems varies on two dimensions that provide higher or lower potential for the IO administration to influence evaluation results. These are the location of the commissioning unit (central evaluation unit vs. decentral operative units) and whether the member state board or the IO senior management controls financial, agenda and staff resources of the organization's evaluation system (external versus internal control over evaluation system). This study therefore asks *whether variation in the institutional design of evaluation systems leads to systematic differences in the positivity of evaluation findings*.

Measuring and comparing whether evaluation results are more or less positive is challenging because most evaluation reports do not have a single outcome metric expressing success or failure. Following recent advances in literature on natural language processing (Eckhard et al., 2023), we compare the share of positive sentence-level assessments per report. Plausibly, bureaucrats whose work is being evaluated should be interested in a greater number of positive assessments as it signals a more successful activity.

For the analysis, we draw on a text corpus with 1082 evaluation reports by nine UN system IOs which were published between 2012 and 2020.[4] We classify close to one million sentences and calculate the share of positive sentences per report. Reports vary along the two dimensions of the institutional design as described above. We use descriptive evidence and regression modeling to estimate differences in the share of positive sentences.

Our findings make important empirical and theoretical contributions to public administration literature and practice. *Empirically*, we find that whether member states or the administration's senior management controls the evaluation system does not substantively affect the positivity of evaluation findings. By contrast, whether operative units commission evaluation themselves, compared with commissioning by central evaluation units, is associated with systematically more positive evaluation findings. These results also hold when including a set of control variables, such as the type of evaluated activity. Hence, going beyond the rich insights provided by interview and survey data used by previous studies, this is the first analysis of the politics of evaluation at the level of text-based evaluation reports and across multiple organizations.

Theoretically, we speak to ongoing disputes between political and 'pragmatic' (functional) perspectives in debates on evaluation and, more broadly, evidence-based policymaking (Cairney, 2016; see MacKillop & Downe, 2022; Sanderson, 2002). We do so by revealing the conditions under which the functional or the political perspective on such a technocratic tool like evaluation is likely to emerge. These conditions relate to the institutional

design of evaluation systems: Evaluation results are systematically more positive in organizational settings where the institutional distance between evaluators and those evaluated is reduced. Institutional design choices influence the production of evidence, such as evaluation, which may explain why policy-makers perceive similar knowledge-tools as functional in one organization, and political in another.

Next, we discuss the state of the art and present the theoretical argument. This is followed by our research design and data. After presenting the empirical analysis, we discuss our results and provide concluding remarks.

A FUNCTIONAL AND POLITICAL PERSPECTIVE ON EVALUATION

Domestically and at the international level, policymakers use evaluations to ensure accountability and organizational learning. Such a *functional perspective* assumes that evaluation reports are value-free technocratic documents that help improving the performance of public organizations (Lee, 2006; Rossi et al., 2004). Evaluation then serves as a final phase in the cyclic policy process, where existing policies are adjusted or new ones are initiated (Anderson, 1975; Boyne et al., 2004). The added value of evaluation thus lies in its potential to allow institutional learning, ensure organizational accountability, and offer objective course-corrections.

Another perspective emphasizes the *politics of evaluation*. Studies accentuate that decision-makers in public organizations may use evaluation results to justify their pre-defined bargaining positions: "Whenever an evaluation affects the future allocation of resources and, hence, a change in power relationships, it is a political activity" (Wergin, 1976, p. 76). Scholars therefore expect that evaluators may be put under pressure "to misrepresent findings" (Pleger et al., 2017, p. 316) or to select "evaluation questions and methods that make convenient outcomes more likely" (van Voorst & Mastenbroek, 2019, p. 626). Surveys among evaluation practitioners (Pleger et al., 2017; The LSE GV314 Group, 2014) indeed find that evaluators are frequently pressured to misrepresent their results by making them "look more positive or less negative than the evaluator thought was warranted" (Morris & Clark, 2012, p. 57).

While most studies draw on domestic level evidence, scholars recently shifted their attention to evaluation in international organizations as well. The literature has focused on policy implementation by the World Bank in particular, because it publishes quantitative performance scores alongside the evaluation reports (Denizer et al., 2013; Dreher et al., 2013; Honig et al., 2022). There is a widespread impression of evaluation as a "gold standard" for assessing organizational performance at the international level (Lall, 2017, p. 246). But there are also studies highlighting the possibility that evaluation findings may be biased (Eckhard & Jankauskas, 2020; Jankauskas & Eckhard, 2023; Malik & Stone, 2018; Raimondo, 2018).

Studies that scrutinize evaluations and test for political biases at the level of reports remain scarce and inconclusive, both for evaluation in domestic and international organizations. At the national level, Vaganay (2016), for example, looks for biases in outcome reporting of government-sponsored policy evaluations, yet analyzes only 13 reports. At the international level, Malik and Stone (2018, 104) show that involvement of multinational corporations in the World Bank's projects leads to 'inflated' evaluation results, caused by "corporate lobbying for disbursements and collusion by Bank staff to influence evaluation". Denizer et al. (2013) use data from 6000 World Bank evaluations to examine correlates of aid-financed project outcomes, acknowledging the possibility that political biases undermine the reliability of results (2013, p. 291). Hence, we still lack comparative empirical research about findings of evaluation reports and their potential biases.

INSTITUTIONAL DESIGN AND EVALUATION BIASES

The dissent between empirical studies that support the functional or the political perspective on evaluation indicates that there may be a neglected confounding factor. Indeed, from an institutionalist perspective, the surrounding organizational context should condition the conduct of evaluation. The core conundrum sustaining the political perspective is that evaluation is supposed to provide *independent* assessment, but it is also a function located *within* the organization. Therefore, scholars have been skeptical of the independence and objectivity of such self-evaluation. In one of the first studies on the topic, Wildavsky even concluded that "evaluation and organization, as it turns out, are to some extent contradictory terms" (Wildavsky, 1972, p. 509). We hypothesize that this is true only under certain configurations of the institutional design.

The argument proceeds along two steps. On the one hand, we know from public choice literature that civil servants working in bureaucracies are generally interested in positive self-portrayal and resource maximization (Dunleavy, 1991; Niskanen, 1994). International administrations, too, have been described as entrepreneurial actors (Knill et al., 2016) who have strong evaluation-related interests that could lead to biases in evaluation outcomes (Gutner & Thompson, 2010; Weaver, 2010). In this regard, positive evaluation findings may serve as a form of "impression management" (O'Brien et al., 2010, p. 441), when administrations use evaluation results for justifying their own initiatives or requesting further funding and support. Negative evaluation results, by contrast, reveal faults in the bureaucratic operations and may lead to a backlash from political principals, legitimacy losses, cuts in funding,

or tensions within the organization. Indeed, as reported above, surveys among evaluators have shown that "being pressured to misrepresent findings is a common occurrence in evaluation" (Morris & Clark, 2012, p. 66).

For instance, evaluated units may be interested in evaluation reports containing positive assessments about their work, such as the following one taken from an evaluation of the UN WOMEN's Tanzania portfolio: "the management structure of the CO [country office] was found efficient and suitable" (UN WOMEN, 2016, p. 10). By contrast, evaluated officials may seek to avoid critical assessments, such as the following one from an evaluation on the Food and Agricultural Organization's (FAO) strategy on nutrition, stating that FAO "failed to address operational issues in sufficient depth and lacked an accountability framework …" (FAO, 2019b, p. 2).

