

Pre-Analysis Plan

Measuring evidence use under social pressure:

A survey experiment on social desirability bias in self-reports of evidence-informed policymaking among Ethiopian civil servants

1. Background and motivation

1.1 EIPM measurement rests on self-reports

Interest in evidence-informed policymaking (EIPM) has grown rapidly in recent years, mostly in low- and middle-income countries (LMICs), and with it, a stream of literature evaluating interventions designed to strengthen public officials' use of research (Langer et al., 2016; Nduku et al., 2024). Across this work, one prominent analytical framework being used has been adapted from the behaviour-change literature on individual capability, opportunity, and motivation (COM-B; Michie et al., 2011). The COM-B framework has become one of the most widely recognized and used theoretical models for operationalizing EIPM at the individual level. For example, the two major syntheses of this literature both adopt it as their central framework: Langer et al. (2016), in their review of reviews, used a COM-B-based framework to classify interventions into six mechanisms of change operating through policymakers' capability, opportunity and motivation to use evidence, while Nduku et al. (2024) applied and refined the same framework in their systematic review of what works to increase evidence use in policy decision-making, with a particular focus on LMICs. Within the COM-B model, capability is typically measured as confidence in interpreting research; opportunity as access to data, time and researcher contacts; and motivation as attitudes, intentions and perceived norms.

What unites this body of work is its reliance on self-reported data. Most studies collect information from participants through online or in-person approaches (Oliver and de Vocht, 2015 in the UK; Armstrong et al., 2014 in Australia).¹ In LMIC settings, Kafiriri and Bondy (2006) administered self-completed questionnaires to Ugandan health workers, while Uneke et al. (2023) surveyed 209 Nigerian policymakers with a Likert-based instrument. Also, the Global Survey of Public Servants (GSPS) aggregated responses from more than 1.3 million civil servants across 23 countries (Schuster

¹ Oliver and de Vocht (2015) combined online surveys with telephone follow-up to elicit definitions of evidence among UK public-health policymakers; Armstrong et al. (2014) deployed the EvIDenT instrument across 45 Australian local governments.

et al., 2023).² The choice is understandable: self-reporting is cheap, scalable, and well-suited to internal constructs such as motivation.

However, self-reported measures are rarely validated against observable behaviour, and researchers often rely primarily on correlations among related survey items to justify them (Landry et al., 2003; Makkar et al., 2016). Moreover, they clearly suffer from the so-called social desirability bias. Social desirability bias (SDB) is the tendency for respondents to provide answers aligned with perceived social norms rather than their actual attitudes or behaviour. Uneke et al. (2023) acknowledged SDB as a limitation in their study given their reliance on self-reports without parallel objective verification. Mehmood et al. (2024), who randomized deputy ministers in Pakistan to econometrics training, relied on self-reports for their key outcomes. These included Likert-scale attitudes towards quantitative evidence and participants' stated willingness to pay for RCT evidence. They were, however, able to validate these self-reports against administrative fiscal data on actual funding recommendations, a step that remains rare in empirical work, as such records are seldom available or accessible to researchers. Some studies incorporate more objective measures, such as document reviews, secondary data, or counts of research citations in policy outputs (Makkar et al., 2016; Mehmood et al., 2024). Even so, such objective indicators typically serve as supplementary validation rather than core outcomes, because the slow and complex nature of policymaking makes behaviour difficult to observe and compare objectively and consistently over time. As a result, what is known about policymakers' engagement with evidence rests largely on what policymakers say they do — the gap this study is designed to address.

1.2 Why EIPM is vulnerable to social desirability bias

The literature distinguishes two mechanisms: impression management, the deliberate distortion of self-reports to project a favourable image, and self-deceptive enhancement, an honestly held but inflated self-assessment (Paulhus, 1991; Kim and Kim, 2016). Both mechanisms are activated by normative pressure, and both are difficult to detect without a comparison instrument. SDB has been documented across a wide range of self-report contexts, including racial attitudes, sexual behaviour, charitable giving, voter turnout, and recently COVID-19 compliance (Tourangeau and Yan, 2007; Gnambs and Kaspar, 2015; Morin-Chassé et al., 2017; Daoust et al., 2021b).

EIPM measurement sits at the intersection of several conditions known to amplify SDB. First, the norm is salient and unambiguous: international donors, training programmes, ministerial communications and academic guidance all describe evidence use as the marker of competent policymaking. Respondents do not need to guess what the “right” answer is. Second, the respondent population is professionally invested in being seen as evidence-informed. Policymakers' identity, promotion prospects and external standing are bound to perceived analytical competence. Lilleker (2003) argues that elite respondents systematically manage impressions in interviews, rewriting

² The GSPS covers both high-income countries (e.g., the United Kingdom, the United States, Australia, Canada) and low- and middle-income countries across the Americas, Europe, Africa and Asia; harmonized microdata are publicly available at globalsurveyofpublicservants.org.

history in their own favour and distancing themselves from poor decisions. Third, the items used to measure EIPM tap exactly the dimensions on which impression management is most attractive: how often the respondent personally seeks evidence, how confidently they interpret it, and how much it shapes the decisions they advise on. Fourth, the populations EIPM interventions usually target are educated professionals with analytical training. This population is traditionally the most prone to SDB: misreporting requires recognizing the norm and cognitively editing one's answer, so more educated and engaged respondents misreport more, not less (Clifford and Jerit, 2015).

