

INFERENCE UNDER MEASUREMENT UNCERTAINTY: Evidence from surveys of tax compliance *

Yannick Markhof[†]

Stephan Dietrich

Rose Camille Vincent

Job Market Paper

ETH Zurich

UNU-MERIT

Maastricht University

UNU-MERIT

Utrecht University

October 27, 2026

[Click here for latest version](#)

Abstract

Empirical research neglects the role of measurement uncertainty—uncertainty arising from discretionary measurement choices between plausible alternatives—when drawing inferences. We show that this can result in fundamentally flawed inference and distort our understanding of economic behavior. In surveys with 3,400 business owners about their tax compliance, validated against administrative tax records, measurement uncertainty dominates sampling uncertainty. In a series of experiments and tests, we trace measurement uncertainty to easily missed biases from cognitive complexity, framing, and enumerator heterogeneity. We develop a Bayesian hierarchical framework that incorporates measurement uncertainty into estimation and offers full-information inference in many applied data settings. When researchers make discretionary measurement choices, conventional standard errors understate the uncertainty around an estimate. This changes how we should draw inferences, what pre-registration can deliver, and the returns to better measurement.

Keywords: Measurement error, Standard error, Survey data, Administrative data, Sensitive issues

JEL Codes: C81, C83, C91, H26, O12

*We gratefully acknowledge the financial support of ETH4D, ICTD, the International Growth Centre (IGC), the J-PAL Governance Initiative, STEG, and UNU-WIDER. We thank the Uganda Revenue Authority (URA), particularly Nicholas Musoke, and Kampala Capital City Authority (KCCA), particularly Ezra Ssebuwufu, for providing us with data access. We are grateful to our local partners, the survey data collection team, and particularly Fred Kasalirwe and Fred Matovu. We are thankful for comments and suggestions received from Lukas Hensel, Eleonora Nillesen, and Philip Wollburg; as well as seminar and conference participants at UC Berkeley, UC Davis, ETH Zurich, the Global Labor Organization, Oxford University/CSAE, ICDE, AFE Chicago, the United Nations University (UNU-MERIT), and the World Bank, among many others. The study was pre-registered in the AEA RCT registry under ID AEARCTR-0012298.

[†]Corresponding author. ymarkhof@ethz.ch

1 Introduction

In this paper, we reveal typically unobservable variation in respondent behavior between different survey measures of the same outcome. We then study what this implies for reliable inference with survey data and other scenarios in which researchers make discretionary choices over multiple plausible measurement approaches. To do so, we combine data from 3,400 interviews with business owners; firm-level administrative data from the registries of two tax authorities that these entrepreneurs are liable to; and a series of survey experiments. This setup allows us to pull back the curtain on respondent behavior and measurement error in ways that are usually impossible (Celhay et al., 2024; De Weerd et al., 2020; Bound et al., 2001). Based on it, we quantify the extent, nature, and implications of a source of inferential uncertainty that is unaccounted for in conventional inference. We call this *measurement uncertainty*, formally distinguish it from *sampling* uncertainty captured in conventional standard errors, and show how to account for it in applied research.

The example running through our study grapples with a common challenge for applied research, measuring sensitive behavior or attitudes in surveys (Bursztyn et al., 2025). Studying sensitive issues—behavior that is illegal, socially stigmatized, or viewed as private and where respondents might therefore hesitate to provide accurate responses—provides an ideal backdrop for studying measurement uncertainty: There is typically no benchmark to ascertain the accuracy of any single measure and economists use a plethora of different measurement approaches for this task (Bursztyn et al., 2025; Crépon et al., 2025). We take this measurement challenge to the context of tax non-compliance in Kampala, Uganda, where we conduct surveys with nearly 3,400 small- and medium-sized businesses and can validate responses against linked administrative records on the same firms.¹ We field several popular measures of sensitive behavior—among them indirect response options such as list experiments and randomized response techniques, incentivizing truthful answers via the *Bayesian Truth Serum* of Prelec (2004), and a simple direct question—and show that these measures suggest vastly different tax non-compliance rates. Starting from this observation, we study whether and why this dispersion does (not) correlate with differences in accuracy, how hard-to-observe measurement errors and common decision heuristics can mislead, and implications for accurate inference.

We have three main results. First, we demonstrate a typically neglected but highly consequential source of inferential uncertainty. We document large *aggregate* differences between the estimates of plausibly equivalent, and at the *respondent level* equally accurate, survey measures of the same target variable. This causes measurement uncertainty (variation *between* multiple plausible measures) to vastly exceed sampling uncertainty (variation *within* a single measure). Responsible for this is the sensitivity of respondent behavior to survey measurement choices. We show that this is easily missed in applied research, gives rise to pervasive and differential measurement errors across a wide range of survey tasks, and purports a fundamentally flawed understanding of the accuracy of any single measurement approach. Single-measure inference

¹Our study context resembles that of recent research aiming to deploy indirect questioning techniques to inform policy on tax evasion (see, e.g., Hoy et al., 2024; Genest-Grégoire et al., 2022; Iraman et al., 2022).

neglects this entirely and results in overconfident inference. Second, we provide experimental and descriptive evidence on the reasons and mechanisms behind measurement uncertainty. We find evidence supporting at least three distinct channels: (i) cognitive complexity is associated with non-strategic errors that scale with respondent burden; (ii) even small variations in question phrasing can activate different response heuristics; (iii) variation in survey implementation between enumerators. Third, a Bayesian hierarchical model, complemented by priors which we collect through the Social Science Prediction Platform, allows us to explicitly account for measurement uncertainty in inference and mitigates the risk of overconfident research conclusions.