On the other hand, an organization's institutional design may shield evaluation processes from the undue influence of those being evaluated. Evaluation research is by definition supposed to be independent and impartial. As a result, evaluations are typically assigned to independent experts who can be sourced from within or outside the organization, or a combination of both. The process starts with the publication of terms of references, which serve as guidelines for the evaluation. External experts, such as companies or academic institutions, apply for the evaluation. Once a team is formed, the evaluation commences with the creation of an inception report that outlines the methodology and data involved. The commissioning unit and other stakeholders provide feedback, which then leads to the actual evaluation research being conducted. A draft report is then prepared, which is subsequently discussed and reviewed. Finally, after incorporating feedback, the final report is published.

However, in practice, organizations vary in the specific design of their evaluation processes, which can grant the units being evaluated varying degrees of influence over the evaluation research and its outcomes. We propose that within IOs, there are two significant dimensions in the institutional design of evaluation systems that determine the proximity between the evaluand and the evaluators (see Figure 1). The closer this proximity is, the greater the potential for administrative influence on the evaluation.

The first dimension refers to the extent to which the political principal—the member state governing board in the context of IOs—exercises control over an organization's evaluation system (see Eckhard & Jankauskas, 2019). These systems consist of an evaluation policy and a centralized evaluation unit. The UN-wide evaluation norms and guidelines formally specify that evaluation units must remain impartial and conduct evaluations "free from undue pressure" (UNEG, 2016, p. 11). Yet, these units inevitably operate under a chain of command in terms of budgetary, personnel and agenda decisions. As one striking feature, IOs differ in terms of who approves such evaluation system resources: In some IOs, it is the member states' collective in the IO governing body who approves the evaluation unit's budget, staff (the head of unit) and agenda (what is to be evaluated). In some IOs, these decisions are taken jointly with the IO administration (management), and in other IOs senior management decides alone. These differences are defined in each organization's evaluation policy (excerpts can be found in Appendix I).

Such institutional control over the evaluation system could enable the administration to realize its strategic evaluation-related interests. The reason is that evaluation units may lean towards that principal (stakeholder) who possesses more resources for sanctioning, rewarding, or controlling them (Eckhard & Jankauskas, 2020; Leeuw & Furubo, 2008, p. 166; O'Brien et al., 2010, p. 432).

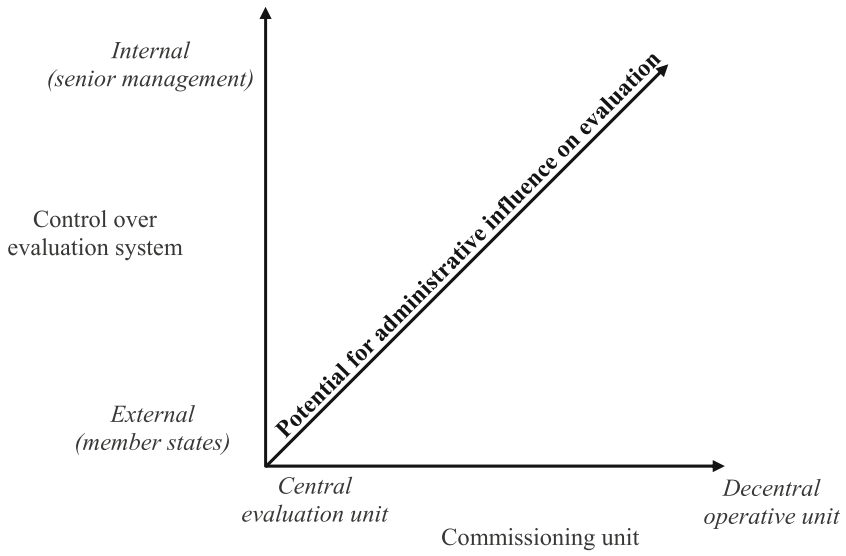

FIGURE 1 Two dimensions of evaluation systems' institutional design in IOs.

Accordingly, if IO administrations maintain their control over evaluation systems, they can signal their bureaucratic interests to evaluation units, thereby explicitly or implicitly pressuring evaluators to present the administration more positively, to ignore more negative findings, to prevent publication of critical reports or to evaluate activities which are known to be functioning well. If, however, member states control IO evaluation resources, they are more involved in the process, and it is harder for administrations to realize evaluation-related interests.

The second institutional design dimension is where the commissioning unit is located in the hierarchy of the IO. Following Wildavsky's (1972) theory on self-evaluating organizations, when there is limited institutional distance between the evaluator and the evaluand, conflicts of interest are more likely to arise. In IOs, evaluations can be commissioned and managed by either the central evaluation unit—or by the IO's operative units that are responsible for policy implementation (see JIU, 2014). When the central evaluation unit commissions an evaluation, its professional and structural independence creates a greater distance between the evaluand and the evaluation team. Conversely, self-evaluation is at its highest when operative units themselves commission evaluations, such as a country office or a substantive unit at headquarters. In such cases, central evaluation units usually only provide remote oversight and quality assurance by commenting on the terms of reference, inception report, and draft final report. Commissioning by operative units minimizes the institutional distance between evaluators and the evaluand and accordingly exposes evaluation reports to the risk of influence by evaluated administrative units.

Overall, we argue that the ability of the evaluated administration to influence evaluation results and make them seem more positive depends on the control of evaluation system resources and the location of the commissioning unit (see Figure 1 above). As explained in more detail below, we can measure the positivity of evaluation findings by calculating the share of positive versus negative assessments about the evaluated activity at the sentence level in an evaluation report (dependent variable). The three hypotheses below summarize our expectations. In the analysis, we also consider a number of potentially confounding factors, such as the activity that is being evaluated.

> **$H_{1.1}$:** Evaluation reports conducted in evaluation systems controlled by IO administrations contain a higher share of positive assessments than those conducted under member states' control.

> **$H_{1.2}$:** Evaluation reports commissioned by decentral operative units contain a higher share of positive assessments than those commissioned by central evaluation units.

> **$H_{1.3}$:** Evaluation reports conducted in evaluation systems controlled by IO administrations and commissioned by decentral operative units contain the highest share of positive assessments.

EMPIRICAL STRATEGY AND MEASUREMENT

We focus on international organizations of the UN system. As in the domestic context, these organizations have seen a proliferation of evaluation in recent years. In the UN system, the number of annual evaluations grew from several dozens in the late 1990s to over 1000 reports since the 2020s.[5] The UN system organizations are similar in terms of their basic structural and management functions. Since they also subscribe to the same system-wide evaluation guidelines of the UN Evaluation Group (UNEG, 2016), they can be treated as a comparable class of cases. Within these cases, we exploit variation in the institutional design of evaluation systems: Who—member states or IO administrations—controls evaluation system resources and who—the central evaluation unit or an operative unit—commissions a given evaluation report.

We estimate the association between these independent variables and the share of positive assessments in evaluation reports. Below, we explain how we measured this variable by compiling a text corpus of evaluation reports and classifying close to one million sentences as containing positive, negative, or neutral assessments of the evaluated activity. Primary focus of the analysis are executive summaries of evaluation reports, as these are the most crucial part of any evaluation (results hold when considering the main text only as well as the whole report). Usually about four to six pages long, they outline key findings of the report. In background conversations with IO officials and member state representatives, we often heard that policymakers are usually unable to read the whole evaluation report (oftentimes several hundred pages long). Thus, well-crafted executive summaries are important for communicating evaluation findings.