There is, in addition, a structural feature of policymaking that further inflates self-reports: evidence is sometimes cherry-picked, invoked rhetorically rather than actually used to inform decisions, or manipulated to support positions already taken (Parkhurst, 2017). Policymakers may invoke research to legitimize positions arrived at through other means, treating the citation rather than the substance of evidence as the relevant signal, for instance. From a survey respondent's point of view, the answer "evidence informs my work" is true in the symbolic sense and may be reported in good faith even where instrumental use is limited.

1.3 Evidence that SDB-corrective designs work — and where they have been tried

A small but well-developed methodological literature shows that careful question design can recover meaningful variation in sensitive responses. Morin-Chassé et al. (2017) embedded a *face-saving preamble and guilt-free response option* in 19 post-election surveys across Europe and Canada and reduced reported turnout by up to 18 percentage points, with the largest movement among educated non-voters. Daoust et al. (2021b) extended the design to compliance with COVID-19 social distancing in three Canadian surveys and recovered up to 11 percentage points of admitted non-compliance; Daoust et al. (2021a) replicated the result across 12 countries; the face-saving option increased reported non-compliance by 9–16 percentage points in every country, with effects consistent across gender, age and education subgroups. Kim and Kim (2016) used a *list experiment* in Korea to show that self-reported public service motivation is inflated by 18–22 percentage points, with larger bias among older respondents. Bernardi and Nash (2023) make a parallel case for ethics measurement: a direct measure of SDB, measured by the impression management subscale developed by Paulhus (1991), remained a significant predictor of self-reported ethical behaviour even when the standard safeguards, such as anonymous self-administered questionnaires and indirect, third-person wording, were in place. Omitting a direct measure of SDB altered which substantive variables appeared significant; yet fewer than 5% of survey-based ethics studies include such a control. Kovács and Veselý (2025) document the rapid growth of population-based survey experiments and their application across policy domains, including education, health, anti-crime, pension and social policy.

The scope of these studies, however, is uneven, and the gap has three layers. The first is geographic — the validated face-saving and list-experimental designs cluster in high-income democracies, such as the United States, Canada, Western Europe, and South Korea (Kim and Kim, 2016; Daoust et al., 2021a). Kovács and Veselý (2025) document the same asymmetry for population-based survey

experiments in public policy: the majority of studies are conducted in developed countries such as Germany, the United States, and Japan, reflecting the concentration of research resources, with contextual factors such as institutional culture receiving limited cross-national comparison. The second is the target population. Existing validations sample ordinary citizens, and the outcomes are citizen-level behaviours, such as turnout, compliance, donations, and vaccination. Professional civil servants are, to our knowledge, absent from this literature, although elite respondents are precisely those argued to manage impressions most actively (Lilleker, 2003). The third is the construct itself. Voting, social distancing or mask-wearing are private behaviours governed by general social norms; evidence use is a professional practice bound to occupational identity, promotion prospects and organizational standing, for which no measurement validation exists in LMIC administrations, the very populations whose EIPM is currently measured most intensively. The Social Desirability Atlas (Bursztyn et al., 2025) confirms that SDB magnitudes vary substantially across behavioural domains and should be treated as a context-dependent distortion, but its underlying validation studies are themselves drawn overwhelmingly from high-income samples.

This asymmetry matters because the dominant interpretation of EIPM survey results in LMICs assumes that, even if absolute levels are inflated, comparisons across experimental arms will net out the bias. That assumption treats SDB as a uniform shift, and only in that special case is it harmless: if all respondents exaggerated equally, only the intercept would move, not the estimated relationships. In practice, the magnitude and pattern of SDB differ systematically across respondents, which makes it an omitted variable; estimates remain unbiased only if the bias is uncorrelated with the explanatory variable of interest (Kim and Kim, 2016). In an experiment, the explanatory variable is treatment assignment, so the netting-out assumption amounts to a strong invariance condition: the treatment must change the underlying construct without simultaneously changing the social meaning of reporting it. The methodological literature identifies this configuration precisely as the case in which SDB distorts treatment effects rather than levels: when a treatment plausibly shifts what respondents perceive as the socially desirable answer, the bias differs between arms and no longer differences out (Bursztyn et al., 2025; de Quidt et al., 2018).