To evaluate the performance of different survey techniques measuring the same target parameter, we compare respondents’ reported business tax non-compliance (in aggregate and individually) to each firm’s true compliance status as observed in administrative tax records. To do so, we obtain access to national and municipal business tax records from the Uganda Revenue Authority (URA, national tax authority) and the Kampala Capital City Authority (KCCA, municipal authority), which we cross-link based on each firm’s unique tax identifier. From the resulting merged registry of over 150,000 unique firms in Kampala, we sample active firms for inclusion in our survey and track them down in an extensive fieldwork effort. For these firms, we thus observe various measures of self-reported tax non-compliance through our survey and, unbeknownst to respondents, true tax non-compliance based on official records. Using partition validation—splitting the sample according to our firm-level non-compliance benchmark and estimating tax non-compliance among true non-compliers and true compliers—we are able to observe how different measurement approaches perform in terms of their true positive rate, false positive rate, and accuracy.²

We show that different measures of tax non-compliance suggest vastly diverging tax non-compliance rates; that, in spite of this, all measures are similarly (in)accurate when validated against the objective benchmark; and that there is therefore considerable measurement uncertainty that comes with picking any single measure, a finding we replicate across two separate surveys. Different measures have different failure modes. The list experiment (LE) and randomized response technique (RRT) substantially oversell their ability to induce truthful reporting as they result in a large increase in non-strategic errors which aggregates mask. Incentivizing truthful answers, as in the Bayesian Truth Serum approach of [Prelec \(2004\)](#), does not meaningfully improve accuracy compared to a non-incentivized direct question. While the LE and RRT increase true positives (decrease false negatives) by 7–11 percentage points (LE) and 20 percentage points (RRT) compared to a simple direct question, they result in an equally large increase in *false* positives (7–11 and 21 percentage points, respectively). As a result, neither approach is any more accurate in terms of overall classification accuracy than the direct question and no technique outperforms a random guess by more than 10 percentage points. This pattern is consistent with an increase in non-strategic misreporting induced by the additional respondent burden of both approaches.

²The true positive rate is the estimated non-compliance rate among tax non-compliers; the false positive rate is the estimated tax non-compliance rate among compliers; and accuracy is the overall correct classification rate.

These results are not driven by issues with the administrative tax registry benchmark we are using. We confirm that they occur in a number of response quality checks that do not use this benchmark at all and in an application using an exogenous, unambiguously observable, and non-sensitive benchmark. Our findings are also consistent across both types of business taxes firms are liable for with the benchmark coming from separate tax authorities’ registries.

Crucially, the failure of common survey techniques to reduce misreporting would have typically been unobservable to applied researchers studying sensitive behaviors. True and false positives are indistinguishable in the absence of an individual-level benchmark, the lack of which motivates the use of these survey techniques in the first place. Further, they both increase the estimated rate of sensitive behavior in line with the common research practice of taking higher estimates of sensitive behavior as evidence of the degree to which any technique was “successful” in inducing truthful reporting (see, e.g., [Chuang et al., 2021](#)). In our case, this would have led us to believe that both the list experiment (+7–11 percentage points higher prevalence of tax non-compliance) and randomized response technique (+20 percentage points) successfully reduced the amount of misreporting in the data compared to the direct question and that the latter was more successful than the former in doing so—whereas the opposite is actually the case. We show that this issue occurs despite passing common internal validity checks for indirect survey methods ([Chuang et al., 2021](#); [Blair et al., 2019](#); [Blair and Imai, 2012](#)).

This first set of findings demonstrates a general empirical challenge: Unobservable variation in the size and sources of measurement error can lead to large disparities between different measures of the same target variable, even without meaningful differences in accuracy—or with no reliable way to adjudicate—between them. In a series of experiments, we study why this occurs in survey data and how it leads to measurement uncertainty.

Simple inversions in question phrasing lead the same question to be perceived and therefore answered very differently by respondents. Randomly asking about tax compliance instead of *non*-compliance produces estimates of the tax non-compliance rate that differ by 47 percentage points in a list experiment and by 9 percentage points (25%) when eliciting respondents’ second-order beliefs. We present evidence that such differences are brought about by changes in response behavior reflecting different forms of non-strategic misreporting rather than more accurate reporting under any one question phrasing.

We also find that enumerators differ in the way they administer an otherwise identical survey. Enumerator effects explain just 9% of variation in data quality in a test that gave enumerators little opportunity to deviate from protocol but as much as 50% of variation where they had a much more prominent role in safeguarding data quality. This highlights that measurement-as-designed may differ from measurement-as-implemented and has implications for real survey design choices that researchers face: While implementation fidelity is hard to observe and enforce, different survey approaches vary in their exposure to implementation heterogeneity.

Our results challenge the validity of inference across a wide variety of cases where measurement accuracy is largely unobservable, multiple plausible approaches exist, and no single measure can be conclusively identified as the most accurate. In these cases, the choice of measurement approach is ambiguous, yet can be highly consequential. This uncertainty is unaccounted for

when researchers make a discretionary choice over multiple plausible approaches, leads traditional confidence intervals to understate true inferential uncertainty, and results in overconfident research and policy conclusions (Manski, 2011).