Explanatory variables: Institutional design of evaluation systems

Who exercises *control over the evaluation system* is defined by the evaluation policy of each organization. Following previous studies that identify key evaluation resources (Azzam, 2010; Eckhard & Jankauskas, 2020; Rossi et al., 2004), we measure which actor—the member state board or the IO administration/ management—takes decisions on the appointment of the head of the evaluation unit, the evaluation budget, and the evaluation agenda. We use a scoring system by coding for each resource the scope of administrative involvement as low, medium, or high. We then sort every IO into one of three

categories denoting either predominant IO public administration (*IPA*), member states (*MS*), or *mixed* control (see Appendix I).

Regarding the *location of the commissioning unit*, IOs have developed their own 'business models' over time and now vary in their share of centraled versus decentral reports (JIU, 2014). Central evaluation units regularly compile a workplan for specific evaluations to be conducted. Decentral evaluations are being conducted on a more ad hoc manner by the operative IO units or at request by a donor. We hand-coded the commissioning unit dichotomously as indicated on each evaluation report (1 decentral and 0 central). See Appendix I for further details.

Dependent variable: The share of positive sentence assessments in evaluation reports

As the dependent variable, we measure the share of positive versus negative assessments of evaluated activities at the sentence level in reports as a proxy for the positivity of evaluation findings. Descriptive background information classifies as neutral. This operationalization follows Eckhard et al. (2023), who also validate that this procedure enables to extract performance measures on the evaluated activity from text-based reports. Written evaluation reports typically break down a broader evaluation question into sub-categories which are then individually assessed. It is plausible, given this structure, that the more positive or negative assessments can be found in the text, the stronger the signal to a reader that the evaluated activity has been successful, or not. Neutral sentences describing the background of the evaluated activity as well as statements on methods, structure, or proceedings are excluded from the measure. This facilitates the interpretation of the outcome variable as the share of positive assessments compared with the share of negative assessments.

The following example sentences from evaluation reports contain a positive, negative, and neutral assessment, respectively:

- *Positive*: "The project was found to be successful in forming partnerships and building synergies with other organizations…" (ILO, 2018, p. 9).
- *Negative*: "FAO has not systematically used its recognized knowledge of the agricultural/rural sector to build strategic and long-term partnerships with key actors working on gender" (FAO, 2019a, p. 3).
- *Neutral*: "The primary goal of the evaluation is to assess the performance—that is, the activities, outputs, and outcomes—of WWAP [World Water Assessment Programme]…" (UNESCO, 2015, p. 1).

For the classification of sentences, we use the deep learning-based language model published by Eckhard et al. (2023). Such deep learning language models advance previous approaches relying solely on handcrafted features to analyze natural language. The model used here has been developed based on the pre-trained language model BERT (Bidirectional Encoder Representations from Transformers), originally published by Google in 2018. The language model classifies sentences with an accuracy of 89%, providing for each sentence the probability that the prediction is correct.

To summarize, we construct our dependent variable based on the predicted probabilities that a sentence in a report is positive or negative. At the level of reports as the unit of analysis, and excluding neutral sentences, we calculate the *share of positive sentence assessments* in a report. This is a continuous measure reaching from 0 to 1, whereby 1 denotes that 100% of assessments are positive.

Control variables

In our regression models, we account for possible confounders that could affect the institutional design of the evaluation system and the share of positive assessments in a report. One possibility is that the original configuration of the evaluation system design is related to the administration's general autonomy from member states. For instance, IOs that enjoy more delegated powers could be more likely to have evaluation units controlled by the administration, and such IOs are also known to perform better than IOs tightly controlled by member states (Honig, 2020; Sommerer et al., 2022). We thus control for *IO autonomy* (as measured by Lall (2017)).

It is also possible that an *IO's overall performance* across its activities affects the share of positive assessments in a single evaluation (Gutner & Thompson, 2010). We add a performance measure provided by Lall (2017) that is based on the average IO performance ratings provided by national donors such as the British or Australian governments.

We also control for the size of an IO considering their *budgets* and *staff* (logged)[6]: larger IOs might have more evaluation capacities and more complex organizational activities, which could affect evaluation results (JIU, 2014).

Another possibility is that not all IO activities are equally easy or difficult to achieve. The UN system IOs conduct evaluation at the level of projects (single activities), programs (a bundle of projects), institutional processes, and broader organization-wide activities, such as thematic areas or corporate strategies. As an example, studies have shown that specific project-level activities are easier to accomplish than broader, corporate-level IO objectives (Feeny & Vuong, 2017, p. 329). We control for *type of evaluated activity*, which we code manually from evaluation reports (see Appendix I).

Finally, we account for unobserved time variant confounders by including year fixed effects. In some models, we additionally include IO fixed effects to control for unit specific heterogeneity.

Case selection and compilation of the text corpus

The UN system's Evaluation Group's member organizations have jointly published several thousand evaluation reports by 2020.[7] Their database however does not contain all reports, which is why we hand-collected from individual organization webpages or sent out written requests to get access. This process is labor intensive, which implies that we had to select a sample of organizations.

For the selection, we considered the control of the evaluation system variable, which was feasible to obtain before manually collecting evaluation reports. Measuring all UNEG organizations, we found that control in eight IOs was predominantly exercised by member states, in eight by the administration and five IOs revealed mixed control (see Appendix I). We then selected three organizations from each group, choosing those who had published the largest number of evaluation reports. The resulting sample of *nine organizations* includes ILO, UNDP, UNICEF (member state control); FAO, UNESCO, WHO (mixed control); and IOM, UNHCR, UN WOMEN (administrative control). From these nine IOs, we gathered *all available evaluation reports* compiled under the same evaluation policy that we identified to be valid in 2020 (see Appendix II for the whole procedure). The final text corpus includes a total of *1082 evaluation reports* published from 2012 to 2020 (Figure 2). After compiling the dataset, we hand-coded the second independent variable: We found that seven IOs publish both central and decentral reports, but there are also two organizations where all reports are commissioned by the central evaluation unit (UNESCO and FAO, both mixed control). Below, we discuss resulting limitations for the analysis.

Raw text scraped from the PDFs was cleaned by applying standard procedures of natural language processing (e.g., the removal of special characters and numbers) and split into sentences. We then used the above language model to classify all 995,743 sentences as containing a positive or negative assessment (or descriptive text). For more details, see Appendix II.

ANALYSIS

We test whether variation in the institutional design of evaluation systems is associated with differences in the share of positive assessments in evaluation reports. We begin with descriptive evidence, present the regression models, and offer robustness tests. The focus of the analysis is on executive summaries, but we also present results for the main body of the reports.

Descriptive evidence

Figure 3a depicts the average share of positive assessments in the executive summary, grouped by the two independent variables. Every dot in the graphic shows the share of positive assessments in an executive summary. Variation between reports is large. The most negative executive summaries contain a share of close to 5% of positive assessments and the most positive summaries 100%. The difference in the positivity of reports between IPA and MS groups is small ($H_{1.1}$): Executive summaries under administrative control contain 54% positive assessments on average compared to 51% in the member state control group (see also Table A III.4 in Appendix III).

