The Review of International Organizations (2023) 18:753–776
https://doi.org/10.1007/s11558-023-09489-1

DATA ARTICLE

The performance of international organizations: a new measure and dataset based on computational text analysis of evaluation reports

Steffen Eckhard[1] · Vytautas Jankauskas[1] · Elena Leuschner[2] · Ian Burton[1] · Tilman Kerl[3] · Rita Sevastjanova[4]

Accepted: 7 March 2023 / Published online: 6 May 2023
© The Author(s) 2023

Abstract
International organizations (IOs) of the United Nations (UN) system publish around 750 evaluation reports per year, offering insights on their performance across project, program, institutional, and thematic activities. So far, it was not feasible to extract quantitative performance measures from these text-based reports. Using deep learning, this article presents a novel text-based performance metric: We classify individual sentences as containing a negative, positive, or neutral assessment of the evaluated IO activity and then compute the share of positive sentences per report. Content validation yields that the measure adequately reflects the underlying concept of performance; convergent validation finds high correlation with human-provided performance scores by the World Bank; and construct validation shows that our measure has theoretically expected results. Based on this, we present a novel dataset with performance measures for 1,082 evaluated activities implemented by nine UN system IOs and discuss avenues for further research.

Keywords Performance · Evaluation · International organizations · Natural language processing · Machine learning

1 Introduction

How well do international organizations (IOs) perform? The literature on this core question of IO research has made huge progress in recent years, thanks also to the increasing availability of independent evaluations[1] published by both international organizations and

[1] Following the UN's definition, evaluation is "an assessment, conducted as systematically and impartially as possible, of an activity, project, programme, strategy, policy, topic, theme, sector, operational area or institutional performance" UNEG (2016). This systematic assessment is typically published in the form of an evaluation report, often up to several hundred pages long.

Responsible editor: Axel Dreher

✉ Steffen Eckhard
steffen.eckhard@zu.de

Extended author information available on the last page of the article

Konstanzer Online-Publikations-System (KOPS)
URL: http://nbn-resolving.de/urn:nbn:de:bsz:352-2-mz3jdaqm1vf66

national governments. Some of these evaluations summarize the performance of evaluated activities in standardized numeric ratings which has enabled researchers to compile major data collections on the performance of IO activities (e.g., Honig et al. 2022). Nevertheless, the rich information contained in the actual evaluation reports – often several hundred pages long – has not yet informed comparative IO research. What is more, standardized performance ratings exist primarily for single projects and for activities implemented by organizations in the field of multilateral financial aid and development assistance. This implies that evaluations on broader thematic topics or institutional activities, as well as those published by organizations in other policy fields, have not yet been considered for IO performance research. Organizations of the United Nations (UN) system, for example, publish around 750 evaluation reports per year, but the majority of these organizations does not publish a performance score.[2]

Against this backdrop, this article breaks new ground by introducing and validating a method suited to extract performance measures from the text of an evaluation report. This provides researchers with an alternative way to measure the performance of an IO activity for those organizations that publish performance scores, and it offers new insights for other organizations which publish evaluation reports without the scores.

First, we introduce the new *IOEval* dataset that contains sentence-level text of 1,082 evaluation reports from nine major UN-system IOs[3] produced between 2012 and 2021. These evaluations assess four different types of IO activities, namely projects, programs, institutional and thematic activities. These activities take place at the country, regional, or global level.

Second, we introduce a procedure that for the first time enables extracting quantitative performance measures from the text of evaluation reports. Our rationale considers the structure of written reports, which typically break down broader evaluation questions into sub-questions that are then individually assessed. Single positive or negative assessments take place at the level of sentences. We expect that *the more positive assessments a report contains, the more positive the performance of the evaluated activity*.

Given that the dataset contains close to one million sentences, we use deep learning for sentence classification. Based on 10,296 hand-coded sentences from 180 reports selected at random, we fine-tune a pre-trained language model (BERT) that ultimately allows classifying sentences as containing a positive or negative assessment of the evaluated activity (or neutral descriptive text). This procedure enables computing the share of positive assessment sentences per report, our performance measure.

[2] Most organizations of the UN system are members of the United Nations Evaluation Group (UNEG) and share common evaluation guidelines, thus following the same principles and criteria. The UNEG website specifies the number of reports published per year: http://www.uneval.org/evaluation/reports (accessed 22 March 2023).

[3] These 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).

We apply three strategies to demonstrate the validity of the novel performance measure (Adcock & Collier, 2001). First, content validation at the report and sentence level shows that the share of positive sentences in a report adequately reflects the underlying concept of performance. We secondly conduct convergent validation by applying our measurement procedure to 661 evaluation reports by the World Bank which were not in the original training data. In addition to the text, these reports also contain a human-provided performance score. We find a strong positive correlation between our own measure and the scores provided by the World Bank evaluators. Lastly, construct validation finds that our measure yields theoretically expected results: Comparing the performance of projects (narrow goals, short time frame, often at sub-national level) and programs (bundle of projects, long-term goals, often at national or regional level), we confirm that the former perform better than the latter across all IOs.

Overall, this article contributes to the study of international organizations by providing a validated, reliable, replicable, and easily scalable performance measure based on the full text contained in IO evaluation reports. We publish the *IOEval* dataset, which enables comparative research on the performance of 1,082 evaluated project, program, institutional or thematic activities at the country, regional or global level by nine UN-system IOs. This helps expanding the study of IO performance – both its causes and consequences – beyond organizations for which human-provided performance scores already exist. As we also publish the language model's algorithm, anybody can replicate our procedure and expand the *IOEval* dataset in the future.

In the following, we first situate our research in the context of studies on IO performance, and then proceed along the structure set out above, ending with a discussion on limitations and opportunities for future research.

2 IO performance and evaluation

In this section, we define the latent concept of organizational performance, discuss how organizations use evaluation for performance measurement, and report to what extent the existing IO literature has used such performance data.

In its simplest terms, the performance of a public organization has been defined as its ability to achieve pre-defined goals (Heinrich, 2012). The Oxford dictionary[4] defines performance as "[t]he accomplishment or carrying out of something commanded or undertaken". As Lipson (2010: 256) put it, it is about "an organization's use of its resources, technology, and relationships with its organizational environment to achieve collective goals".

Public sector organizations have long attempted to measure their performance "along a set of key indicators" (Poister et al., 2015: 7). But since not all questions of organizational goal accomplishment can be answered at the level of indicators,

[4] Oxford English Dictionary, www.oed.com (accessed 22 March 2023).

organizations also use evaluation to "analyz[e] the performance data" (Poister et al., 2015: 9). Evaluations therefore contain analyses of a broad range of data sources to assess the performance of an organization in a specific activity.

Academic research, too, has attempted to measure IO performance, employing a broad range of concepts and definitions of the term (for an overview, see Gutner & Thompson, 2010). Some studies address performance at the level of institutional design choices, staffing practices or management processes (Graham, 2014; Heinzel, 2021). Others capture performance at the level of outputs, for instance by counting and characterizing IO governing body resolutions (Sommerer et al., 2021; Tallberg et al., 2016). Some IOs, most notably in the field of development assistance, publish performance metrics for single projects which have been used extensively as a reference for IO performance (Heinzel, 2022; Honig, 2020; Honig et al., 2022; Lall, 2017). Lastly, there are also attempts to measure performance at the level of societal outcomes, examining the extent to which IOs contribute to managing the global economy (e.g. Parízek, 2020), fostering good governance (e.g. Honig & Weaver, 2019), or protecting human rights (e.g. Lebovic & Voeten, 2009).