An EIPM intervention approaches the worst case, because its explicit content is the norm itself: participants are taught that evidence use is the marker of competent policymaking, so treated respondents exit the intervention with both potentially changed behaviour and demonstrably heightened pressure to report changed behaviour, two channels that are observationally equivalent at endline. The inferential problem therefore cuts both ways. A positive effect may reflect respondents learning what to say rather than what to do; a null may mean the intervention failed, or that it succeeded while ceiling effects in baseline self-reports mask the gain. Without a measurement-validation step, neither result is interpretable. The measurement step is thus not a methodological footnote; it is a precondition for inference.

2. Research gap

This study addresses the following three important gaps in the existing literature, as defined above.

First, the EIPM measurement literature has converged on self-reported data as its standard and has rarely subjected this standard to within-study validation. Studies catalogue SDB as a residual limitation but do not redesign the instrument to identify it. As a result, many of the descriptive patterns that motivate EIPM interventions, such as consistently high motivation scores alongside lower capability and opportunity scores, may be a reflection of the limitations of self-reported measures rather than the underlying behaviour itself. This study attempts to untangle the two different influences on the final self-reported data.

Second, the survey-experimental toolkit for correcting SDB is well developed but has been applied almost exclusively to general-population samples in high-income countries on consumer-level behaviours. Application to elite, professional, LMIC samples is rare, and to our knowledge no published study applies face-saving designs to EIPM measurement in an African civil service.

Third, policymakers are a population in which SDB pressures are likely to be unusually intense; professional identity, perceived competence, and proximity to donor expectations all push in the same direction, yet they are systematically under-represented in the methodological literature on SDB correction.

The asymmetry between where SDB correction is studied (high-income countries, citizens) and where EIPM self-reports are most used (LMICs, civil servants) is the methodological gap this study addresses. Existing work establishes the building blocks for this study but has not combined them. Survey methodology research shows that self-reports of socially desirable behaviours are systematically inflated and that question design can mitigate this (Tourangeau and Yan, 2007). Face-saving preambles have been validated experimentally, but in general-population surveys on compliance behaviour rather than among public officials (Daoust et al., 2021a, who show guilt-free wording increases reported non-compliance across 12 countries). Within public administration, SDB has been documented in civil servants' self-reports of public service motivation (Kim and Kim, 2016), while the EIPM literature continues to rely on self-report instruments whose susceptibility to such bias remains untested (e.g., the SAGE tool; Makkar et al., 2016). This study connects these strands: to our knowledge, it provides the first experimental estimates of how civil-servant self-reports of evidence use respond to established SDB-mitigation techniques, in an active policymaking setting in a low-income country. Beyond the point estimates, it documents the magnitude and heterogeneity of question-design effects in a design that future EIPM studies in comparable settings can replicate directly.

3. Research questions

The study addresses two pre-specified research questions that measure the extent of SDB observed in policymakers in Ethiopia:

RQ1 (Normalizing preamble). Does a short normalizing preamble that legitimizes non-ideal answers reduce mean self-reported scores on three core COM-B EIPM items (**capability** to interpret quantitative evidence, **opportunity** to use evidence in daily work, and **motivation**/attitude toward evidence as an input to policy quality), relative to a direct-question control? To what extent does the magnitude differ across components?

RQ2 (Heterogeneity). Are the preamble effects larger among respondents for whom the normative pressure to appear evidence-informed is plausibly stronger, specifically, more senior officials, officials in technical roles, and officials with higher educational attainment, and smaller among respondents for whom the baseline pressure is weaker?

4. Hypotheses

Based on the RQs, we pre-specify the following directional hypotheses. The associated effect-size ranges are best-guess priors derived from the SDB-correction literature in adjacent domains; they are reported alongside the primary point estimates rather than treated as mechanical decision rules. Where no informative prior is available, we state that explicitly.

4.1 Primary hypotheses

H1 (Preamble reduces self-reported EIPM). Respondents assigned to the normalizing-preamble arm will report lower mean values on a pre-specified EIPM-COM-B index than those assigned to the direct-question control arm. Expected direction: negative. Expected magnitude: 0.15–0.35 standard deviations on the standardized index, informed by Daoust et al. (2021b) and Morin-Chassé et al. (2017). H1 is rejected only if the index-level estimate is non-negative or statistically indistinguishable from zero.

H1a-H1c (Component-level effects). The same negative effect holds at the level of each COM-B component item, capability (H1a), opportunity (H1b), and motivation (H1c). We expect the largest movement on the motivation component, which is the most normatively loaded; the smallest on capability, which is partly anchored to verifiable analytical practice; and an intermediate effect on opportunity, which depends on respondents' interpretation of contextual constraints. These differential expectations are stated for transparency but are not themselves the basis for primary inference.

4.2 Heterogeneity hypotheses

H2a (Seniority). Preamble effects are larger among senior officials (Senior Expert/Lead Expert and above) than among junior or expert-level officials, because the normative cost of admitting non-ideal evidence use is higher for those whose roles explicitly involve advising on policy.