By contrast, the differences between central and decentral reports are more pronounced ($H_{1.2}$). Figure 3b shows that executive summaries in reports conducted by decentral operative units contain on average 55% of positive assessments, whereas reports conducted by central evaluation units contain 46% of positive assessments.

Figure 4 shows the share of positive assessments in an executive summary per control group and commissioning unit. Reports with most positive summaries are commissioned by a decentral operative unit in an organization where the evaluation system is under administrative control, which points towards our theoretical expectations ($H_{1.3}$).

As shown in Figure A III.1, results are comparable when considering the main text of reports. To test our hypotheses more rigorously and exploit the within organization variation of commissioning unit over time, we turn towards regression analysis in the next section.

Regression models

For the regression analysis, we focus on the association between the operative commissioning unit and positivity of executive summaries. We discuss differences in evaluation system control alongside as only the data concerning the commissioning unit follow a panel structure that allows to control for unobserved IO specific and time variant confounders. We include a set of additional control variables to account for confounding factors, as discussed above.

We run several ordinary least squares (OLS) models at a unit of analysis of executive summary in an evaluation report (Table 1). All models in the main table include standard errors clustered by IO (see Table A III.5 in Appendix III for all coefficients).

Model 1 shows the bivariate relation of decentral evaluation report with the share of positive assessments in an executive summary as the dependent variable. In model 2, we include both year and IO fixed effects (FE) to control for unobserved time variant and time invariant unit specific heterogeneity. The point estimate remains 0.093 and is statistically significant at a p-value below 0.01. The model suggests that the executive summaries in evaluation reports are 9 percentage points more positive if they were commissioned by an IO's operative unit (decentral report).

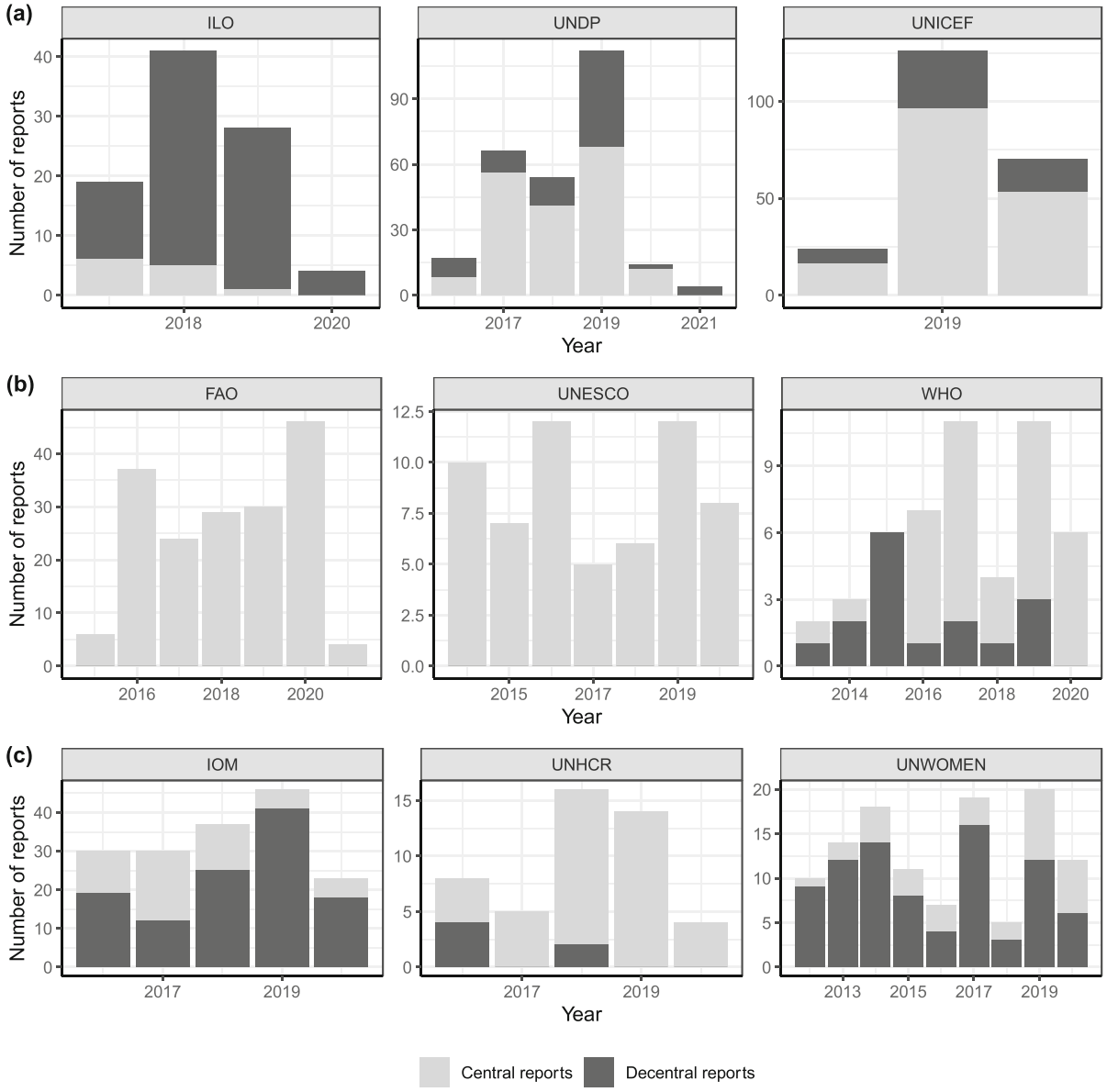

FIGURE 2 The evaluation report text corpus showing the number of central and decentral reports per IO. **Group a** shows organizations where evaluation systems are predominantly controlled by **member states**, **group b** shows **mixed** cases and **group c** cases with **administrative** control over evaluation systems.

Figure 5 plots the predicted means of the share of positive assessments in an executive summary based on model 2. The bars show the 95% prediction intervals around the predicted mean. In line with Table 1, the figure shows that a change in the commissioning unit from decentral (left hand side) to central shifts the predicted mean in the share of positive assessments in an executive summary from 58% to 49%.

To account for differences in control over evaluation systems, and considering alternative explanations, we include the variable regarding MS, mixed or IPA control as well as several control variables in model 3 in Table 1. Controls account for IO performance, de facto IO autonomy, an IO's size (in terms of budget and staff), as well as the type of evaluated activity (project, program, institutional, or thematic). We drop IO fixed effects, as the indicators for IO performance and autonomy are constant per organization. The estimate for decentral reports remains statistically significant. Differences in evaluation system control are not significantly distinct from each other and are not associated with the positivity of an executive summary. Thus, model 3 adds to the descriptive insight that who controls evaluation system resources does not systematically affect evaluation results in executive summaries ($H_{1.1}$).