Evaluation, as defined by the UN system (see introduction), can attempt to measure performance at either one of these levels, depending on the exact evaluation research question. The UN Evaluation Group's classification distinguishes four main evaluation types. Evaluation can measure performance of 1) single *projects*; 2) *programs*, which typically contain a bundle of projects; 3) *institutional processes* or activities, such as human resources, public relations, or procurement; and 4) broader *thematic* questions, such as an IO's support to youth or gender equality.[5] When assessing one of these activities, evaluators consider the performance along all or several of the UN Evaluation Group's six evaluation criteria. These are the relevance, coherence, effectiveness, efficiency, impact as well as sustainability of an IO activity (UNEG, 2016: 10).

IO evaluations are typically conducted or supervised by an IO's central evaluation unit. These are established as independent entities within the organization (hence they are sometimes named *independent* evaluation units). Which IO activities will be evaluated is not random, but subject to stakeholder consultations which usually involve IO management and member state representatives in the IO governing bodies (Eckhard & Jankauskas, 2019). The aim is to generate "the most relevant, useful and timely information" about a wide spectrum of IO activities (UNEG, 2016: 21). Sometimes it is also a decentral operative unit that commissions the evaluations, with the central evaluation unit providing guidance and oversight. The actual evaluation research is conducted by a team of experts, either internal or external consultants, or a mix of both. The process of drafting the final evaluation report involves feedback and stakeholder consultation.

Some IOs also publish a standardized performance rating alongside their evaluation reports. Using this data, Honig et al. (2022) published the hitherto most comprehensive dataset covering project ratings for more than 20,000 projects by

[5] Some IOs use additional categories, such as 'impact evaluation' or 'synthesis of evaluations' but these only rare exceptions.

12 bilateral and multilateral aid agencies. Scholars can easily use such standardized performance ratings for their analyses. For instance, Honig et al. (2022) studied the link between transparency and performance for more than 20,000 projects, Denizer et al. (2013) utilized 6,000 World Bank project evaluations to study micro and macro correlates of aid project outcomes, and Bulman et al. (2017) looked into 3,797 World Bank and 1,322 Asian Development Bank projects (see also Buntaine & Parks, 2013; Dreher et al., 2013; Feeny & Vuong, 2017; Geli et al., 2014).

So far, however, mostly organizations in the field of multilateral financial aid and development assistance have published standardized performance ratings. This means that there is still a lack of comparative insights for other policy fields, such as health, humanitarian aid or social policy. In this regard, two key obstacles currently prevent scholars from unleashing the potential that IO evaluation reports offer. First, there is no comprehensive empirical basis upon which scholars could build their analyses, since internal IO evaluation reports are scattered between IOs, timeframes, and evaluation types and levels. Second, given that many IOs do not provide standardized ratings for each evaluation, scholars lack analytical tools to extract performance measures from large numbers of reports. Both reasons explain why thousands of internal evaluations from the UN system remain neglected in IO research.

To be sure, evaluations of IO activities can also be conducted *externally*, such as those produced by donors who seek to scrutinize how and to what end IOs employ their financial contributions. For example, a bilateral donor or multilateral development bank may sponsor a project by an UN agency and may publish the resulting evaluation data. However, such external donor evaluations usually only assess those projects which are funded by or of interest to the evaluating donor. This "eye of the beholder" problem (Gutner & Thompson, 2010: 233) biases the available performance data, considering that it is primarily Western donors who provide the bulk of funding for IO activities.[6] Also, these reports are usually not public.

Therefore, the remainder of this article introduces a novel dataset of internal IO evaluation reports along with a method suited to extract performance measures from their text.

3 The IOEval dataset

The UN Evaluation Group maintains a database with over 20,000 evaluation reports (2023) published by its 21 member organizations. But the database does not store all reports. There are often missing documents or links to reports in other databases that are not accessible. We therefore hand-collected reports, which is labor intensive. Thus, for this first version of the dataset, we limited the selection to reports from nine major UN system IOs: ILO, UNDP, UNICEF, FAO, UNESCO, WHO,

[6] But note that when aggregating project performance scores at an organizational level, donor evaluation results seem to be largely aligned with one another. Individual governmental evaluation scores also correlate with the results of MOPAN surveys (Lall (2017: 260).

IOM, UNHCR, and UN WOMEN.[7] We chose these IOs to gain a diverse set of organizations while ensuring that each published a large number of reports: First, these IOs vary in their policy fields, staff size and constellations, as well as budgetary scale and scope (see Appendix Table A I.1). Second, each IO published between 43 and 244 reports between 2012 and 2021. Third, as members of the UN Evaluation Group, these IOs are subject to the same system-wide evaluation norms and guidelines (UNEG, 2016), making their evaluation reports comparable. To verify that the definition of evaluation and the evaluation criteria indeed matched the UN Evaluation Group standards, we restricted the data collection for each organization to the period for which we could access their evaluation policies.

We proceeded in several rounds to compile the dataset: First, we web-scraped the UN Evaluation Group's data repository, which however yielded missing entries in the data due to broken links and missing PDFs; second, we manually collected additional reports from individual IO websites or requested reports directly from evaluation units (approximately 74% of the whole dataset). All PDF-documents were converted into raw text using Optical Character Recognition. Raw text was cleaned by applying standard procedures of natural language processing (e.g., removal of special characters and numbers) and split into sentences. The final *IOEval* dataset includes a total of 1,082 evaluation reports published from 2012 to 2021 and 995,743 distinct sentences, indicating their order in the original report. For further details on data collection and cleaning see Appendix I.

In addition, the *IOEval* dataset also includes metadata variables at the level of reports: report *title*, *publication date*, *evaluation type* (project, program, institutional or thematic), *evaluation level* (country (specifying its name), regional, global), and *commissioning unit* (centralized or decentralized). At a sentence level, we specify to which *text section* a sentence belongs (executive summary, main text, appendix). See Table A II.1 in the Appendix for further details and examples.

4 Measuring performance based on evaluation reports

To extract a performance measure from the text-based reports, we consider the structure of the text. Each report typically breaks down a broader evaluation question into smaller sub-questions that are being analyzed and assessed, sometimes with a mix of research methods. The structure can also follow the UNEG's six evaluation criteria as introduced above. The main location of individual assessments is the analytical or findings section of a report, where the different sub-questions are raised, discussed and answered. In addition, summaries of the main findings are provided in the executive summary, introduction, recommendation section, and the conclusion. It is plausible to expect that the more positive or negative assessments

[7] Without selection, subsequent data processing would not be feasible due to a huge amount of data generated.

a report contains, the more positive or negative should the overall judgement about the evaluated activity be.

This logic can be further extended to the level of sentences. Evaluation reports in our dataset contain on average 850 sentences. Around half of these sentences contain no assessments. They are descriptive and provide information on the evaluation's background, structure or methodology. The other half of the sentences contain assessments, i.e., either positive or negative judgements about the evaluated activity. Our central claim is that *the more positive (or negative) sentence-level assessments a report contains, the more positive (or negative) the overall performance of the evaluated activity.* An important limitation in that regard is that this measure treats all sentences equally. It is well possible that some sentences contain several judgements or are more important than others. We account for this in the validation below.