We propose an alternative approach that provides full-information inference and honest uncertainty quantification given the available evidence base. Instead of betting on a single measure, we take multiple measures, along with prior information, and use the extent of disagreement between them to quantify measurement uncertainty, not just sampling uncertainty.

To implement our approach, we use a Bayesian hierarchical model (BHM) and adapt the framework of Bayesian meta-analysis (Meager, 2019; Bandiera et al., 2021; Vivalt, 2020) to the problem of pooling multiple measures within a single study.³ Our approach shares the spirit of Bayesian model averaging (Chatfield, 1995; Draper, 1995) according to which uncertainty over the correct specification should propagate to inferential uncertainty rather than be assumed away. It also connects to established bounding approaches for survey inference under measurement error (de Quidt et al., 2018; Lee, 2009) and recent work on including extraneous prior information in inference (Iacovone et al., 2025). We demonstrate how to generalize these insights to observational and descriptive estimates of hard-to-observe behavior and attitudes, quantities that are ubiquitous in economic research and essential inputs to policy.

The BHM reveals four core findings about inference under measurement uncertainty. First, total uncertainty is grossly understated in a single measure’s standard error and on average 2.4–2.9 times larger when aggregating what the evidence base really implies in our setting. Second, much of the adjustment in overconfident inference is already achieved by collecting one or two additional measures of the outcome of interest. This matters as the set of plausible measurement design variations is potentially large but space in surveys is limited. Third, like any aggregation technique, the BHM cannot promise unbiasedness by turning poor measurement into good measurement. Instead, it provides full-information inference and quantifies uncertainty given all available evidence. Fourth, the BHM communicates inferential uncertainty in a form directly usable for decision problems which we demonstrate in the context of tax enforcement policy. For example, we estimate, given data and priors, that a cautious policymaker would deploy additional resources to municipal tax enforcement for break-even rates below 14 percent non-compliance and to enforcement of national business tax for break-even rates up to 22 percent.⁴

Our paper makes three substantial contributions to survey-based inference of economic and behavioral parameters. First, we provide insights into measurement error in surveys that is typically unobservable but highly relevant for reliable inference. In a series of experiments, we identify the extent and sources of resulting measurement uncertainty in survey data. We show that respondent behavior is sensitive to even small changes in survey design and that this can substantially alter the prevalence estimates of behavior and attitudes in surveys, the conclusions drawn about the validity of different measurement approaches, and lead to overconfident policy advice (Manski, 2011). As such, we demonstrate that measurement uncertainty is a quantita-

³In meta-analysis as well as in our case, modeling heterogeneity guards against false precision: about effects in new contexts for meta-analysis and about the estimate itself in our case.

⁴The break-even rate is the tax non-compliance rate above which enforcement starts paying for itself.

tively important source of uncertainty in survey estimates which receives a disproportionately small amount of attention from economists (Chatfield, 1995). Relative to the principal error paradigm in survey methodology, the total survey error (TSE) framework (Biemer, 2010), measurement uncertainty sits one level above. Where the TSE concerns the sources of survey error (bias and variance) *conditional on a chosen survey design*, measurement uncertainty describes the inferential uncertainty inherent in the choice among multiple plausible designs in the first place. We formalize this relationship in a hierarchical model of inference under measurement uncertainty.

Second, we demonstrate that shifting from a single measure to multiple measures of a variable whose accuracy is hard to verify yields a more faithful reflection of the true inferential uncertainty around a point estimate. We formalize measurement uncertainty as the between-measure variance of latent measurement biases and show that conventional single-measure inference is the special case that assumes this variance away. The resulting framework is simple and estimable, and it extends beyond surveys to any setting where researchers choose among plausibly equivalent measurement or analytical approaches. Our approach makes already established heuristics in the evaluation of causal inference in economics explicit and unifies ideas from the literatures on Bayesian evidence aggregation (Meager, 2019; Bandiera et al., 2021; Vivalt, 2020) and bounding approaches to survey estimation (de Quidt et al., 2018; Lee, 2009). It also allows for a way to incorporate non-standard, but potentially valuable, data such as expert judgments, qualitative insights, or unstructured data into the estimation process (Iacovone et al., 2025). As such, we provide a principled empirical framework to address two frontier issues in the production of credible research: “dueling certitudes” in policy analysis (Manski, 2011) and the “humans-in-the-loop” issue (Stevenson and Fischman, 2026).⁵ An extensive appendix offers practical implementation guidance to applied researchers.

Third, we provide the most comprehensive validation study of popular survey tools for a common issue in economics research, collecting accurate data on sensitive issues (Bursztyn et al., 2025). Over the last ten years, the tools we validate have been used over a dozen times in the top economics journals and hundreds more times outside these journals.⁶ Compared to most previous evidence, we validate techniques against a respondent-level, real-world benchmark (Celhay et al., 2024; Bound et al., 2001). We achieve this by gaining access to the tax records of two independently acting tax authorities that firms in our survey are liable to, cross-linking

⁵Dueling certitudes describes “contradictory predictions made with alternative assumptions” (Manski, 2019) where each prediction is plausible under its own set of assumptions. The resulting uncertainty is seldom recognized and instead hardens into an all-or-nothing choice between competing specifications. Similarly, uncertainty from “humans-in-the-loop” arises when researchers make discretionary choices that substantially impact findings but that are unaccounted for in the existing econometric paradigm (Stevenson and Fischman, 2026).