H2b (Technicality of role). Treatment effects differ between respondents in technical or analytical positions (e.g., analysts, statisticians, monitoring-and-evaluation officers, researchers) versus respondents in managerial or administrative positions. Officials in technical roles engage with

quantitative evidence as part of their routine work and may therefore be better placed to absorb and apply the intervention content; at the same time, their baseline familiarity with evidence may leave less scope for improvement.

H2c (Education). Effects are larger among respondents with a postgraduate degree (Master's, PhD) than among those with a Bachelor's or Diploma, consistent with Clifford and Jerit (2015) and Morin-Chassé et al. (2017).

4.3 Outcome variables

All outcomes in the experiment are recorded on 1–5 sliders, anchored at semantic endpoints (e.g., “Not at all” / “To a very great extent”). The three outcome items are:

- **Y1 — Capability:** “How confident are you in interpreting and applying quantitative evidence (e.g., statistics, statistical significance, impact evaluations) to inform policy decisions?” (1 = Not at all confident, 5 = Extremely confident).
- **Y2 — Opportunity:** “To what extent does your work environment enable you to use research evidence in your daily tasks?” (1–5, Not at all / To a very great extent).
- **Y3 — Motivation:** “To what extent do you believe that using research evidence improves the quality and credibility of policy decisions in your directorate?” (0–5, Not at all / To a very great extent).

The estimation will use a pre-specified COM-B index defined as the simple unweighted mean of the three standardized questions, where each item is standardized using the mean and standard deviation of the direct (control) arm:

$$Index_i = (1/3) \cdot (z(Y1_i) + z(Y2_i) + z(Y3_i)).$$

This makes the index interpretable and minimizes the influence of item-specific variance. The secondary specification reports the three items separately. We pre-specify the index as the primary outcome because (i) the components measure facets of a single underlying construct (COM-B EIPM) and (ii) component-wise reporting otherwise inflates the family-wise error rate.

4.4 Relation to the main cRCT

The SDB experiment is embedded in the registration form of *Bridging the Gap*, a cluster-randomized controlled trial (cRCT) evaluating an EIPM intervention among approximately 350 mid-level policymakers in trade-related directorates of Ethiopian federal ministries and agencies. In the main trial, directorates are randomly assigned to a capacity-building arm, a capacity-building plus co-production and mentoring arm, or a control condition, to identify the effect of capacity-building relative to business-as-usual and the added effect of co-production and mentoring. The SDB experiment is methodologically and analytically separate from the cRCT in which the registration form is embedded. Random assignment to the preamble is performed at the individual level by the survey platform, independent of the directorate-level assignment to capacity-building, co-production or control arms in the main trial. We do not draw inferences about cRCT treatment effects in this pre-

analysis plan, nor do we interpret SDB-experiment estimates as informative about the main trial's primary outcomes.

5. Experimental design

To test the research questions, a short experimental module is embedded in the registration form administered to all individuals who agree to participate in the wider Ethiopia EIPM study. The registration form is otherwise a standard intake instrument collecting basic demographics, ministry and directorate affiliation, seniority, and a short set of attitudinal items on evidence use. Crucially, the SDB experiment is independent of the main cRCT treatment assignment: registration responses are collected before any cluster-level randomization to capacity-building, co-production, or control arms, and individual-level assignment to the SDB conditions described below is uncorrelated with any subsequent cRCT exposure. The two studies share data collection but are statistically and conceptually distinct.

5.1 Experiment design: normalizing preamble

The experiment randomizes individuals into two groups based on whether three matched EIPM items (capability, opportunity, and motivation) are presented in a standard direct form or preceded by a short normalizing preamble. The three items, drawn from the COM-B operationalization in the wider study, are identical in question and response format across two groups; only the presence and content of the preamble differ.³ The design follows the face-saving template of Morin-Chassé et al. (2017) and Daoust et al. (2021b), adapted to a measurement context in which the sensitive behaviour is *admitting limited use of or limited confidence in research evidence*.

The preamble legitimizes non-ideal answers without prescribing them. For the question on capability, the preamble notes that policy staff often work under time pressure and come from diverse academic and technical backgrounds, and that even experienced officials may find it challenging to interpret quantitative evidence. For opportunity, it notes that the ability to use evidence depends on data access, time constraints, internal processes, and leadership support, which vary across teams and periods. For motivation, it notes that policymakers place different levels of importance on research evidence depending on experience, priorities, and the context in which decisions are made. Each preamble is short (one or two sentences) and does not direct the respondent to any specific answer; it only signals that variation in answers is normal and that no answer is professionally compromising.