In contrast, models 2 and 3 provide strong evidence for the positive association between the location of the commission unit and positivity of evaluation reports

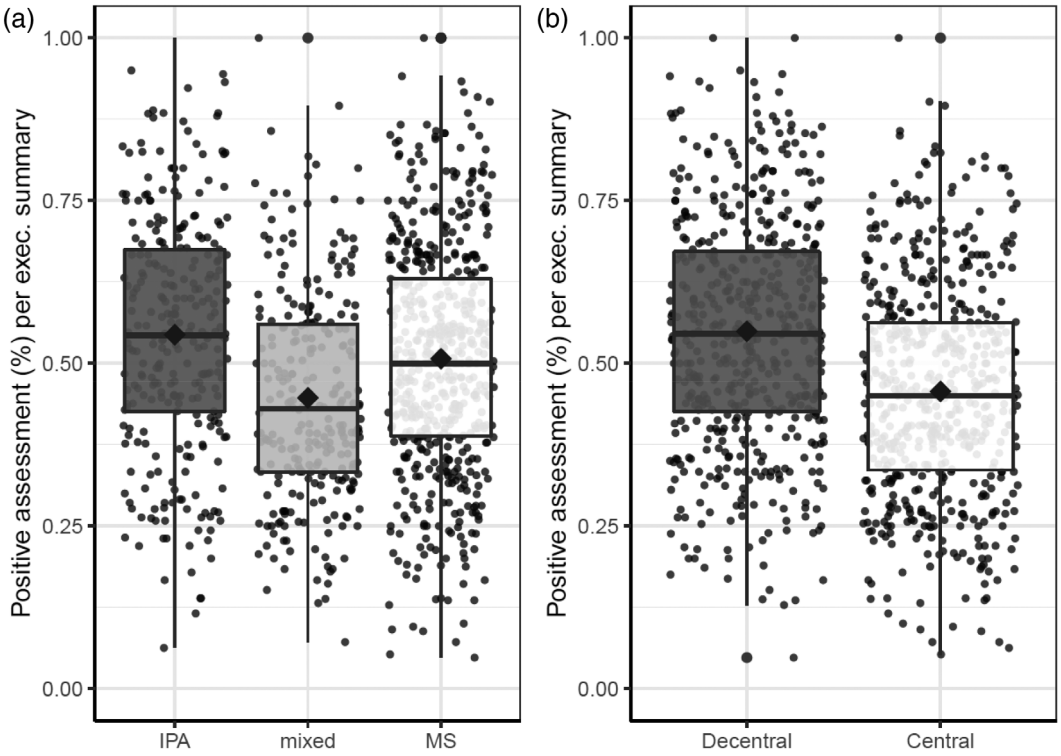

FIGURE 3 Average share of positive assessments in an executive summary. Figure 3a differentiates between control over IO evaluation systems, whereas Figure 3b differentiates between commissioning from decentral or central evaluation units. Horizontal black lines indicate the median, rhombuses the mean.

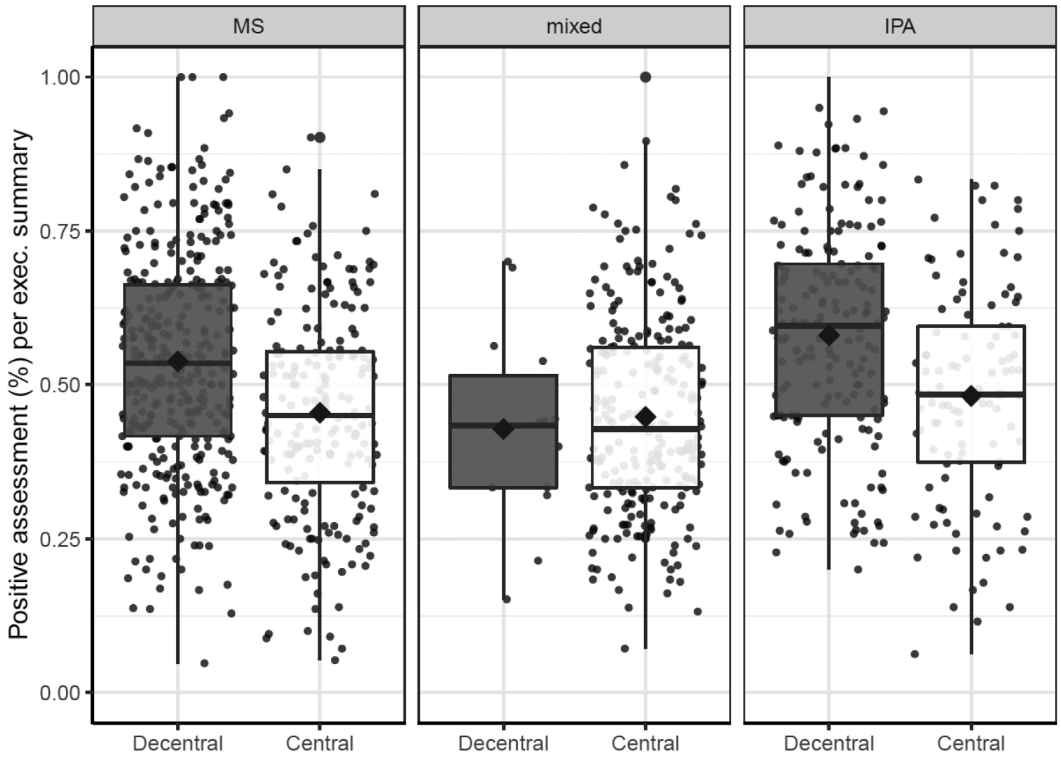

FIGURE 4 Average share of positive assessments in an executive summary per control group and commissioning unit. Horizontal black lines indicate the median, rhombuses the mean.

TABLE 1 OLS regression models.

	Dependent variable Share of positive assessments			
	(1)	**(2)**	**(3)**	**(4)**
Decentral report	0.092***	0.090***	0.049***	0.070***
	(0.018)	(0.023)	(0.018)	(0.004)
Mixed control			−0.028*	−0.023
			(0.016)	(0.017)
IPA control			−0.030	0.001
			(0.038)	(0.036)
Mixed control × decentral				−0.006
				(0.033)
IPA control × decentral				−0.062**
				(0.025)
Constant	0.457***	0.491***	2.098**	2.284**
	(0.017)	(0.014)	(0.960)	(0.975)
Year FE	No	Yes	Yes	Yes
IO FE	No	Yes	No	No
Controls	No	No	Yes	Yes
Observations	1082	1082	1075	1075
Adjusted R^2	0.065	0.118	0.156	0.159
F Statistic	76.261*** (df = 1; 1080)	9.015*** (df = 18; 1063)	11.458*** (df = 19; 1055)	10.635*** (df = 21; 1053)

Note: Standard errors are clustered by IO. Complete results with all variables in Appendix III, Table A III.5.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

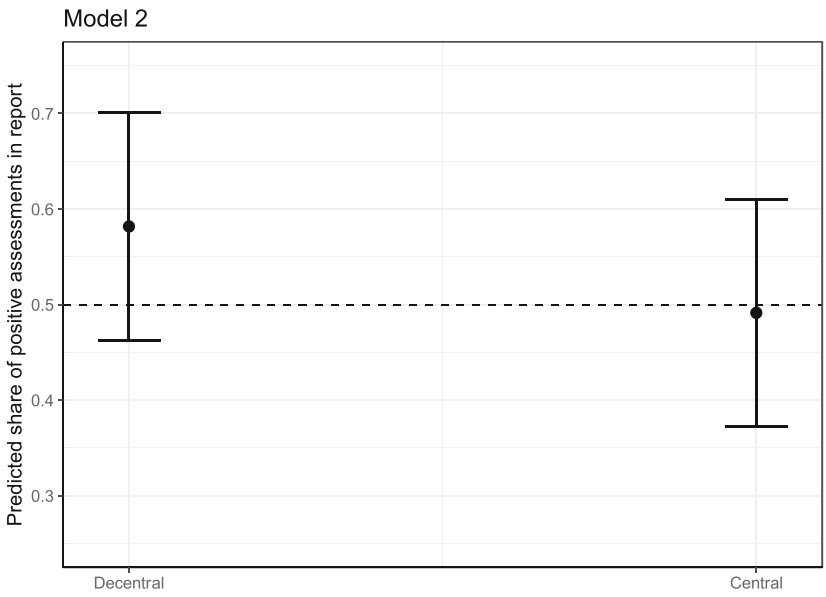

FIGURE 5 Prediction plot for model 2 from Table 1 with 95% prediction intervals around the prediction.