⁶We use the tool described in Goldsmith-Pinkham (2026) available at <https://paulgp.com/econlit-pipeline/search.html> (accessed April 28, 2026) to search for articles using either list experiments, the randomized response technique, or the Bayesian Truth Serum in the following journals: *American Economic Review*, *Quarterly Journal of Economics*, *Econometrica*, *Journal of Political Economy*, *Review of Economic Studies*, *AEJ: Applied Economics*, *AEJ: Economic Policy*, *AEJ: Microeconomics*. The search terms were: “list experiment” or “item count technique” or “list randomization” or “randomized response technique” or “randomized-response technique” or “bayesian truth serum”. We subsequently manually validate each match.

both at the firm level, and tracking down firms in an intensive fieldwork effort. In this context, we show how and why survey tools break down when validated at the respondent level. Such issues are typically not observable to researchers and make the ‘more is better’ assumption—the common heuristic to take higher estimated prevalence of sensitive behavior to indicate more valid measurement—untenable in an applied setting.⁷

Measurement uncertainty arises wherever researchers choose among multiple plausible measurement approaches, and its consequences are not specific to our application. Our paper adds to a long tradition in economics that points to a disconnect between the seeming ubiquity of measurement errors in (survey) data and their often too cavalier treatment (see, e.g., [Bound et al., 2001](#); [Manski, 2011](#); [Stevenson and Fischman, 2026](#)). Similar examples of substantial, non-strategic misreporting can be found in contexts very different from ours, such as simple questions on social assistance receipt in the United States ([Celhay et al., 2024](#); [Meyer et al., 2015](#)) and foundational topics such as the elicitation of economic preferences ([Gillen et al., 2019](#)). Specifically for sensitive behaviors, evidence of non-strategic reporting errors is emerging from a small number of other validation studies with individual-level benchmarks, for both the list experiment ([Kuhn and Vivyan, 2022](#)) and the randomized response technique ([Höglinger and Jann, 2018](#); [Höglinger and Diekmann, 2017](#)).⁸ The mechanisms we identify as giving rise to measurement uncertainty—cognitive complexity, framing effects, and implementation heterogeneity—operate across many contexts (see, e.g., [Artinger et al., 2022](#); [Bertrand and Mullainathan, 2001](#); [Oprea, 2024](#); [West and Blom, 2016](#)) and relate to behavioral phenomena widely studied by economists. Not least, the issue measurement uncertainty describes, and the solution we develop, extend naturally to other settings such as choosing between different remotely sensed earth observation products ([Josephson et al., 2026](#)) or data analyzed independently by separate research teams ([Menkveld et al., 2024](#)). The fundamental problem measurement uncertainty speaks to is therefore widely relevant for economists and firmly anchored in the mainstream economic literature, yet rarely analyzed systematically. Standard empirical practice lacks flexible, inexpensive tools that extend to field settings where different measurement approaches suffer from fundamentally different biases rather than independent random error. Unlike methods that correct measurement error under structural assumptions on its form, the tool we provide quantifies the uncertainty it creates between plausible measurement choices. It rests on weaker assumptions, uses the available data efficiently, and reports what the available evidence collectively implies.

Our findings have three general implications for empirical economics. First, where researchers make discretionary choices over several plausible measurement approaches, conventional standard errors overstate how much a study has pinned down. Such choices are pervasive

⁷For example, in a context similar to ours, a recent study concluded that a double list experiment provided “arguably [...] among the best estimates of the levels of tax evasion by firms in Indonesia” and proposed this approach as a tool to inform tax policy in developing countries ([Hoy et al., 2024](#)). Similarly, another study concludes based on the ‘more is better’ assumption that estimates from a list experiment allow for “the first analysis providing prevalence rates arguably net of response bias of men’s GBV-related opinions and behaviors in West Africa” ([Bertelli et al., 2025](#)).

⁸This evidence comes from online or mail-in surveys in high-income countries and does not study a real-world sensitive behavior. Still, this indicates that the patterns we document are not artifacts of our specific setting.

in social science. When each is made with unwarranted confidence, the disagreements that follow harden into the dueling certitudes [Manski \(2011\)](#) criticizes. Committing to one measurement approach then reveals a single point in the range of results the study could plausibly have produced, similar to how a single experimental site reveals one point in the distribution of program impacts found elsewhere ([Vivalt, 2020](#)). Second, quantifying measurement uncertainty should be part of the default empirical toolkit where it closes a gap in pre-registration practice. Large between-measure variation for the same target variable showcases the scope for substantial specification searching in empirical work and directly relates to the concerns raised by [Christensen and Miguel \(2018\)](#). While pre-registration eliminates specification searching by committing to a measure ex-ante, committing to a single measure is precisely what keeps measurement uncertainty out of view. The approach we propose in this paper is therefore complementary to pre-registration: Researchers should pre-specify their main measurement approach, plausible alternatives as robustness checks, and the rule for aggregating them such that measurement uncertainty is accounted for in inference. Third, the returns to better measurement are larger than identification-focused practice recognizes, and our framework is what makes them visible. Standard inference quantifies but a fraction of what poor measurement costs and instead assumes much of the cost away ([Dillon et al., 2020](#)). Once measurement uncertainty is quantified, better measurement carries a clear payoff. Validated, best-practice measures are in this sense a public good for inference. Each one shrinks the choice set of plausible measures and therefore the measurement uncertainty every subsequent study faces. Mapping measurement uncertainty across topics and settings would then direct methods research to where its return is highest, improving the reliability, comparability, and policy relevance of empirical work in economics and beyond ([Manski, 2019](#); [Meyer et al., 2015](#); [Stevenson and Fischman, 2026](#)).