Respondents in the control (direct) arm answer the three items without any face-saving preamble; the questions are otherwise identical. All six items (three direct, three with the face-saving preamble) are measured on the same response scale used throughout the registration form: a six-point Likert-type rating ranging from 1 ("Not at all") to 5 ("To a very great extent"), or from 1 ("Not at all

³ The exact preambles and related questions can be found in Appendix A.

confident”) to 5 (“Extremely confident”) for the capability item. Responses are recorded as whole numbers only. The scale is presented as a slider with verbal anchors at the two ends and no visible numeric values or category labels. This presentation is intended to mitigate central tendency bias: with no salient midpoint or numbered categories on display, respondents are less likely to default to a middle response option.

Before the experimental items are displayed, all respondents answer four questions on evidence use, presented in identical wording and order regardless of treatment assignment (registration items 26–29): the extent to which research evidence influences their understanding of key policy issues, how often they personally seek research evidence in their work, how often evidence is used within their directorate when developing or revising policies, and the extent to which research evidence has influenced directorate policies over the past two years. These items are not part of the experimental manipulation. As they are administered before respondents encounter the experimental items, responses cannot be affected by treatment assignment, making them valid pre-treatment covariates.

5.2 Placement and flow within the registration form

The experimental module is positioned at a fixed point in the registration form, after the demographic, ministry and seniority block and after the four upstream attitudinal items. Direct items (capability, opportunity, motivation) appear at registration items 30–32 for respondents in the direct arm; the preamble items appear at the corresponding positions for respondents in the preamble arm.

The questionnaire went through multiple rounds of feedback checks and bench testing. Based on the feedback, the flow and understandability of the questions were tested and improved, and the terms and language were adapted to fit the Ethiopian policymaking context. This process also confirmed that each respondent sees only one arm of the experiment and that items 26–29 always appear before the experimental manipulation.

5.3 Randomization procedure

Randomization is implemented by the online survey platform (SurveyMonkey) on which the registration form runs. Assignment is at the individual level, with equal allocation probabilities.

The platform draws a uniform random number for each respondent and assigns the respondent to the direct arm (registration items 30–32 with no preamble) or to the preamble arm (preamble + same items) with probability 0.5 each. Respondents do not see both arms due to the ID-based nature of the platform.

We do not stratify the randomization by directorate or ministry: with two arms and an expected sample of ~350–400 respondents, stratification on the institutional variables of interest (ministry, seniority, education) provides little efficiency gain and complicates the survey form’s implementation logic. We will instead verify covariate balance ex post and adjust for stratification-relevant variables in the regression analysis.

6. Sampling and context

6.1 Setting

We use the cRCT registration cohort solely as our sampling frame. Eligible respondents are mid-level civil servants working in directorates with mandates related to trade, agriculture, labour and skills, industry, planning and development, finance, revenue and customs, investment, and standards. Formative consultations conducted by the implementing partners (the Ethiopian Economics Association and C4ED) identified ~70 directorates across these ministries as potentially eligible for the wider trial; the registration form is the entry point for individuals from these directorates.

Two contextual features of this setting are relevant for the SDB study. First, Ethiopia is in the middle of an active reform cycle, finalizing WTO accession, implementing AfCFTA commitments, and operationalizing the 10-Year Development Plan, in which the discourse of evidence-informed decision-making is institutionally prominent. The normative pressure to appear evidence-informed is plausibly elevated relative to a steady-state policymaking environment. Second, the participating institutions span a wide range of plausible evidence cultures, from explicitly analytical bodies (e.g., policy and strategy units, planning and development directorates) to operational and regulatory units (e.g., customs, licensing). This heterogeneity is useful for the heterogeneity analysis, as described in Section 7.2.

6.2 Target population and expected sample

The target population is mid-level officials engaged in policy analysis, technical advisory work, or the preparation of policy-relevant documents within these directorates. Individuals in purely administrative or support roles are excluded by design. Senior officials are eligible, and a small number are expected to register.

The expected sample size is approximately 350–400 individuals across ~70 directorates, consistent with the wider cRCT registration plan. Approximate per-arm samples are therefore 175–200 individuals. The sample is non-probabilistic: participation is voluntary, conditional on directorate-level institutional agreement, and recruited through formal letters of support from ministry leadership. This produces a sample that is plausibly positively selected on motivation and engagement, which we treat as a limitation rather than a target of inference.

7. Estimation strategy

Our default specification is an intention-to-treat (ITT) linear regression with directorate-clustered standard errors, deliberately kept parsimonious given the sample size and the descriptive nature of the question. We do not attempt to recover compliance-adjusted local average treatment effects: assignment compliance is essentially mechanical (the platform displays the assigned items) and non-compliance is operationally implausible in this design.

7.1 Primary specifications

The primary estimation equation is:

$$Index_i = \alpha + \beta \cdot Preamble_i + \gamma' X_i + \varepsilon_i,$$

where Index is the standardized COM-B index defined in Section 4.3, Preamble equals one for respondents assigned to the normalizing-preamble arm and zero otherwise, X is the pre-specified vector of covariates from Section 7.4, and standard errors are clustered at the directorate level using the cluster-robust standard errors. The same equation is estimated separately for the three individual components of the COM-B index as secondary outcomes.