The *IOEval* dataset contains close to one million sentences, which exceeds the scope of sentences that can reasonably be classified by human coders. For the classification of sentences as containing a positive or negative assessment (or neutral descriptive text), we therefore utilize recent breakthroughs in natural language processing techniques in the area of deep learning-based contextualized language models. In particular, we employ a state-of-the-art language model BERT (Bidirectional Encoder Representations from Transformers)[8] that was developed by Google in 2018 (Devlin et al., 2019). These models have been developed to conduct classification tasks, and they can be *fine-tuned* to more specific applications (see Zhuang et al., 2021), an approach which has become the standard for deep learning in natural language processing (Huo & Iwaihara, 2020; Sun et al., 2019). For instance, scholars used a fine-tuned BERT model for classifying textual review and social media (e.g., Twitter) data (Chiorrini et al., 2021; Pota et al., 2021). However, according to our knowledge, no prior work has applied a fine-tuned BERT model on lengthy political report data.

Fine-tuning requires input data. We used manually coded (labeled) text from 180 evaluation reports selected at random from the nine IOs in the *IOEval* dataset. After establishing that the nature of sentences in executive summaries is not different as to how sentences in the main text of the reports are written, we coded all sentences in the executive summaries. This enabled to cover a broader variety of reports (and therefore topics and activities) compared to an approach that would have manually coded full reports. Overall, the 180 executive summaries contained 10,296 sentences. The coding involved three coders who first coded the same set of reports to establish a common understanding. Once the inter-coder agreement exceeded 80%, coding proceeded individually but weekly meetings were held to clarify questions and maintain the inter-coder agreement.

[8] The base model used for the fine-tuning is a BERT uncased 12-layer model provided by HuggingFace (2021).

Each sentence was labeled as either positive or negative assessments of the evaluand; or as neutral when it contained no performance judgement.[9] Below we give examples for typical sentences:

- *Positive assessment sentence*: "Effective management and efficient allocation and use of resources have contributed to the achievement of results grounded in EU normative standards…" (UN WOMEN, 2020: 16).
- *Negative assessment sentence*: "However, the lack of an institutional host or anchor for the awareness-raising campaign strategy … casts some doubts with regards to the sustainability of the campaign" (IOM, 2019: 3).
- *Descriptive sentence (neutral)*: "Two distinct training approaches … were compared in this evaluation – the foundational and the enhanced ECD kit interventions" (UNICEF, 2018: 8).

To implement the fine-tuning, all sentences were tokenized using the BERT tokenizer, and the tokens were converted to their identifiers according to the BERT's dictionary. We fine-tuned the model on around 90% of the labeled data (the rest was used for validation, see below). The transfer-learning step took approximately 0.5 h for four epochs using google collab's graphics processing unit. The output of the fine-tuned BERT are three values, i.e., logits that get normalized using the softmax function to retrieve prediction probabilities for positive, negative or neutral assessment. To determine the predicted class, we took the maximum among these three probability values. In other words, the class label with the highest probability is the predicted class label for the particular input sentence. For more details on the manual coding rules and procedures, see Appendix III

Overall, the language model enabled us to classify 995,743 sentences comprised in the 1,082 reports of the *IOEval* dataset, predicting for each sentence whether it contains a positive or negative assessment of the IO activity under evaluation – or neutral text. For each evaluation report, we then computed a simple proportional calculation where the ratio of positive sentences in a report to negative sentences is used to construct a continuous measure reaching from 0 to 1, whereby 1 denotes that 100% of sentence-level assessments (excluding neutral) are positive. Figure 1 provides an overview of the resulting *IOEval* dataset, showing year coverage and positive assessment share at the report level for each IO.

[9] In some instances, sentences contained contradictory judgements (e.g., "One the one hand…, on the other hand…"). In such cases, we coded the dominant statement which hence informed the language model.

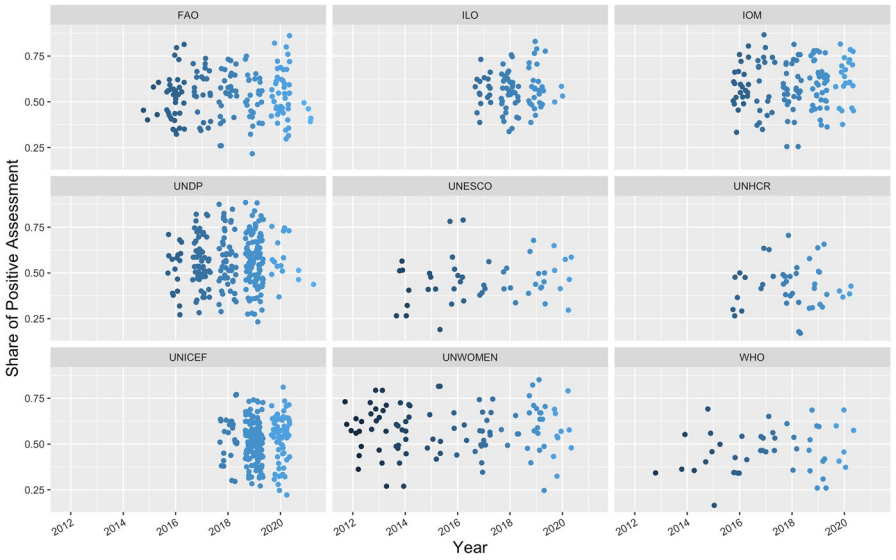

Fig. 1 IO performance at report level. The figure plots individual reports in our dataset and their share of positive findings (y-axis) for IOs over time (x-axis)

5 Validation

The performance of IO activities is a latent concept, and as such its true values are unknown (Kellstedt et al., 1993). Certain frameworks have been proposed to estimate the validity of novel measures for latent concepts (Adcock & Collier, 2001; see also Lührmann et al., 2020; Weidmann & Schutte, 2017). Following these, we report findings from content, convergent, and construct validation of our novel performance measure.

5.1 Content validation

Content validation examines how a measure converges with underlying concepts (Adcock & Collier, 2001). In this case, we address the "adequacy of content" (Adcock & Collier, 2001: 538) both quantitively and qualitatively.

First, we quantitively assess the extent to which the algorithm accurately predicts sentences as compared to a human's decision regarding the evaluated activity's performance. For that, our human coders manually labeled approximately 2,000 sentences as positive, negative, or neutral assessments from randomly selected evaluation reports' executive summaries which were not in the original training data. The hand-coded sentences were equally distributed among the three code dimensions (666 sentences for each group comprising of either positive, negative or neutral assessments). Then, the same sentences were classified by the algorithm and the results were compared. Supporting the model's validity, its accuracy (i.e., the human-algorithm coding agreement) on this test data reaches

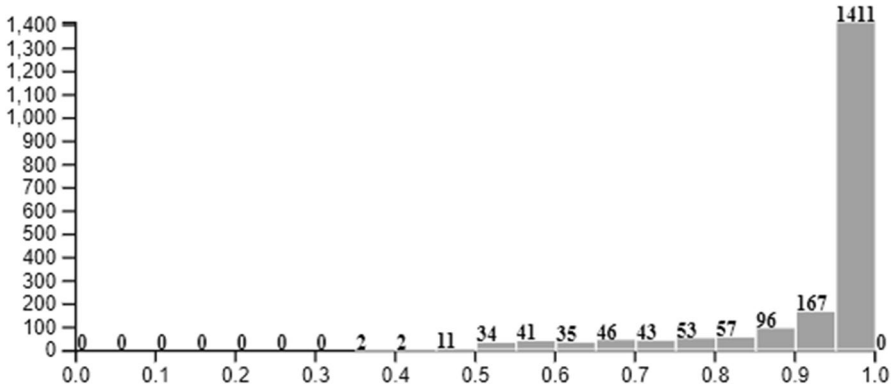

Fig. 2 Probability scores for the predicted label (x-axis) for all test data sentences (y-axis, number of sentences)

89% (87% for positive assessment sentences, 90% for negative assessment sentences, and 92% for neutral sentences).