2 Survey and administrative data

We draw on two distinct data sources: administrative tax records obtained from municipal and national tax authorities and two surveys covering a total of 3,400 firms spread across Kampala.

2.1 Administrative data

Our first data source is administrative tax records at both levels at which businesses in Kampala are liable for business tax: the municipal and national levels. First, businesses must obtain a trade license from the Kampala Capital City Authority (KCCA), with the amount to be paid depending on the location within Kampala and the industry of the business.⁹ The licenses are due for renewal every 12 months (typically in January, as licenses used to be issued by the calendar year). Firms with multiple physical or business branches (for example, a restaurant with multiple locations in Kampala or a hotel with a beauty salon on its premises) are required to obtain a separate trade license for each branch. Second, every business is subject to national

⁹For example, retailers, who make up the largest group of our sample, would be required to pay up to UGX 210,000 (USD 57) in annual trade license fees.

business tax levied by the Uganda Revenue Authority (URA).¹⁰ The amount of tax due is based on businesses' turnover. Firms below a turnover of 150 million Ugandan shillings (USD 40,000) are subject to less strict reporting requirements and declare presumptive taxes that use turnover estimates to calculate the tax burden. Firms above the threshold declare corporate income tax and are subject to more stringent reporting requirements. Turnovers are usually declared per fiscal year (July–June), and payment is due by the end of the calendar year.

Data can be linked between both tax levels (municipal and national) at the firm level through a unique identifier, the Taxpayer Identification Number (TIN). For a subset of firms, no TIN was documented in KCCA's records. In these cases, we retrieve the TIN by matching the firm name and contact details (phone number and email address) to the universal TIN registry. As a result, all surveyed firms are uniquely identified by their TIN in KCCA's and URA's records, thus facilitating the cross-record link. For more information on the administrative tax data and extensive documentation of the cross-record linkage exercise see [Dietrich et al. \(2026\)](#).

2.2 Firm survey data

Our second data source comes from two surveys of businesses in Kampala that we conducted in September–October 2023 and November–December 2024. The surveys targeted firms with a TIN and a turnover between UGX 20,000,000 (approximately USD 5,400) and UGX 300,000,000 (approximately USD 80,000).¹¹ Firms were sampled at random within Kampala from the list of businesses that had paid for a KCCA (municipal) trade license at least once since 2019 (2023 survey) or since 2020 (2024 survey).¹²

As the exact business address was typically not available and often dozens of businesses cluster in roughly the same location, we used the help of local leaders from the village chairman's office to track down businesses sampled for our study. At the beginning of each interview, enumerators confirmed again that the business matched the one sampled and corrected any inaccuracies regarding firm name, contact details, and location. Due to occasional inaccuracies in location information, refusals (primarily from employees who lacked the authorization to speak on behalf of the business), and non-responses (often occurring when business owners were unreachable despite repeated attempts via in-person visits, email, and phone calls), we successfully interviewed 38% of initial survey targets in 2023 and 55% in 2024.¹³ As we had

¹⁰We focus on tax filings in the case of national business tax since every business has to *file* tax to the URA, but the smallest businesses (below 10 million shillings (about USD 2,700) turnover) do not need to *pay* a positive amount. Either way, we estimate based on past turnover and businesses' characteristics that all firms in our sample exceed the exemption threshold by at least a factor of two.

¹¹We obtain turnover estimates for each firm from recent tax filings. In case a firm did not recently file presumptive or corporate income tax, we use information from the KCCA trade license data on the trade license fee charged to the firm and the firm's grade, both of which are positively correlated with turnover size.

¹²We make this restriction for several practical reasons. First, the KCCA data is the source of detailed contact information (phone numbers, email, village/street name) of the firms in our sample, which we needed to track down the firms for interviews. Second, records where no KCCA trade license had been obtained in recent years are more likely to pertain to inactive firms. Sampling these firms would have sharply increased the rate of non-response to our survey. These restrictions mean that we do not sample firms that are completely informal; that is, they have never obtained a trade license from the authorities.

¹³For comparison, the most recent World Bank Enterprise Survey in Uganda achieved a response rate

expected relatively high non-response rates based on the survey pilot, we compiled an extensive replacement list alongside our primary target sample. This allowed us to substitute businesses when interviews were not feasible and reach our desired sample size.

The final sample consisted of 3,383 businesses (1,460 in 2023 and 1,923 in 2024)¹⁴, with 50% categorized as retail shops, 11% engaged in wholesale trade, 8% boutiques and beauty salons, 5% food and hospitality, and the remaining businesses distributed across various other industries (Table A.1). These businesses were geographically dispersed across all five divisions of Kampala. Most businesses were small-to-medium in size, with a median of two paid employees and a median reported annual sales slightly exceeding UGX 50 million. Our respondents primarily included business owners (58%) or employees who are knowledgeable about the tax affairs of the business. In 87% of cases, we interviewed the person who reports being the most knowledgeable person about the tax matters of the business. In the remaining cases, we interviewed another person affiliated with the business (typically another employee) but verified through self-assessment and a knowledge questionnaire that they were sufficiently familiar with tax matters to provide valid responses (see Section 3.2.3).