7.2 Heterogeneity

We estimate three pre-specified heterogeneity contrasts for each primary outcome, by adding an interaction between the treatment indicator and a binary moderator to the primary specification:

- Seniority: Senior Expert or above vs. junior/expert (H2a).
- Technical role: respondents in technical or analytical positions (e.g., analysts, statisticians, monitoring-and-evaluation officers, researchers) vs. respondents in managerial or administrative positions, based on self-reported job function at baseline (H2b).
- Education: Master's or PhD vs. Bachelor's or Diploma (H2c).

Heterogeneity estimates are reported with multiple-testing corrections within the heterogeneity family. We treat heterogeneity results as exploratory and explicitly do not condition on them in any primary inference.

7.3 Secondary and exploratory analyses

We pre-specify the following secondary analyses, the results of which will be reported regardless of the primary findings:

- Unadjusted estimates (no covariates), to demonstrate that the primary effect does not depend on covariate adjustment.
- Robustness to alternative clustering: Besides the main specification with directorate-clustered standard errors, robustness checks use ministry-level clustered and respondent-level robust standard errors.
- Sensitivity to differential attrition/non-response in the experimental items, using Lee (2009) bounds.

Any analysis not described above will be flagged as exploratory in the final paper. We will pre-register a code repository containing the cleaned analysis dataset structure and the estimation scripts as part of the AEA RCT Registry submission.

7.4 Covariates

We record and pre-specify a parsimonious set of covariates collected at the start of the registration form, before any experimental item:

- Gender (binary)
- Age (years)
- Education level (Diploma / Bachelor's / Master's / PhD or equivalent)
- Ministry or agency (categorical, 12 institutions)
- Directorate (categorical)
- Seniority level (Junior Expert/Officer; Expert; Senior/Lead Expert; Desk Head/Team Leader/Department Head; Director/State Minister/CEO)
- Tenure in current ministry or agency (registration item 22; five bands: Less than 1 year; 1 to 3 years; More than 3 to 5 years; More than 5 to 10 years; More than 10 years). In estimation, tenure enters as four indicator (dummy) variables with “Less than 1 year” as the omitted reference category; each coefficient measures the difference in the outcome for that tenure band relative to staff with less than one year of tenure.
- Baseline evidence-engagement index: the unweighted mean of the four standardized upstream items (registration items 26–29: influence of evidence on policy understanding, frequency of personally seeking evidence, frequency of evidence use within the directorate, and influence of evidence on directorate policies over the previous two years). Combining the four items into a single pre-specified index preserves degrees of freedom and captures baseline evidence engagement as one construct; the index enters the estimation as a single continuous covariate.

The first seven items above are baseline characteristics, fixed before randomization, and may be used both for balance assessment and as pre-specified regression covariates. The four upstream attitudinal items appear in the same form in both arms but are administered after the consent question and before the experimental block. We pre-specify these covariates to improve precision because they capture baseline attitudes that are likely to be predictive of the outcome.

8. Statistical power

The experiment randomizes treatment at the individual level: each respondent is independently assigned, within the survey instrument, to the direct-question arm or the face-saving preamble arm with equal probability. The expected sample is 350 respondents across approximately 70 directorates (average cluster size ≈ 5); we present power calculations at three reference values to reflect uncertainty in the realized sample. The mean for the given variable is selected from the Mehmood et al. (2024) paper, where respondents were asked “How important do you think quantitative analysis is in public policy making?”. It is the closest study to our context that also asks about the importance of evidence in policymaking.

Minimum detectable effects (MDEs) are computed with the standard two-sample formula for a standardized outcome, assuming 80% power and a two-sided test at $\alpha = 0.05$. As treatment is randomized independently at the individual level, the intra-cluster correlation of treatment assignment is approximately zero, so clustering (from the larger cRCT) implies no meaningful variance inflation and no cluster design effect is required.

Given a sample of 350 (we also evaluate a smaller starting sample of 300 in rows 3 and 4 of Table 1), we expect to be adequately powered, as the MDEs for all samples lie between 0.18% and 0.20%. In SD units, this is a change of between 0.38 and 0.43 (a medium-sized effect given Cohen's definition). We also calculate for higher attrition (in the second and fourth rows) and find that our sample should still be able to detect effects, if any are present.

Table 1. Minimum detectable effects under alternative sample sizes

Outcome	Mean	SD	Total sample endline	Total sample baseline	Change in SD units	Target mean	% change
Ratings	2.745	0.013	231	350	0.38	2.75	0.18
Ratings	2.745	0.013	210	350	0.40	2.75	0.18
Ratings	2.745	0.013	198	300	0.42	2.75	0.19
Ratings	2.745	0.013	180	300	0.43	2.75	0.20

To benchmark these values: Morin-Chassé et al. (2017), pooling 19 wording experiments on reported voter turnout, find a treatment effect of 7.6 percentage points. Daoust et al. (2021b) find 2–3 percentage points for a preamble-only treatment and 8–11 percentage points for their strongest treatment combining a preamble with guilt-free response options.