In addition, for each predicted sentence the algorithm gives us a probability score which shows how confident the model is regarding the allocated code (positive, negative, neutral). The histogram in Fig. 2 shows the probability distribution for the predicted labels. The chart reveals that most of the class labels are predicted with a high probability (>=95%) indicating high model confidence. This yields no indication of a systematic bias. In the Appendix (Tables A IV.1–2), we show example sentences for rare cases when the model's probability for a prediction did not exceed 50%.

As the second approach to assess content validity, we employ content analysis of extreme case reports. This aims to establish whether reports that contain a very high or low share of positive sentences indeed represent cases of very successful or unsuccessful IO activities. We selected four reports located at the far end of our scale from the group of executive summaries we coded manually when training the language model. Below, we demonstrate that these executive summaries indeed contain accounts of very successful or unsuccessful activities.

Regarding the two most negative reports, UNESCO's evaluation of its field reform in Africa (UNESCO, 2015) has only 16% positive sentences in the executive summary, compared to 84% negative sentences.[10] This indicates very poor performance. Manual inspection yields that this is also reflected in the content of the executive summary, which explains that "achievements thus far have been limited" (5), that the field reform "was not complemented by a strategy (…) or a robust implementation plan with clear targets and deliverables" (5), and that the "overall

[10] In line with our performance measure, the percentage values in this section refer to assessment sentences, thus excluding neutral sentences, such as descriptive text.

leadership, monitoring and oversight over the reform was ambivalent, uncoordinated and uneven" (6). Next, the evaluation of UNHCR's efforts to phase down its presence in Angola, Botswana and Namibia also has received only few (13%) positive sentences (UNHCR, 2018). Here, evaluators strongly criticized UNHCR's lack of strategy and guidance for the phase down process. We quote two exemplary sentences below:

> "Nevertheless, when the 2013 decision [on the phasedown, *the authors*] was made, the intended outcomes were formulated only in terms of office structures and presences. The decision did not include a transparent analysis of underlying assumptions and preconditions that could have guided field offices; as a result, appropriate strategies, with clear indicators, operational milestones and roadmaps were not developed (…); and could not be used to support the review of progress in subsequent years" (iv).

By contrast, two reports ranging at the other end of the scale, with the highest share of positive sentences, paint a completely different picture. The executive summary of the FAO's evaluation of its project on water use and management in the Sana Basin contains 100% positive assessment sentences. The report indeed consistently emphasizes the "competence and determination of the project team", saying that it "met its objectives" and "provided an effective model" for future regulation (FAO, 2018). Similarly, the IOM's evaluation of its project on human security also contains only positive assessments. In the report, evaluators state that the project was "relevant", "effective", and "efficiently implemented both in terms of operations and financially" (IOM, 2018). Overall, this illustrative content analysis of extreme cases indicates that the low or high share of manually coded evaluation reports clearly and adequately reflects activities that contain highly positive or negative performance assessments.

Overall, these findings demonstrate that the language model classifies the content of sentences with a very high accuracy and the share of positive sentences adequately represents the performance-related content of reports. Given that predicted probabilities for each sentence are known (model's confidence), the insecurity of the estimation can be incorporated as a control variable in future analyses.

5.2 Convergent validation

Convergent validation compares how a novel measure correlates with a similar concept (Adcock & Collier, 2001). In this case, we compare the text-based performance measure with an exogenous performance metric of IO activities. The World Bank (WB) offers such alternative data in the form of a manual outcome rating for projects provided by the organization's Independent Evaluation Group (IEG). Their standardized rating procedure aims to provide coherent and consistent performance scores to allow comparison over time and between countries (IEG, 2022). If the shares of positive assessments, as identified by our model, correlate with the IEG ratings (in the same sample of WB reports), then our confidence that these shares indeed reflect project performance increases. Note that there is a discussion about the validity of the

IEG performance scores (e.g., Malik & Stone, 2018), which is why we expect no perfect correlation.

To compare our model results with the IEG ratings, we collect a sample of WB evaluation reports which all contain standardized IEG performance ratings. We focus on the most general metric termed "Outcome", which measures "the extent to which the operation's major relevant objectives were achieved, or are expected to be achieved, efficiently" (IEG, 2014: 5). Hence, it generally aligns to the evaluation objectives specified by the UN Evaluation Group. There are six possible ratings for this metric: *"Highly Satisfactory", "Satisfactory", "Moderately Satisfactory", "Moderately Unsatisfactory", "Unsatisfactory", "Highly Unsatisfactory"*.

We collected available WB reports published between 2012 to 2021 in line with the timeframe used for the compilation of our main *IOEval* dataset. The WB dataset consists of 661 reports in total, which includes two different report types: 473 are so-called Implementation Completion and Results Report Reviews (ICRRs), which the IEG drafts for all WB projects based on interviews and document analysis. 189 documents are Project Performance Assessment Report (PPARs). These reports are conducted on 20% of all ICRRs but involve much more in-depth research.[11] The report types thus vary in length, scope and how long after the project end they were performed but, crucially, utilize the same ratings framework. After applying the same data-preprocessing steps as used above, the reports were separated into their composite sentences (255,732 in total)[12] for feeding into our language model. In turn, average shares of positive assessments per evaluation report were calculated and correlated with the IEG ratings.

Figure 3 shows at a report level (represented by dots on the plot) how the share of positive assessments per report (y-axis) corresponds with the associated IEG ratings (x-axis). Density estimates of the distribution show that reports associated to each WB class are approximately normally distributed around progressively median points and that these distributions are statistically distinct (Figure A V.3 in the Appendix). Clearly visible at this level is a steady positive correlation between the two variables, with each higher rating resulting in a distribution

[11] ICRRs are available at: https://documents.worldbank.org/en/publication/documents-reports. PPARs are available at: https://finances.worldbank.org/Other/IEG-World-Bank-Project-Performance-Ratings/rq9d-pctf (both accessed 22 March 2023). We began with collecting all available PPAR pdfs with evaluation year between 2012 and 2021, and then matched the corresponding ICRRs. Some ICRRs were drafted prior to 2012, which expands the timeframe of our dataset beyond that year. To gain more documents, we added all available ICRRs for 2020.

[12] To ensure the quality of the analysis, certain restrictions were placed on the length and number of sentences. After removal of neutral sentences, if a report had less than 30 sentences, it was excluded. This was chosen as a minimum to ensure true reflection of the report. Likewise, sentences over 90 words long were excluded as these were most probably multiple sentences incorrectly merged during the cleaning process.