($H_{1.2}$). Executive summaries tend to be about two pages long and contain on average 60 assessment sentences. Thus, substantially, our models predict that an evaluation report commissioned by an IO's operative unit contains on average five to six positive assessment sentences more than a central report by the independent evaluation unit.

We move on to the interaction between operative commissioning unit and administrative control over the evaluation system in model 4 (see $H_{1.3}$). While the coefficient denoting this interaction is statistically significant at a p-value below 0.05 and results point in the expected direction, effect sizes are small after inclusion of controls

(see also Figure A III.2 in the Appendix). Further robustness tests described below show that these results do not hold across different model specifications.

Overall, the findings from Table 1 and descriptive insights in the section above suggest that differences in who controls an IO's evaluation system affects the reporting of positive assessments to limited extent. Thus, the results do not provide evidence for Hypothesis 1.1. Reports that are commissioned by an IO's operative unit itself, however, are significantly more positive, allowing us to maintain Hypothesis 1.2. We further reject Hypothesis 1.3 as the analysis provides only indicative but not robust support for the interaction. In the following, we probe these findings with a series of robustness tests.

Robustness tests

We run a series of additional models at the report as well as at the sentence level. Results remain highly comparable and are displayed in Appendix III. We first check whether our results hold for the main text of an evaluation report (excluding the executive summary) as well as reports as a whole (Tables A III.6 and A III.7).

Next, we turn towards the explanatory variables. We re-run our analysis with a continuous indicator capturing the control over evaluation systems (Table A III.8, see Appendix I for information on the continuous measure). Table A III.9 displays the results when including the percentage of staff employed outside an organization's headquarters to account for the degree of decentralization of an IO that could affect whether a report is decentral. In addition, we show that results are robust when excluding FAO and UNESCO that have only published central reports within our time of observation (Table A III.10).

Further, we test an alteration to our outcome and include neutral sentences. The dependent variable changes to a measure that captures the positivity of a report compared to *all* other remaining sentences. We re-run our main analysis with the share of positive assessments compared with negative and neutral sentences (Table A III.11) and negative compared to positive and neutral sentences (Table A III.12). In addition, we replicate our main analysis including evaluation reports that were identified to have overly long or short executive summaries or main texts (Table A III.13).

Next, we re-run all discussed model specifications at the sentence level in Tables A III.14 to A III.18 to include weights for the language model's classification accuracy per sentence. In these models, the dependent variable is a bivariate measure indicating whether a sentence is positive (1) or negative (0).

Last, we address that our outcome measure treats all sentences as equally important. We check that results are robust when including the relative order in which s sentence appears to account that sentences at the very beginning of an executive summary might be more important to the reader than other sentences (Table A III.19).

DISCUSSION

In the analysis, we scrutinized whether differences in the institutional design of IO evaluation systems are systematically associated with differences in the share of positive assessment sentences across 1082 IO evaluation reports. On the one hand, our findings show no evidence that the *control structure* of IO evaluation systems (IPA vs. MS control) is associated to such biases. Descriptive evidence shows only small differences between IOs in which either member states or the administration controls evaluation system resources, these differences turn out insignificant in the regression analysis. We therefore reject Hypothesis 1.1.

On the other hand, however, we do find evidence that reports commissioned by decentral operative administrative entities are systematically more positive than those commissioned by central evaluation units. For executive summaries, that is, sections that decision-makers are most likely to read, the predicted difference is more pronounced (9 percentage points) than differences in the main text (4.8 percentage points). These results are robust to a series of additional tests. Therefore, we hold that the empirical findings support Hypothesis 1.2 related to the commissioning evaluation unit. But we cannot confirm Hypothesis 1.3 postulating an interaction between the two independent variables.

As a limitation, our findings apply only to specific operationalizations of the two explanatory variables, that is, control over evaluation system and the commissioning unit of an evaluation report. It is possible that additional political dynamics are at play, such as individual factors linked to the evaluation team drafting a report. This analysis was unable to consider such factors. Furthermore, the measurement of the dependent variable also reveals a limitation: Whereas the measure treats all sentences equally, it is possible that certain sentences are more salient than others. Although we considered the order of sentences as a robustness check, there may be other differences that only a qualitative in-depth analysis could capture. As another limitation, we scrutinized the political perspective on evaluation only from one possible perspective, namely that evaluated units may seek to have their performance portrayed in the most positive light. In addition to this, other substantive political interests may also manifest in evaluation reports.

Regarding generalization, this article focused on evaluation in the context of international public administrations. These are public service organizations operating at the international level, guided by Weberian principles of bureaucracy (Bauer et al., 2017; Eckhard & Ege, 2016; Stone & Moloney, 2019; Thorvaldsdottir et al., 2021). Furthermore, evaluation principles and guidelines (i.e., the OECD-DAC criteria) applied in the UN system also guide

the conduct of evaluation elsewhere. And last but not least, pressure by evaluated units on evaluators has been frequently documented for the domestic level, too (Azzam, 2010; Bjornholt & Larsen, 2014; Morris & Clark, 2012). It is therefore plausible that our findings generalize to other types of public service organizations. But there are also relevant differences between domestic and international bureaucracies. Most importantly, the political principal of an international public administrations comprises a collective of member states, which potentially causes idiosyncratic political dynamics between the principal and agent, which have been well documented by a rich literature in international relations (Hawkins et al., 2006; Jankauskas, 2022; Trondal et al., 2010). It is possible that this leads to higher degrees of politicization of IO activities, including evaluation, or more pronounced administrative interests in the evaluation results. Further research is therefore needed to explore whether our results replicate beyond the international level, and how exactly the ideal institutional distance between evaluator and evaluand manifests in a domestic context.

CONCLUSION

This article set out to address a missing puzzle in the discussion that emerged from Wildavsky's (Wildavsky, 1972, p. 509) longstanding argument that evaluation and organization are contradictory terms. This literature on the politics of evaluation put forth interview and survey data, showing that evaluators are frequently confronted with undue pressure from members of an organization to present evaluation results more positively. We expand this discussion by arguing that the institutional design of an organization conditions whether evaluation outcomes are functional or politicized. And we break new ground by offering the first comparative study on evaluation politics with data gained from a text analysis of evaluation reports.