9. Missing data and exclusions

Because the experimental questions are required and administered early in the registration form, missing data are expected to be minimal. Individuals who do not provide consent are automatically exited from the survey and are not included in the analysis sample. Duplicate submissions will be removed, retaining only the first valid response. We will conduct sensitivity analyses excluding implausibly fast responses, but otherwise expect little attrition or item non-response. No outcome imputation will be performed, and all exclusions will be reported transparently.

10. Ethical considerations

This study embeds a low-risk survey experiment within the registration form of the wider EIPM study. Participants provide informed consent before entering the experimental section, while the specific experimental manipulations are not disclosed in advance to avoid influencing responses, consistent with standard survey-experimental practice. The experiment involves no deception, poses minimal risk, and has no consequences for participants or their institutions. Ethical approval has

been obtained from the relevant national institutional review board (IRB) in Ethiopia, and all analyses will use de-identified data in accordance with established data-protection and confidentiality protocols. Participation does not affect access to any study benefits, and participants will be provided with a debrief after data collection is completed.

11. Limitations

We flag four limitations that are intrinsic to the design and that will frame the interpretation of results.

Selection. Participation is voluntary and recruitment passes through institutional gatekeepers. The sample is plausibly positively selected on motivation, openness to research engagement, and seniority of relationships with the implementing partners. The estimates therefore characterize the responsiveness of SDB to face-saving manipulations in this population of relatively engaged officials, not in Ethiopian policymaking writ large. We expect, if anything, the SDB inflation in this sample to be lower than in a fully representative frame, on the grounds that engaged respondents are likely to be both higher on actual evidence use and more comfortable acknowledging variation in their practice.

Single context. All estimates come from a single country, a single sector cluster (trade and adjacent ministries), and a single instrument. External validity to other LMIC civil services is a hypothesis, not an outcome, of this study. Where possible, we report estimates at the ministry level so that readers can locate the magnitudes within the institutional variation present in our sample.

Self-report against self-report. The study compares two self-report instruments to each other; it does not benchmark either against an external behavioural anchor. We can show that the preamble changes reporting, but we cannot show whether the change reflects movement toward or away from the latent construct.

Survey-platform inflation. Survey experiments operate in a stylized environment in which the treatment is more salient than in real-life information environments. The estimated effects of the preamble are likely to be upper bounds on what an equivalently worded item would produce in a real policymaking workflow.

We deliberately avoid two additional designs that the SDB literature uses, the list experiment (e.g., Kim and Kim, 2016) and the randomized-response technique, on the grounds that they are difficult to embed in a short professional registration form without raising the cognitive cost of the form to levels that would risk attrition. The face-saving preamble is the most defensible design given our implementation constraints.

References

Armstrong, R., Waters, E., Moore, L., Dobbins, M., Pettman, T., Burns, C., Swinburn, B., Anderson, L., & Petticrew, M. (2014). Understanding evidence: A statewide survey to explore evidence-

- informed public health decision-making in a local government setting. *Implementation Science*, 9(1), 188.
- Bernardi, R. A., & Nash, J. (2023). The importance and efficacy of controlling for social desirability response bias. *Ethics & Behavior*, 33(5), 413–429.
- Bursztyn, L., Haaland, I. K., Röver, N., & Roth, C. (2025). The social desirability atlas (No. w33920). National Bureau of Economic Research.
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.
- Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*, 79(3), 790–802.
- Daoust, J.-F., Bélanger, É., Dassonneville, R., Lachapelle, E., Nadeau, R., Becher, M., Brouard, S., Foucault, M., Hönnige, C., & Stegmüller, D. (2021a). A guilt-free strategy increases self-reported non-compliance with COVID-19 preventive measures: Experimental evidence from 12 countries. *PLOS ONE*, 16(4), e0249914.
- Daoust, J.-F., Nadeau, R., Dassonneville, R., Lachapelle, E., Bélanger, É., Savoie, J., & van der Linden, C. (2021b). How to survey citizens' compliance with COVID-19 public health measures: Evidence from three survey experiments. *Journal of Experimental Political Science*, 8(3), 310–317.
- de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11), 3266–3302.
- Gaines, B. J., Kuklinski, J. H., & Quirk, P. J. (2007). The logic of the survey experiment reexamined. *Political Analysis*, 15(1), 1–20.
- Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1), 33–48.
- Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, 47(4), 37–67.
- Grimmond, J., Brown, S. D., & Hawkins, G. E. (2025). A solution to the pervasive problem of response bias in self-reports. *Proceedings of the National Academy of Sciences*, 122(3), e2412807122.
- Kapiriri, L., & Bondy, S. J. (2006). Health practitioners' and health planners' information needs and seeking behavior for decision making in Uganda. *International Journal of Medical Informatics*, 75(10–11), 714–721.
- Kim, S. H., & Kim, S. (2016). Social desirability bias in measuring public service motivation. *International Public Management Journal*, 19(3), 293–319.
- Kovács, R., & Veselý, A. (2025). Survey experiments in public policy: A systematic literature review. *Review of Policy Research*, 43(2), e70040.