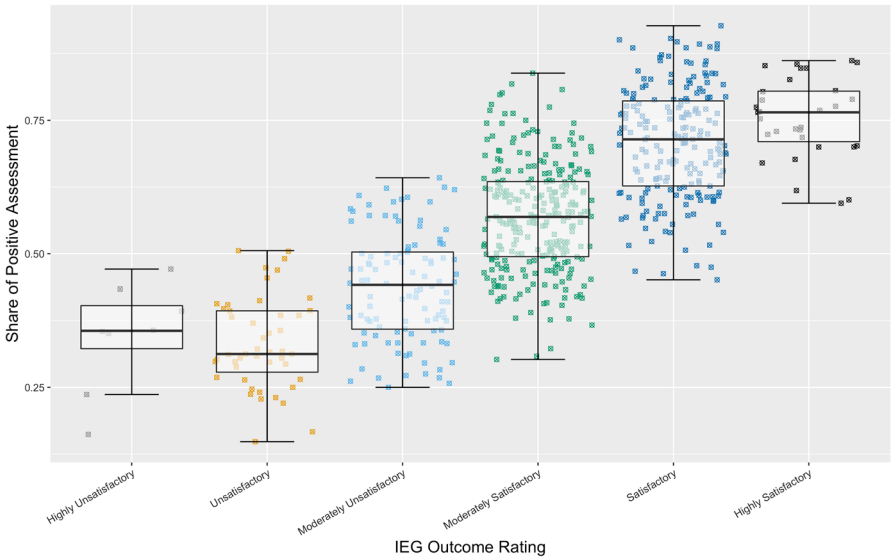

Fig. 3 Share of positive assessment (y-axis) across IEG Ratings (x-axis) by report. Note that IEG Outcome Rating is an ordinal variable with no values in between, the points have been spread randomly to improve interpretability

centered around a higher median point for the positive assessment share variable, save for the lowest IEG outcome rating level of "Highly Unsatisfactory". There is also a strong positive correlation (Spearman's rank correlation[13]: $r(648) = 0.76$, $p = 2.2e\text{-}16$).

Whilst this initial examination supports the convergent validity of our novel measure with the manual metric of performance by the IEG, it also shows that the distribution of the lowest rating level does not follow the expected trend. However, statistical power analysis shows that the number of reports rated as "Highly Unsatisfactory" is too small to give an accurate representation of the distribution for this rating level.[14] While we still include this underpowered rating level in the main analysis, separate models in the Appendix exclude this level and show that the differences are negligible (see Table A VI.2).

To further investigate to what extent each IEG rating category matches our text-based measure, we report findings from an ordinary least squares (OLS) regression analysis in Table 1. Model 1 shows the bivariate association between the IEG ratings, taken as a continuous variable between 1 ('Highly Unsatisfactory') and 6 ('Highly Satisfactory') and our continuous positive assessment share measure. We

[14] Statistical power analysis supports that the number of observations obtained for the lowest rating level is insufficient. The computed statistical power for "Highly Unsatisfactory" is 0.648, which is below the threshold of 0.8 as suggested by the literature (Bausell and Li (2002: 40) See also Appendix V.

[13] This type of test was chosen due to its superiority when dealing with a continuous-ordinal variable relationship (Khamis, 2008). For the calculation, we used the *cor.test* function within the *stats* package in R.

Table 1 Models of IEG outcome rating and average assessment relationship. IEG Outcome treated as continuous value (1–6). Restricted: neutral sentences removed; 95% confidence intervals displayed in brackets below the coefficients (alternative models reported in Appendix VI yield highly comparable results)

Model Summary			
	Dependent variable:		
	Positive Assessment Share		
	(1)	(2)	(3)
IEG Rating (numeric)	0.117***	0.115***	0.114***
	(0.109, 0.124)	(0.108, 0.123)	(0.106, 0.123)
Report Type (PPAR)		-0.042***	-0.043***
		(-0.059, -0.025)	(-0.064, -0.022)
Constant	0.106***	0.124***	0.066
	(0.074, 0.138)	(0.092, 0.156)	(-0.153, 0.285)
Year FE	No	No	Yes
Country FE	No	No	Yes
Observations	657	657	657
R^2	0.578	0.593	0.699
Adjusted R^2	0.578	0.592	0.628

Note: *p < 0.1; **p < 0.05; ***p < 0.01

find that each higher level of IEG Rating is associated with a statistically significant mean increase in predicted positive assessment share of a report. The finding also holds for other model specifications as reported below.

In Model 1, that has no other control variables, each rating level is associated with an increase of 0.117 in share of positive assessments at a 99% confidence level. With an R^2-value of 0.58, the measure explains substantive variation dependent variable. This supports the claim that there is a consistent positive relationship between our novel measure and the manual measure produced by the IEG. This convergence, in turn, supports the argument that our text-based performance measure is aligned with the IEG measure.

A limitation, also visible in Fig. 3 above, is that there is overlap between the rating levels. Especially for the ratings "Satisfactory" and "Highly Satisfactory". This might also be reflected in the R^2-value showing that in Model 1 the IEG scores account for approximately 58% of the variation in the positive assessment share, and not more (see also the prediction plot in Appendix VI, Figure A VI.2). Thus, reports with a relatively high share of positive assessments indicate high performance but map only partially into a specific IEG rating category. It is important to stress that this does not undermine the positive correlation between both measurements detected above. But it shows that our novel measure does not allow to predict each IEG score in a deterministic way.

The reason is that the performance of IO activities is a latent concept. There is no certainty that either the IEG scores or our own sentence-based measure are fully accurate. On the one hand, the IEG measure could be biased. One study that

investigated the IEG outcome scores by re-rating performance based on a second reading of the evaluation reports found some deviations (Malik & Stone, 2018; see also Weaver, 2010). On the other hand, our text-based measure could also be a source of bias. Evaluators could be discouraged to report concrete negative examples because they know that reports will be published with their names on them. In our sample, IEG evaluations are on average slightly more positive compared to UN evaluation reports (Appendix V, Figure A V.1). But whether the remaining gap between both measures is caused by our text-based measurement or the IEG's measurement remains up for further investigation, as highlighted in the discussion.

To investigate what other factors could influence the relationship, we include additional covariates to the regression analysis. First, Model 2 investigates whether there may be a systematic flaw in our text-based input data. As mentioned above, there are two types of IEG evaluation reports, the shorter and more standardized Implementation Completion and Results Report Reviews (ICRRs), and the much longer and more in-depth Project Performance Assessment Report (PPARs). It is possible that the length or accuracy of these report types accounts for the remaining gap between the two measures. In Model 2, the coefficient remains statistically significant but there is a slight difference (-0.042) between both report types in our sample. The R^2 remains close to that of Model 1, which suggests that the type of report does not seem to drive the gap.

A second possibility is that there are biases on the side of human coders of the IEG. For example, it could be that a group of IEG employees who provide the scores for a range of reports from a given country or at a given time could systematically deliver more or less positive performance grades. In Model 3, we therefore add year fixed effects indicating the year of report completion and the country/region in which the evaluation was conducted (see Appendix VI for full table). The R^2 figure increases substantively with the inclusion of these controls, meaning that the model accounts for just under 70% of the variation in the dependent variable, while coefficients remain highly comparable to Model 1.

Overall, our results thus support the proposition that our measure for performance tracks that of the human coders at the IEG.

5.3 Construct validation

The third step is construct validation, i.e., whether a measure has theoretically expected results. For that, we draw on the different types of evaluations in the *IOEval* dataset. These types refer to the different IO activities being assessed: projects, programs, institutional, or thematic activities. For program and project type activities, management literature and literature on IO performance yield the expectation that projects, with their much more narrow and attainable goals, should on average have higher performance scores compared to the more ambitious and complex program activities.

In order to achieve long-term policy goals, modern public organizations employ strategic management frameworks that structure their work along a hierarchical logic, stretching from abstract goals to concrete action (Maylor et al., 2006; Poister et al., 2015): At the top is the organizational strategy, outlining a vision with respect to a

policy field. Programs specify mid- and long-term goals and objectives. Projects are concrete activities with specific goals that are carried out within a short time frame. This has implications for the success chances of projects and programs as a study by Shao et al., (2012: 46) found: "project success is focused on project deliverables, whereas program success is concerned with delivering benefits and strategies."