Based on the operationalization and analysis presented above, we show that whether the member states board or the IO administration controls the evaluation system is not associated with systematic differences in the positivity of reports' findings. But we do find that evaluation reports commissioned by an IO's operative units are on average more positive than evaluation reports commissioned by central evaluation units. These remaining differences cannot be accounted for by a range of plausible factors, such as the general performance of an organization or the type of the evaluated activity.

We therefore conclude that the increased positivity of decentral evaluation reports emerges due to a lack of independence of evaluators, who are either pressured to present findings more positively, or do so in anticipation of the interests of commissioning units—sometimes maybe even unconsciously. Both these dynamics have been identified from survey and interview-based studies (Azzam, 2010; Morris & Clark, 2012). Our study therefore contributes to the literature by revising Wildavsky's initial claim: *Evaluation and organization are contradictory terms, but only in the absence of institutional safeguards that enhance the distance between evaluator and those evaluated*.

Our findings also speak to ongoing disputes between political and 'pragmatic' (functional) perspectives on evidence-based policymaking (Cairney, 2016; see MacKillop & Downe, 2022; Sanderson, 2002) by revealing the mechanisms that may lead to evaluation as evidence being perceived in either way. An overly positive portrayal of evaluated activities may undermine trust in evaluation results. And it may undermine the ability of policymakers to gain a realistic understanding of policy impact and administrative performance, which are necessary conditions for organizational and policy learning.

For international public administrations, we found that central evaluation units successfully protect the independence of evaluation, even if IO management takes key budgetary, staff and agenda setting decisions on the evaluation system (rather than member states). This implies that evaluation units can be de facto independent, even if they are de jure part of the organization they assess.

This finding should also be relevant for other public organizations which consider where to allocate the evaluation function in their institutional architecture. This may even include domestic bureaucracies. Studies on evaluation in the domestic context point to the presence of outside lobbying or political pressure on evaluators, which may be reduced or avoided by reducing the distance between the evaluators and those evaluated, for instance, through the installation of a central evaluation unit.

ACKNOWLEDGMENTS

We thank the following individuals who contributed to the success of this research. For their excellent research support at various stages of this project, we would like to thank Ian Burton, Svana Burger, Alina Becker, Marisa Brecht, Roberto Daniele Cadili, Tom Milson, and Phillip Rothe. We presented this paper at various venues and would like to thank Estelle Raimondo, Nicholas Charron, Eva Thomann, Ronny Patz, Mirko Heinzel, Tom Biersteker, Laurin Friedrich, Nicolai Dose, and Catherine Weaver for providing comments, as well as all attendants of the Public Administration and Public Policy seminar series in Gothenburg and the Management and Public Administration seminar series at the University of Konstanz and Zeppelin University. The German Research Foundation (DFG) provided funding for this project (grant no. EC 506/2-1), which we gratefully acknowledge. Last but not least, we thank the editors and three anonymous reviewers at PAR for their helpful and constructive comments during the review process. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Steffen Eckhard https://orcid.org/0000-0002-5320-0730
Elena Leuschner https://orcid.org/0000-0003-1364-473X

ENDNOTES

[1] For instance, the United States government passed the Evidence-Based Policymaking Act in 2018, https://www.congress.gov/115/plaws/publ435/PLAW-115publ435.pdf; the German government installed a National Institute for Evaluation (Deval) in 2012 to offer evaluation services in the field of development assistance (https://www.deval.org/en/about-us/the-institute/goals-and-functions); and the European Commission adopted the Better Regulation toolbox in 2017, https://ec.europa.eu/info/sites/info/files/better-regulation-toolbox_2.pdf.

[2] For an overview, see https://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm.

[3] For an overview of UN system organizations, see https://www.un.org/en/about-us/un-system.

[4] The nine organizations are the International Labor Organization (ILO), the UN Development Program (UNDP), the UN International Children's Emergency Fund (UNICEF), the Food and Agricultural Organization (FAO), the UN Educational, Scientific and Cultural Organization (UNESCO), the World Health Organization (WHO), the International Organization for Migration (IOM), the UN High Commissioner for Refugees (UNHCR) and the UN Entity for Gender Equality and the Empowerment of Women (UN WOMEN).

[5] As per number of published evaluation reports in the database of the UN Evaluation Group (http://www.uneval.org/evaluation/reports).

[6] Data retrieved from UN System Chief Executives Board for Coordination: https://unsceb.org/.

[7] See previous footnote 5.

REFERENCES

Anderson, James E. 1975. *Public Policy-Making*. New York: Praeger.

Azzam, Tarek. 2010. "Evaluator Responsiveness to Stakeholders." *American Journal of Evaluation* 31(1): 45–65.

Bauer, Michael W., Christoph Knill, and Steffen Eckhard, eds. 2017. *International Bureaucracy: Challenges and Lessons for Public Administration Research*. London: Palgrave Macmillan UK.

Bjornholt, Bente, and Flemming Larsen. 2014. "The Politics of Performance Measurement: 'Evaluation Use as Mediator for Politics'." *Evaluation* 20(4): 400–411.

Boyne, George A., Julian S. Gould-Williams, Jennifer Law, and Richard M. Walker. 2004. "Toward the Self-Evaluating Organization? An Empirical Test of the Wildavsky Model." *Public Administration Review* 64(4): 463–473.

Cairney, Paul. 2016. *The Politics of Evidence-Based Policy Making*. London: Palgrave Macmillan.

Denizer, Cevdet, Daniel Kaufmann, and Aart Kraay. 2013. "Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance." *Journal of Development Economics* 105: 288–302.

Dreher, Axel, Stephan Klasen, James R. Vreeland, and Eric Werker. 2013. "The Costs of Favoritism: Is Politically Driven Aid less Effective?" *Economic Development and Cultural Change* 62(1): 157–191.

Dunleavy, Patrick. 1991. *Democracy, Bureaucracy and Public Choice: Economic Explanations in Political Science*. London: Harvester Wheatsheaf.

Eckhard, Steffen, and Jörn Ege. 2016. "International Bureaucracies and their Influence on Policy-Making: A Review of Empirical Evidence." *Journal of European Public Policy* 23(7): 960–978.

Eckhard, Steffen, and Vytautas Jankauskas. 2019. "The Politics of Evaluation in International Organizations: A Comparative Study of Stakeholder Influence Potential." *Evaluation* 25(1): 62–79.

Eckhard, Steffen, and Vytautas Jankauskas. 2020. "Explaining the Political Use of Evaluation in International Organizations." *Policy Sciences* 53(4): 667–695.

Eckhard, Steffen, Vytautas Jankauskas, Elena Leuschner, Ian Burton, Tilmann Kerl, and Rita Sevastjanova. 2023. "The Performance of International Organizations: A New Measure and Dataset Based on Computational Text Analysis of Evaluation Reports." *Review of International Organizations* 1–24. https://doi.org/10.1007/s11558-023-09489-1.

FAO. 2019a. "Evaluation of FAO's Work on Gender: Office of Evaluation. Rome: WFP. https://www.fao.org/publications/card/en/c/CA3755EN

FAO. 2019b. Evaluation of the Strategy and Vision for FAO's Work in Nutrition. Rome: WFP. https://www.fao.org/documents/card/en/c/CA3762EN/.

Feeny, Simon, and Vu Vuong. 2017. "Explaining Aid Project and Program Success: Findings from Asian Development Bank Interventions." *World Development* 90: 329–343.