- Landry, R., Lamari, M., & Amara, N. (2003). The extent and determinants of the utilization of university research in government agencies. *Public Administration Review*, 63(2), 192–205.
- Langer, L., Tripney, J., & Gough, D. (2016). The science of using science: Researching the use of research evidence in decision-making. EPPI-Centre, UCL.
- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3), 1071–1102.
- Lilleker, D. G. (2003). Interviewing the political elite: Navigating a potential minefield. *Politics*, 23(3), 207–214.
- Makkar, S. R., Brennan, S., Turner, T., Williamson, A., Redman, S., & Green, S. (2016). The development of SAGE: A tool to evaluate how policymakers' engage with and use research in health policymaking. *Research Evaluation*, 25(3), 315–328.
- Mehmood, S., Naseer, S., & Chen, D. L. (2024). Training policymakers in econometrics [Working paper].
- Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*, 6, 42.
- Morin-Chassé, A., Bol, D., Stephenson, L. B., & Labbé St-Vincent, S. (2017). How to survey about electoral turnout? The efficacy of the face-saving response items in 19 different contexts. *Political Science Research and Methods*, 5(3), 575–584.
- Mutz, D. C. (2011). *Population-Based Survey Experiments*. Princeton University Press.
- Nduku, P., Ategeka, J., Madonsela, A., Mdlalose, T., Stevenson, J., Shisler, S., Pande, S., & Mahlanza-Langer, L. (2024). Protocol: What works to increase the use of evidence for policy decision-making: A systematic review. *Campbell Systematic Reviews*, 20(4), e1435.
- OECD. (2020). *Building Capacity for Evidence-Informed Policy-Making: Lessons from Country Experiences*. OECD Publishing.
- Oliver, K. A., & de Vocht, F. (2015). Defining 'evidence' in public health: A survey of policymakers' uses and preferences. *European Journal of Public Health*, 26(3), ckv082.
- Panagopoulos, C. (2010). Affect, social pressure and prosocial motivation: Field experimental evidence of the mobilizing effects of pride, shame and publicizing voting behavior. *Political Behavior*, 32(3), 369–386.
- Parkhurst, J. (2017). *The Politics of Evidence: From Evidence-Based Policy to the Good Governance of Evidence*. Routledge.
- Paulhus, D. L. (1991). Measurement and control of response bias. In Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Academic Press.
- Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282.

- Schuster, C., Mikkelsen, K. S., Rogger, D., Fukuyama, F., Hasnain, Z., Mistree, D., Meyer-Sahling, J., Bersch, K., & Kay, K. (2023). The Global Survey of Public Servants: Evidence from 1,300,000 public servants in 1,300 government institutions in 23 countries. *Public Administration Review*, 83(4), 982–993.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883.
- Uneke, C. J., Okedo-Alex, I. N., Akamike, I. C., Uneke, B. I., Eze, I. I., Chukwu, O. E., Otubo, K. I., & Urochukwu, H. C. (2023). Institutional roles, structures, funding and research partnerships towards evidence-informed policy-making: A multisector survey among policy-makers in Nigeria. *Health Research Policy and Systems*, 21(1), 36.
- Weiss, C. H. (1979). The many meanings of research utilization. *Public Administration Review*, 39(5), 426–431.

Appendix A: Exact wording of experimental items

A.1 Direct arm (control)

Capability (Y1). “How confident are you in interpreting and applying quantitative evidence (e.g., statistics, statistical significance, impact evaluations) to inform policy decisions?” Slider 0–5; anchors: Not at all confident / Extremely confident.

Opportunity (Y2). “To what extent does your work environment enable you to use research evidence in your daily tasks?” Slider 0–5; anchors: Not at all / To a very great extent.

Motivation / Attitude (Y3). “To what extent do you believe that using research evidence improves the quality and credibility of policy decisions in your directorate?” Slider 0–5; anchors: Not at all / To a very great extent.

A.2 Normalizing-preamble arm (treatment)

Capability preamble. “Policy staff often work under time pressure and come from different academic and technical backgrounds. As a result, even experienced officials may sometimes find it challenging to interpret quantitative evidence.” Followed by the capability item above, identical wording.

Opportunity preamble. “In many ministries, the ability to use evidence depends on factors such as access to data, time constraints, internal processes, and leadership support, which can vary across teams and periods.” Followed by the opportunity item above, identical wording.

Motivation / Attitude preamble. “Different policymakers place different levels of importance on research evidence depending on their experience, policy priorities, and the context in which decisions are made.” Followed by the motivation/attitude item above, identical wording.