Interviews were conducted in person via computer-assisted personal interviewing (CAPI) and collected data on a number of business characteristics, practices, attitudes, and knowledge of the tax system. Interviews were framed to respondents as collecting information on business practices and the economic climate for firms in Kampala. In 2023, participants received a compensation of UGX 26,000 (control group for Bayesian Truth Serum), or an amount of UGX 22,000 with a further variable amount of UGX 5,500 on average (treatment group for Bayesian Truth Serum). In 2024, where the questionnaire was shorter, all respondents received UGX 10,000.

2.3 Prediction survey

To construct an informative, independent prior for the Bayesian hierarchical model, we run a survey on the Social Science Prediction Platform (SSPP) and collect predictions of the tax non-compliance rate (DellaVigna et al., 2019). Access to our prediction survey was public and participation anonymous. Respondents were experienced academic predictors, familiar with the Sub-Saharan African context, or had relevant research experience. Among the 36 respondents, 87% had previously used the SSPP for research, 67% had conducted research or lived in Sub-Saharan Africa before, and 31% had conducted research on tax non-compliance and informality. After receiving contextual background information on business taxation in Kampala and our survey sample, each expert¹⁵ provided a point prediction of the non-compliance rate for KCCA and URA. In addition, experts provided a measure of their uncertainty for each prediction.

of 41% for a final sample size of 600 businesses (World Bank, 2013).

¹⁴In total, we conducted 3,413 interviews but in 30 cases (19 in 2023 and 11 in 2024), the business interviewed did not match the one sampled and could not be unambiguously identified in the administrative data. We are therefore unable to link these businesses to our benchmark measure and exclude them from the analysis.

¹⁵We use ‘expert’ to collectively refer to respondents of the prediction survey as a descriptive shorthand without intending this as a claim on their actual expertise.

Specifically, experts reported a value v such that they believed with 90% certainty that the true non-compliance rate would fall within $\pm v$ percentage points of their prediction. We interpret this as the half-width of the expert’s subjective 90% credible interval. Table A.2 reports descriptive statistics on the predictions made by the prediction survey’s respondents.

3 Experimental design

Our survey experiments cover different survey techniques to measure sensitive issues, allow us to quantify measurement error at the respondent level, and investigate mechanisms. Each experiment is short and embedded within the broader questionnaire. Experiments are spread across the length of the questionnaire amid other questions surrounding business practices, knowledge, and characteristics.

When testing different ways to measure tax non-compliance, the ordering of experiments within the questionnaire is such that we field those techniques first that make the issue of interest (tax non-compliance) least salient and those that preserve the anonymity of individual answers; and move to direct questioning last. This safeguards that, even if respondents were able to draw links between our different experiments, they do not have any incentives to adjust their answers to be consistent with information they provided previously. In the 2023 survey, we randomize the order in which we field the first two experiments (list experiment and randomized response technique) as both, in theory, provide equal privacy protection and salience of the matter of interest. In the 2024 survey, we only conduct a list experiment. In effect, these measures mitigate concerns of order and spillover effects between different measurement techniques. Enumerators walked each respondent through an innocuous example before the start of each technique to help respondents understand the instructions.

Table 1 provides an overview of our survey experiments. We next detail each of them.

3.1 Measures of tax non-compliance

3.1.1 Administrative benchmark

The sensitive behavior of interest for our study is tax non-compliance defined at the extensive margin.¹⁶ We define this based on whether

- (i) the firm paid for at least one KCCA trade license in the previous year (municipal level)¹⁷

¹⁶This is consistent with most applied research on sensitive behaviors that is concerned with *whether* the respondent engaged in the behavior or holds a certain attitude. Further, intensive margin tax compliance is much more complicated to measure. While we observe the amount of tax paid by each firm, we cannot audit firms’ books to ascertain whether their tax declaration was accurate in the first place. Similarly, while we observe the number of trade licenses a firm paid for, we do not observe whether there are any additional branches unknown to the tax authorities that the firm did not obtain a trade license for or any previously active branches that have closed.

¹⁷We take this conservative definition of tax compliance because it is much harder to ascertain the correct number of trade licenses a firm should have obtained (see footnote 16). Among firms in our sample, 92% report having only a single branch.

TABLE 1: Overview of survey experiments

Type	Experiment name	Sample size	Tax level
2023 survey			
Survey measure	Double list experiment (LE)	1,458	Municipal or national (50/50)
Survey measure	Randomized response technique (crosswise, RRT)	1,458	Municipal or national (50/50)
Survey measure	Bayesian Truth Serum (BTS)	1,320	Both (random order)
Survey measure	Direct question (DQ)	1,569	Both (random order)
2024 survey			
Survey measure	Double list experiment	1,741 (90% sample)	Municipal or national (50/50)
Survey measure	Direct question	1,907	Municipal or national (50/50)
Survey measure	2 nd order belief	1,718	Municipal or national (50/50)
Survey measure	Enumerator guess	1,923	Municipal or national (50/50)
Validity test	Dummy list experiment	1,909	N/A
Quantify error	Consistency check	1,864	N/A
Quantify error	Recall check	1,923	N/A
Mechanisms	Mental models (framing)	1,922 (90/10 split)	Municipal or national (50/50)
Mechanisms	Implementation fidelity (enumerators) [†]	1,923	N/A

Note: Overview of survey measures of tax non-compliance and experiments to isolate non-strategic measurement error and identify mechanisms. [†] No explicit randomization of enumerators to interviews.