Gutner, Tamar, and Alexander Thompson. 2010. "The Politics of IO Performance: A Framework." *The Review of International Organizations* 5(3): 227–248.

Hawkins, Darren G., David A. Lake, Daniel L. Nielson, and Michael J. Tierney, eds. 2006. *Delegation and Agency in International Organizations*. Cambridge: Cambridge University Press.

Honig, Dan. 2020. "Information, Power, and Location: World Bank Staff Decentralization and Aid Project Success." *Governance* 33(4): 749–769.

Honig, Dan, Ranjit Lall, and Bradley C. Parks. 2022. "When Does Transparency Improve Institutional Performance? Evidence from 20,000 Projects in 183 Countries." *American Journal of Political Science*: 1–21. https://doi.org/10.1111/ajps.12698.

ILO. 2018. "Promoting Worker Rights and Competitiveness in Egyptian Export Industries: EGY/11/06/USA." https://www.ilo.org/eval/Evaluationreports/WCMS_445897/lang--en/index.htm.

Jankauskas, Vytautas. 2022. "Delegation and Stewardship in International Organizations." *Journal of European Public Policy* 29(4): 568–588.

Jankauskas, Vytautas, and Steffen Eckhard. 2023. *The Politics of Evaluation in International Organizations*. Oxford: Oxford University Press.

JIU. 2014. "Analysis of the Evaluation Function in the United Nations System: United Nations Joint Inspection Unit." JIU/REP/2014/6.

Knill, Christoph, Steffen Eckhard, and Stephan Grohs. 2016. "Administrative Styles in the European Commission and the OSCE-Secretariat: Striking Similarities despite Different Organisational Settings." *Journal of European Public Policy* 23(7): 1057–76.

Lall, Ranjit. 2017. "Beyond Institutional Design: Explaining the Performance of International Organizations." *International Organization* 71(2): 245–280.

Lee, Barbara. 2006. "Theories of Evaluation." In *Evaluationsforschung: Grundlagen Und Ausgewählte Forschungsfelder*, 3rd ed., edited by Reinhard Stockmann, 137–176. Muenster: Waxmann Verlag.

Leeuw, Frans L., and Jan-Eric Furubo. 2008. "Evaluation Systems: What Are they and why Study Them?" *Evaluation* 14(2): 157–169.

MacKillop, Eleanor, and James Downe. 2022. "What Counts as Evidence for Policy? An Analysis of Policy Actors' Perceptions." *Public Administration Review* Early view. https://doi.org/10.1111/puar.13567.

Malik, Rabia, and Randall W. Stone. 2018. "Corporate Influence in World Bank Lending." *The Journal of Politics* 80(1): 103–118.

Morris, Michael, and Brittany Clark. 2012. "You Want me to Do What? Evaluators and the Pressure to Misrepresent Findings." *American Journal of Evaluation* 34(1): 57–70.

Niskanen, William A. 1994. *Bureaucracy and Public Economics*, 2nd ed. Aldershot, Hants: Elgar.

O'Brien, Terri, Sheila Payne, Mike Nolan, and Christine Ingleton. 2010. "Unpacking the Politics of Evaluation: A Dramaturgical Analysis." *Evaluation* 16(4): 431–444.

Pleger, Lyn, Fritz Sager, Michael Morris, Wolfgang Meyer, and Reinhard Stockmann. 2017. "Are some Countries more Prone to Pressure Evaluators than Others? Comparing Findings from the

United States, United Kingdom, Germany, and Switzerland." *American Journal of Evaluation* 38(3): 315–328.
Raimondo, Estelle. 2018. "The Power and Dysfunctions of Evaluation Systems in International Organizations." *Evaluation* 24(1): 26–41.
Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman, eds. 2004. *Evaluation: A Systematic Approach*, 7th ed. Thousand Oaks, CA: Sage Publication.
Sanderson, Ian. 2002. "Making Sense of 'What Works': Evidence Based Policy Making as Instrumental Rationality?" *Public Policy and Administration* 17(3): 61–75.
Sommerer, Thomas, Theresa Squatrito, Jonas Tallberg, and Magnus Lundgren. 2022. "Decision-Making in International Organizations: Institutional Design and Performance." *Review of International Organizations* 17: 815–845.
Stone, Diane, and Kim Moloney, eds. 2019. *The Oxford Handbook of Global Policy and Transnational Administration*. Oxford, United Kingdom: Oxford University Press.
Taylor, David, and Susan Balloch, eds. 2005. *The Politics of Evaluation: Participation and Policy Implementation*. Bristol: The Policy Press.
The LSE GV314 Group. 2014. "Evaluation under Contract: Government Pressure and the Production of Policy Research." *Public Administration* 92(1): 224–239.
Thorvaldsdottir, Svanhildur, Ronny Patz, and Steffen Eckhard. 2021. "International Bureaucracy and the United Nations System." *International Review of Administrative Sciences* 87(4): 695–700.
Trondal, Jarle, Martin Marcussen, Torbjorn Larsson, and Frode Veggeland. 2010. *Unpacking International Organisations: The Dynamics of Compound Bureaucracies*. Manchester: Manchester University Press.
UN WOMEN. 2016. "Country Portfolio Evaluation: Tanzania, Strategic Note 2014–2016." Final Report.
UNEG. 2016. *Norms and Standards for Evaluation*. New York: United Nations Evaluation Group.
UNESCO. 2015. "Evaluation of the World Water Assessment Programme: IOS/EVS/PI/142 REV." Paris: UNSECO. https://unesdoc.unesco.org/ark:/48223/pf0000234429.
Vaganay, Arnaud. 2016. "Outcome Reporting Bias in Government-Sponsored Policy Evaluations: A Qualitative Content Analysis of 13 Studies." *PLoS One* 11(9): e0163702.
van Voorst, Stijn, and Ellen Mastenbroek. 2019. "Evaluations as a Decent Knowledge Base? Describing and Explaining the Quality of the European Commission's ex-Post Legislative Evaluations." *Policy Sciences* 52(4): 625–644.
Weaver, Catherine. 2010. "The Politics of Performance Evaluation: Independent Evaluation at the International Monetary Fund." *Review of International Organizations* 5(3): 365–385.
Wergin, John F. 1976. "The Evaluation of Organizational Policy Making: A Political Model." *Review of Educational Research* 46(1): 75–115.
Wildavsky, Aaron. 1972. "The Self-Evaluating Organization." *Public Administration Review* 32(5): 509.

AUTHOR BIOGRAPHIES

Steffen Eckhard is a Professor of Public Administration and Public Policy at Zeppelin University. He is also affiliated with the Center of Excellence "The Politics of Inequality" at the University of Konstanz and a Fellow at the Global Public Policy Institute (GPPi) in Berlin.
steffen.eckhard@zu.de

Vytautas Jankauskas is a Post-Doctoral Research Fellow at Zeppelin University and Associate Research Fellow at the University of Munich (LMU).
vytautas.jankauskas@zu.de

Elena Leuschner is a PhD candidate at the Department of Political Science at the University of Gothenburg.
elena.leuschner@gu.se

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Eckhard, Steffen, Vytautas Jankauskas, and Elena Leuschner. 2024. "Institutional Design and Biases in Evaluation Reports by International Organizations." *Public Administration Review* 84(3): 560–573. https://doi.org/10.1111/puar.13705