The idea of hierarchically structured policy activities has also affected how the international community approaches major policy change, such as the fight against climate change or the strive towards sustainable development goals. Broader goals (and indicators) are agreed upon politically. Organizations and other actors then design policy programs and project activities in order to enable transformation pathways towards these goals. Such a hierarchical logic of programs and projects also structures the internal management of most UN system organizations. For example, the UNDP's operational policy specifies[15] that "UNDP's results are outlined in Country Programmes, Regional Programmes and the Global Programme. […]. Programmes are operationally implemented through projects with multi-year or annual work plans."

Success chances are not equally distributed between projects and programs. To account for this, IO performance literature has used the metaphor of a pyramid. Gutner and Thompson (2010: 236), for instance, expect good performance to "trickle up," with success at each lower stage serving as building blocks for success as we move up the pyramid." Thereby, the success chances are higher for projects, with their more attainable goals. Programs by contrast depend on the success of their downstream projects. This has also been shown empirically by aid effectiveness literature. In their study of 1,600 Asian Development Bank projects and programs, Feeny and Vuong (2017: 329) find that "projects are more likely to be successful than programs". For the construct validation, the expectation therefore is that due to their more limited scope and objectives, IO activities at the project level should on average perform better than activities at the program level.

Figure 4 shows the difference in means of positive assessments between project and program level activities for the UN organizations in the *IOEval* data set. Program level activities contain on average 53% positive assessments, whereas projects contain on average 59% positive assessments. This difference in the average share of positive assessments is statistically significant, as displayed in Table 2 as well as in Appendix VII. Aid effectiveness literature highlight that most variation in performance ratings can be explained based on country-level determinants (Denizer et al., 2013). Unsurprisingly, including year, country and IO fixed effects therefore reduces the difference between both groups. But in line with Feeny and Vuong (2017: 329), project reports remain 0.038 percentage points more positive at a p-value below 0.01.

Consistent with the theoretical expectation, we therefore find that projects, which have more attainable goals, perform better than programs. Being able to reproduce such an established expectation based on our proposed performance measure thus serves as

[15] Available at https://popp.undp.org/SitePages/POPPChapter.aspx?TermID=e3cd9ba4-cada-4d5e-9fc7-94547251b9ec (accessed 22 March 2023).

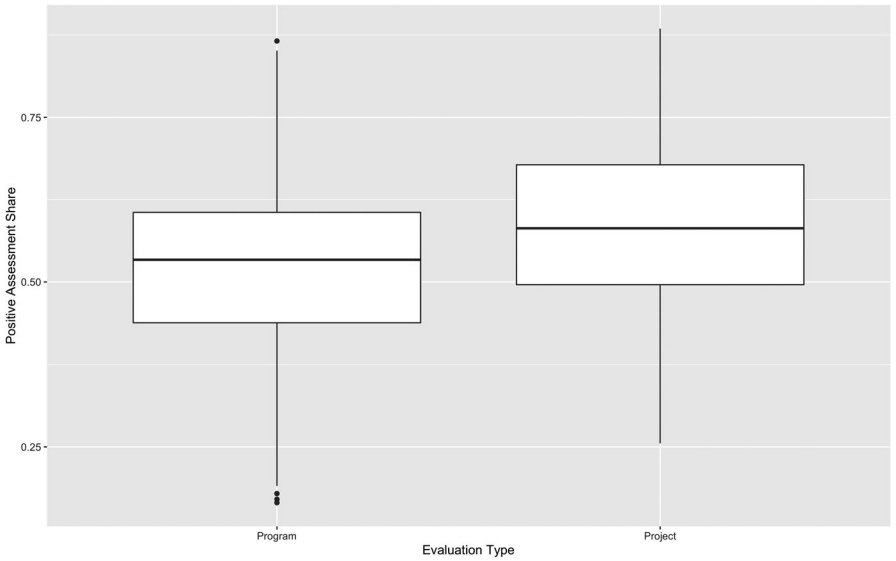

Fig. 4 Difference in means of the share of positive assessments between program level activities (left y-axis) and project level activities (right y-axis) displayed in boxplots

Table 2 Models of evaluation type rating and average assessment relationship. Program evaluations used as base level for evaluation type. 95% confidence intervals displayed in brackets below the coefficients

Model Summary		
	Dependent variable:	
	Positive Assessment Share	
	(1)	(2)
Evaluation Type (Project)	0.059***	0.038***
	(0.042, 0.075)	(0.017, 0.058)
Constant	0.526***	0.415***
	(0.515, 0.538)	(0.298, 0.532)
Year FE	No	Yes
Country FE	No	Yes
IO FE	No	Yes
Observations	925	925
R^2	0.050	0.345
Adjusted R^2	0.049	0.191

Note: *p < 0.1; **p < 0.05; ***p < 0.01

construct validation. Furthermore, the repeated finding of differences in outcomes for projects and programs, in combination with our data set, opens a new and interesting route for research: Efforts to achieve long term policy goals, such as climate change or sustainable development, could benefit from an in-depth understanding of how certain combinations of programs and projects enable or obstruct transformation pathways.

6 Discussion

To summarize, we combined three validation strategies to scrutinize the validity of our novel measure for the performance of IO activities. Content analysis of extreme case reports finds that a very high or low share of positive sentences corresponds with reports that present IO activities as very successful or unsuccessful. Furthermore, we show that our language model classifies sentences with an accuracy of 89%. Our performance measure also converges strongly with an alternative metric offered by the World Bank's IEG. And it has theoretically expected results. We therefore argue that these analyses act as a suitable validation measure under the content, convergent and construct validation framework utilized in previous research (Adcock & Collier, 2001; Lührmann et al., 2020; Weidmann & Schutte, 2017). We maintain our initial proposition on the text-based performance measure, stating that *the more positive assessments an evaluation report contains, the more positive the performance of the evaluated IO activity*. An advantage compared to existing approaches is that our measure takes the full evaluation report into consideration, and it produces a continuous measure rather than a categorical one.

Although the language model classifies sentences with a very high accuracy and although the report-level measure correlates highly with alternative performance scores, there are remaining limitations.

First, the process of generating the share of positive assessments has several opportunities for introducing error, such as when sentences are separated incorrectly or inaccurately classified. This could lead to particularly positive or negative parts of text being subsumed in other text or just going undetected. Moreover, there is currently no weighting process applied to the sentences. A long sentence that contains lots of comments on major aspects of a project contributes as much weight to the positive assessment share as a short sentence that pertains to something of minor significance to the overall project. Future refinements of the language model could however improve its accuracy, for instance by weighting sentences.

Secondly, regarding convergent validation, the positive assessment shares of individual reports overlap to some extent across IEG rating levels, causing a gap in the correlation. However, this gap is reduced if control variables are added to the model. Hence, it might emerge due to certain biases in how the IEG assigns scores to its project evaluation reports (Malik & Stone, 2018; see also Weaver, 2010) (although future research should scrutinize this claim). Nevertheless, considering the positive assessment shares are distributed normally around a mean that correlates well with the IEG measure, we deem the underlying convergent validation not impinged.

A third question refers to the extent to which our measure applies to unseen evaluation reports. The language model was trained on the UN reports (excluding the World Bank) and we show that it also performs well in predicting the World Bank's IEG outcome ratings. This supports that there is sufficient consistency between the language used by the IEG and UN evaluators. However, these individuals form a relatively homogenous epistemic community of experts who work on evaluation in international politics. There may be other evaluation cultures at other types of organizations or policy fields. While we are

optimistic that generalization is possible, at least as long as evaluations follow the UNEG (or OECD-DAC) criteria, further research is needed to substantiate this claim.