- (ii) the firm filed presumptive or corporate income tax to the URA in the previous year (national level)¹⁸

For our main analysis, we pool results across tax levels for higher power and because the distinction between different tax types is secondary to the methodological questions we are attempting to answer. Throughout, we ascertain the robustness of our results to conducting the analysis separately for municipal and national business taxes.

3.1.2 Double list experiment (LE)

In a list experiment (or “item count technique”, [Raghavarao and Federer \(1979\)](#), [Miller \(1984\)](#)), respondents are presented with a list of behaviors or attitudes and are asked to provide the number of items they agree with, or that apply to their business. Businesses are randomized into two groups: one for which the list comprises four baseline statements (short list group); and one for which the list comprises the same four baseline statements plus a fifth statement containing

¹⁸In each survey, we therefore ask about tax non-compliance for the year with the most recent tax filing deadline: 2022 for the survey in 2023 and 2023 for the survey in 2024.

the behavior or attitude of interest (long list group). In our case, this additional statement is about whether the business did not file national business tax (did not pay for a trade license) in the previous year. By only asking respondents to reveal the number of items that are true for their business—but not their answer to any individual item—the list experiment introduces random noise into the measurement process. This makes individual tax non-compliance status unobservable and safeguards anonymity while still allowing us to back out an estimate of the aggregate tax non-compliance rate as the simple difference in item counts between the long list and short list group. In a double list experiment, this process is performed twice, each time for a different list of base items and switching which respondents get the short list and which get the long list. In order to reduce respondent burden and limit the scope for strategic behavior in response to the experiment, we only implement the double list experiment for one of two tax levels for each respondent. We provide more detail on the design of the list experiment in Supplementary Appendix C.1.

3.1.3 Randomized response technique (RRT)

In an RRT (Warner, 1965), respondents are presented with the sensitive statement of interest and a randomization device. Respondents are instructed to adjust their answer to the outcome of the randomization device (such as telling the truth if they roll a six with a die and lie otherwise). The randomization device’s outcome remains hidden to the experimenter but its probability distribution in aggregate is known. This makes individual behavior unobservable to the researcher while allowing the researcher to form an aggregate estimate of the behavior of interest.

Different variants of the RRT exist depending on the choice of the randomization device. We implement a crosswise design which has distinct advantages in terms of comprehensibility and ease of administration in our context (see Supplementary Appendix C.2). Respondents are presented with two statements, one of which asks whether their birth month is in May, June, or July and one asking about tax non-compliance. They are then asked whether either 0 or 2 (no or both) statements are true *or* one statement is true.¹⁹ This follows the canonical example described by Heck et al. (2018) in which the statement on the birth month of the respondent functions as the “randomization device” with known aggregate probability and unobserved private outcome.

We then back out an estimate of the tax non-compliance rate according to

$$s = \frac{r + p - 1}{2p - 1} \quad (1)$$

where s is the estimated share of respondents whose business did not file tax (pay for a trade license), r is the share of respondents who report that 0 or 2 statements are true, and p is the probability of being born between May and July.

Similar to our approach for the double list experiment, we conduct the RRT only once with

¹⁹To facilitate compliance with the experiment’s instructions, enumerators presented respondents with a handout and two cards of different colors (pink and green). They instructed respondents to show the pink card if none or two statements were true and the green card if exactly one statement was true.

respect to one of the two tax levels (municipal or national business tax). Respondents randomized into completing the LE with respect to the firm’s compliance with municipal taxes completed the RRT regarding national taxes and vice versa. All respondents, therefore, completed both the double LE and RRT in 2023, but no respondent completed any single experiment more than once; or both with respect to the same tax level.

We provide more information on how we administered the RRT in Supplementary Appendix C.2.

3.1.4 Bayesian Truth Serum (BTS)

The Bayesian Truth Serum (BTS) method was introduced by [Prelec \(2004\)](#) and has been used primarily in online experiments (see, e.g., [Frank et al., 2017](#); [Weaver and Prelec, 2013](#)). The goal of the experiment is to give respondents a financial incentive to answer sensitive questions about their business’s tax non-compliance truthfully. This works by calculating a score that is in expectation higher for respondents who give honest and accurate answers and paying them a financial reward based on their score.

The experiment simply involves asking each respondent two questions for each tax level. First, respondents are asked a direct question on whether their business did not pay municipal (file national) business tax in the previous year. Immediately after, they are asked to provide their best guess of how many other businesses in our survey will *say* that they did not pay municipal (file national) tax in the previous year. Respondents in the BTS treatment group are told that their answers will be scored according to a mathematical formula that gives a higher score for more truthful and more accurate answers and that they will receive a financial reward depending on their score.

Respondents are only informed that their answers will be used to calculate a score that rewards truthful answers and that they will receive a bonus based on this score. This information is the key mechanism through which the BTS method works. The effectiveness of the technique thus depends on whether respondents deem it credible that they will be rewarded for accurate answers and on the reward being sufficiently high to induce truth-telling behavior. Respondents are not, however, informed about the specific mechanism of the BTS and do not need to be able to follow Bayesian reasoning for the BTS to be effective in inducing truthful responses.