With these remaining limitations, we suggest that our text-based performance score should not necessarily be understood as a measure that is superior to previous numeric performance scores, such as by the IEG. In that sense, we do not claim that it is able to capture the 'true value' of the latent concept of performance more accurately than a human coder does. However, given the way it is constructed, based on full evaluation reports and a classification of sentence-level assessments, it offers an alternative data source – or, when no performance ratings exist (e.g., as in evaluations from the nine UN system IOs in our dataset) – a valid novel source on the performance of IO activities. The key advantage is that it provides a highly *reliable* and *replicable* measure that can be applied consistently to any evaluation report (bearing in mind the above limitations). It ensures *transitivity*, which means that a report containing more positive assessment sentences on an activity is also rated higher on the performance metric, compared to a report with fewer positive assessment sentences. And our measure is also *scalable*, enabling researchers to easily expand the existing *IOEval* dataset with new evaluation reports.

7 Conclusion: Areas for application and future research

As its main contribution, this article develops an original method to extract the performance information from text-based evaluation reports by classifying sentence-based assessments as positive, negative, or neutral, and by calculating the share of positive assessments per evaluation report. We demonstrated this method's validity for IOs in the UN system by means of content, convergent, and construct validation. Moreover, we publish a novel dataset of IO evaluation reports, with performance measures on 1,028 evaluated project, program, thematic, and institutional activities from nine UN system IOs. It contains cleaned text at the level of close to one million sentences, as well as the probability values for our classification of sentences as positive or negative assessments, or neutral descriptive information. We also publish the language model which means that anybody can expand the *IOEval* dataset.

This offers a range of exciting opportunities for future research and practitioner application. First, the dataset and the model enhance our ability to study the *causes* of IO performance. Existing studies have already pointed to relevant performance-affecting factors like transparency (Honig et al., 2022; Marchesi & Masi, 2021), level of control and autonomy (Honig, 2019; Lall, 2017), unilateral donor influence (Watkins, 2022), or decentralization of IO staff (Honig, 2020) and their competence (Bulman et al., 2017; Heinzel, 2022; Heinzel & Liese, 2021). However, these studies focus mainly on the field of international and bilateral development assistance, broadly defined. Fewer insights exist for organizations in other fields, such as humanitarian aid, health, and social policy. By treating our model's performance score as a dependent variable, scholars can explore factors explaining the successes and failures of IO activities for an extended set of organizations and policy fields. Furthermore, in addition to *project* performance scores, reports in the *IOEval* dataset also cover activities at the *program*, *institutional,* and *thematic* level. Factors explaining performance can be explored using our metadata variables

which account for a range of contextual factors (years, types of activity, report level as country, regional, global).

New insights can also be gained by employing additional methods of (computational) text analysis to our text corpus. Keyword-in-context or topic modelling analyses are just some of the simpler methods that enable extracting information on the factors that account for IO performance (see, for instance, Cormier & Manger, 2022). Especially for the work of the UN system IOs at the country level, this presents a rich source of data for comparative analyses.

To be sure, our data allows comparing performance tendencies both within and between IOs, yet this should be done with caution. So far, we lack information on the underlying evaluation case selection. While project and program evaluations are oftentimes part of the regular project management life-cycle, organizations may commission institutional or thematic evaluation precisely for areas where their performance is *weaker*. The evaluation data should therefore not be used at the aggregate IO level in the sense of a general IO performance score. Comparisons between organizations are hence possible for certain types of activities and when considering the lack of information about evaluation case selection as a limitation.

Second, the dataset and the model also allow to study the *consequences* of IO performance. For example, existing research has asked how IOs affect states' domestic spending (Stubbs et al., 2020), conflict recurrence or economic recovery (Flores & Nooruddin, 2009). Treating our performance score as an independent variable, scholars could similarly explore further impacts of IO performance, for instance, regarding IO funding patterns (see Patz & Goetz, 2019) or IO survival and member state contestation (see Borzyskowski & Vabulas, 2019; Eilstrup-Sangiovanni, 2020).

Lastly, policymakers, too, can use the novel text-based performance metric. To the extent that their organization uses quantitative scores, they can investigate potential gaps between performance metrics as an additional quality check. But as we argue above, most IOs do not even have such performance information. In these cases, our model could be used to identify outlier reports, for example, extremely negatively or positively evaluated activities. Such insights should contribute to learning and accountability for evaluation systems in IOs. After all, the average evaluation in the UN system costs around 500,000 USD (OIOS, 2019) which certainly warrants some attention as to the quality and impact of their findings.

As Gutner and Thompson (2010: 234) write, "understanding and explaining the performance of international organizations is uniquely difficult – and uniquely interesting". We thus hope that the introduced dataset and the language model algorithm – both of which we make available to practitioners and the academic community – helps to overcome at least some of these difficulties.

8 Author contribution statement

Steffen Eckhard and Vytautas Jankauskas led the development of the research design, conceptualization, and theory (50/50%). Elena Leuschner and Ian Burton led the text analysis (50/50%). Tilman Kerl and Rita Sevastjanova led the training of the BERT language model (50/50%). The order of authors reflects the significance of the authors' contributions.

9 Data availability statement

All data generated and analyzed during the current study, including the language model developed as part of this study, are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0SI2VX.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11558-023-09489-1.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest/competing interest statement The research leading to these results received funding from the German Research Foundation (DFG) under Grant Agreement No EC 506/2–1.
The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adcock, R., & Collier, D. (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review, 95*(3), 529–546.

Bausell, R. B., & Li, Y.-F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical, and social sciences*. Cambridge University Press.

Bulman, D., Kolkma, W., & Kraay, A. (2017). Good countries or good projects? Comparing macro and micro correlates of World Bank and Asian Development Bank project performance. *Review of International Organizations, 12*(3), 335–363.

Buntaine, M. T., & Parks, B. C. (2013). When Do Environmentally Focused Assistance Projects Achieve their Objectives? Evidence from World Bank Post-Project Evaluations. *Global Environmental Politics, 13*(2), 65–88.

Cormier, B., & Manger, M. S. (2022). Power, ideas, and World Bank conditionality. *Review of International Organizations, 17*(3), 397–425.

Chiorrini A, Diamantini C, Mircoli A, et al. (2021) Emotion and sentiment analysis of tweets using BERT. *In EDBT/ICDT Workshops*.

Denizer, C., Kaufmann, D., & Kraay, A. (2013). Good countries or good projects? Macro and micro correlates of World Bank project performance. *Journal of Development Economics, 105*, 288–302.

Dreher, A., Klasen, S., Vreeland, J. R., et al. (2013). The Costs of Favoritism: Is Politically Driven Aid Less Effective? *Economic Development and Cultural Change, 62*(1), 157–191.

Eckhard, S., & Jankauskas, V. (2019). The politics of evaluation in international organizations: A comparative study of stakeholder influence potential. *Evaluation, 25*(1), 62–79.

Eilstrup-Sangiovanni, M. (2020). Death of international organizations: The organizational ecology of intergovernmental organizations, 1815–2015. *Review of International Organizations, 15*(2), 339–370.

FAO (2018) *Final Evaluation of the Project on Decentralized Supply and Water Use Management in the Sana's Basin to Sustain Water Resources and Rural Livelihoods*.

Feeny, S., & Vuong, V. (2017). Explaining Aid Project and Program Success: Findings from Asian Development Bank Interventions. *World Development, 90*, 329–343.

Flores, T. E., & Nooruddin, I. (2009). Financing the peace: Evaluating World Bank post-conflict assistance programs. *Review of International Organizations, 4*(1), 1–27.

Graham, E. R. (2014). International Organizations as Collective Agents: Fragmentation and the Limits of Principal Control at the World Health Organization. *European Journal of International Relations, 20*(2), 366–390.

Geli P, Kraay A and Nobakht H (2014) *Predicting World Bank Project Outcome Ratings*.

Gutner, T., & Thompson, A. (2010). The politics of IO performance: A framework. *The Review of International Organizations, 5*(3), 227–248.

Heinrich, C. J. (2012). Measuring Public-Sector Performance and Effectiveness. In: Pierre J and Peters BG (eds) *The SAGE handbook of public administration:* Los Angeles: Sage, pp. 32–49.

Heinzel. M. (2021). Divided loyalties? The role of national IO staff in aid-funded procurement: Early view. *Governance*. DOI: https://doi.org/10.1111/gove.12650.

Heinzel, M. (2022). International Bureaucrats and Organizational Performance: Country-Specific Knowledge and Sectoral Knowledge in World Bank Projects. *International Studies Quarterly* 66(2).

Heinzel, M., & Liese, A. (2021). Managing performance and winning trust: How World Bank staff shape recipient performance. *Review of International Organizations, 16*(3), 625–653.

Honig, D. (2019). When Reporting Undermines Performance: The Costs of Politically Constrained Organizational Autonomy in Foreign Aid Implementation. *International Organization, 73*(1), 171–201.

Honig, D., Lall, R., Parks, B. C. (2022). When Does Transparency Improve Institutional Performance? Evidence from 20,000 Projects in 183 Countries. *American Journal of Political Science*: 1–21.

Honig, D. (2020). Information, power, and location: World Bank staff decentralization and aid project success. *Governance, 33*(4), 749–769.

Honig, D., & Weaver, C. (2019). A Race to the Top? The Aid Transparency Index and the Social Power of Global Performance Indicators. *International Organization, 73*(03), 579–610.

HuggingFace (2021). Available at: https://huggingface.co/ (accessed 23 August 2021).

Huo, H. & Iwaihara, M. (2020). Utilizing BERT Pretrained Models with Various Fine-Tune Methods for Subjectivity Detection. *In Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*: 270–284.

IEG (2022). IEG Methodology. Available at: https://ieg.worldbankgroup.org/methodology (accessed 24 March 2022).

IEG (2014). *Guidelines for Reviewing World Bank Implementation Completion and Results Reports*.

IOM (2018), *Final project evaluation: Building Sustainable Peace and Promoting Human Security of Cross-border Communities and Mobile Populations through Integrated Border Security and Management and Community-Based Peacebuilding Activities*.

IOM (2019). *Action to Support the National Coordinating Committee on Combating and Preventing Illegal Migration and Trafficking in Persons (NCCPIM&TIP) to Create a Safe and Secure Environment in Egypt: Evaluation Report*.

Kellstedt, P., McAvoy, G. E., & Stimson, J. A. (1993). Dynamic Analysis with Latent Constructs. *Political Analysis, 5*, 113–150.

Khamis, H. (2008). Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography, 24*(3), 155–162.

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

Lebovic, J. H., & Voeten, E. (2009). The Cost of Shame: International Organizations and Foreign Aid in the Punishing of Human Rights Violators. *Journal of Peace Research, 46*(1), 79–97.

Lipson, M. (2010). Performance under ambiguity: International organization performance in UN peacekeeping. *Review of International Organizations, 5*(3), 249–284.
Lührmann, A., Marquardt, K. L., & Mechkova, V. (2020). Constraining Governments: New Indices of Vertical, Horizontal, and Diagonal Accountability. *American Political Science Review, 114*(3), 811–820.
Malik, R., & Stone, R. W. (2018). Corporate Influence in World Bank Lending. *The Journal of Politics, 80*(1), 103–118.
Marchesi, S., & Masi, T. (2021). Delegation of implementation in project aid. *Review of International Organizations, 16*(3), 655–687.
Maylor, H., Brady, T., Cooke-Davies, T., et al. (2006). From projectification to programmification. *International Journal of Project Management, 24*(8), 663–674.
OIOS (2019). United Nations Evaluation Dashboard 2016–17: IED-19–002.
Parízek, M. (2020). *Negotiations in the World Trade Organization: Design and Performance*. Routledge.
Patz, R., & Goetz, K. H. (2019). *Managing Money and Discord in the UN: Budgeting and Bureaucracy*. Oxford University Press.
Poister, T. H., Aristigueta, M. P., & Hall, J. L. (2015). *Managing and measuring performance in public and nonprofit organizations*. Jossey-Bass.
Pota, M., Ventura, M., Catelli, R., et al. (2021). An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian. *Sensors* 21.
Shao, J., Müller, R., & Turner, J. R. (2012). Measuring Program Success. *Project Management Journal, 43*(1), 37–49.
Sommerer, T., Squatrito, T., Tallberg, J., et al. (2021). Decision-making in international organizations: Institutional design and performance. *Review of International Organizations*. https://doi.org/10.1007/s11558-021-09445-x
Sun, C., Qiu, X., Xu, Y., et al. (2019). How to Fine-Tune BERT for Text Classification? *China National Conference on Chinese Computational Linguistics*: 194–206.
Stubbs, T., Reinsberg, B., Kentikelenis, A., et al. (2020). How to evaluate the effects of IMF conditionality. *Review of International Organizations, 15*(1), 29–73.
Tallberg, J., Sommerer, T., Squatrito, T., et al. (2016). The performance of international organizations: A policy output approach. *Journal of European Public Policy, 23*(7), 1077–1096.
UNEG (2016).*Norms and Standards for Evaluation*. New York: United Nations Evaluation Group.
UNESCO (2015). *Lessons Learned from UNESCO's Field Reform in Africa: Evaluation Office*.
UNHCR (2018). *Evaluation of UNHCR's country operations in Angola, Botswana and Namibia: Assessment of phasing down UNHCR presence during the period 2012–2016*.
UNICEF (2018) *Early Childhood Development Kit: Humanitarian Evaluation*.
UN WOMEN (2020). *Ending Violence Against Women in the Western Balkans and Turkey: Final Evaluation*.
von Borzyskowski, I., & Vabulas, F. (2019). Hello, goodbye: When do states withdraw from international organizations? *Review of International Organizations, 14*(2), 335–366.
Watkins, M. (2022). Undermining conditionality? The effect of Chinese development assistance on compliance with World Bank project agreements. *Review of International Organizations*, 17(4), 667-690.
Weaver, C. (2010). The politics of performance evaluation: Independent evaluation at the International Monetary Fund. *Review of International Organizations, 5*(3), 365–385.
Weidmann, N. B., & Schutte, S. (2017). Using night light emissions for the prediction of local wealth. *Journal of Peace Research, 54*(2), 125–140.
Zhuang, F., Qi, Z., Duan, K., et al. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE, 109*(1), 43–76.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Steffen Eckhard[1] · Vytautas Jankauskas[1] · Elena Leuschner[2] · Ian Burton[1] · Tilman Kerl[3] · Rita Sevastjanova[4]

Vytautas Jankauskas
vytautas.jankauskas@zu.de

Elena Leuschner
elena.leuschner@gu.se

Ian Burton
ian.burton@zu.de

Tilman Kerl
tilman.kerl@tuwien.ac.at

Rita Sevastjanova
rita.sevastjanova@uni-konstanz.de

1 Zeppelin University, Am Seemooser Horn 20, 88045 Friedrichshafen, Germany

2 University of Gothenburg, Box 100 405 30, 412 96 Göteborg, Sweden

3 Vienna University of Technology, Wiedner Hauptstraße 76 Stiege 2, 1040 Vienna, Austria

4 University of Konstanz, 78457 Constance, Germany