After data collection for the survey had been completed, we calculated each respondent’s score based on their answers. Respondents in the treatment group were then rewarded with airtime, depending on their score. On average, respondents received UGX 5,500 in airtime.

Supplementary Appendix C.3 details the calculation of the BTS score.

3.1.5 Direct questioning (DQ)

To obtain a direct, self-reported measure of tax non-compliance, respondents were asked the following questions:

1. *In 2022 [2023], did this business NOT pay for a new or renewed trade license?*
2. *In 2022 [2023], did your business NOT file corporate or presumptive income tax to the*

In order to avoid priming respondents for our experiments and in line with recommended practice (Blair and Imai, 2012), we field these questions later in the questionnaire after all other experiments are already completed. For a randomly selected sub-sample in 2023, these questions are part of the incentive experiment (*Bayesian Truth Serum*, Section 3.1.4). Our ‘clean’ estimate of self-reported compliance thus comes from the control group not selected to take part in the incentive experiment. In the 2023 survey, the order of the two questions was randomized. In the 2024 survey, respondents were randomized into either being asked about local trade license payments or national business tax filings.

3.1.6 Second-order beliefs

As part of the 2024 survey, we also collected respondents’ second-order beliefs on the tax non-compliance rate among businesses similar to theirs. After answering the direct question, respondents were asked one of

1. *Imagine 10 businesses similar to yours. How many of the 10 do you think did NOT pay for a new or renewed trade license to the KCCA in 2023?*
2. *Imagine 10 businesses similar to yours. How many of the 10 do you think did NOT file corporate or presumptive tax to the URA in 2023?*

3.1.7 Enumerator guess

After having finished each interview, we also asked enumerators to complete a short module with questions about the interview process, respondent, and business as part of the 2024 survey. As part of this module, enumerators were asked about the respondent’s attitude (cooperative or not) and attentiveness during the interview, and the primary language that the enumerator used to communicate with the respondent²¹. In addition, we asked enumerators to provide their own best guess on the tax non-compliance status of the business along with the confidence they have in this guess.²² Depending on whether the respondent had previously been randomized into answering questions regarding trade licenses or national business taxes, enumerators were asked

²⁰We decided to ask about tax *non*-compliance (instead of tax compliance) as our standard phrasing to be consistent with the literature on sensitive questions that typically asks whether the respondent has engaged in the sensitive behavior of interest and for the direct question to mirror the sensitive statements presented to respondents during the list experiment and randomized response technique. In Section 4.5, we measure the causal effect of this choice on response behavior.

²¹56% of interviews were conducted in English, 39% in Luganda, and 5% in another local language.

²²The idea behind this approach was that enumerators might have been able to gather insights during the interview process that are not reflected in respondents’ survey responses. For example, during fieldwork debriefs for the 2023 survey, enumerators regularly told anecdotes that were suggestive of a firm’s tax compliance behavior. This included information respondents had shared with them outside the interview or observations enumerators made while present at the business.

1. *Please take a moment to reflect on the interview. In your opinion, do you think this business filed corporate or presumptive tax to the URA [paid for a new or renewed KCCA trade license] in 2023?*

3.2 Validation of survey techniques for sensitive behavior

3.2.1 Identification and characterization of measurement errors

To identify, quantify, and characterize measurement error, we compare self-reported tax non-compliance under different survey techniques to the administrative benchmark. In each case, we are interested in three metrics:

1. **True positive rate** (TPR)²³: the share of true tax non-compliers correctly identified as non-compliers
2. **False positive rate** (FPR): the share of true tax compliers wrongly identified as non-compliers
3. **Correct classification rate** (CCR)²⁴: the share of correctly identified tax compliers and non-compliers in the sample

The LE and RRT do not allow us to recover self-reported tax non-compliance at the individual firm level but only yield an estimate of the aggregate non-compliance rate. However, we can still estimate the TPR, FPR, and CCR using ‘partition validation’ (Höglinger and Jann, 2018; Kuhn and Vivyan, 2022). This involves two steps: (i) splitting the sample into true tax non-compliers and true tax compliers according to the administrative benchmark; (ii) for each group, estimating the tax non-compliance rate using the respective survey measure. The true positive rate for the survey measure is the estimated tax non-compliance rate among true tax non-compliers; the false positive rate is the estimated non-compliance rate among true compliers; and the correct classification rate is estimated as

$$CCR = TPR \times NC + (1 - FPR) \times (1 - NC) \quad (2)$$

where TPR is the true positive rate, FPR is the false positive rate²⁵, and NC is the true tax non-compliance rate.

Techniques for which accuracy is statistically significantly different from 100% may suffer from two sources of misreporting. First, they may be subject to *strategic* misreporting which occurs when respondents have an incentive to misreport their true tax non-compliance status. This applies to respondents whose firm does not comply with taxes and who may therefore underreport tax non-compliance relative to the benchmark. Second, any technique may be subject to *non-strategic* misreporting occurring even in the absence of incentives to misreport. For example, respondents may be inattentive, confused, or otherwise satisfice in providing their

²³The true positive rate is sometimes also called ‘sensitivity’ or ‘recall’.

²⁴Also simply called ‘accuracy’

²⁵ $(1 - FPR)$ is therefore the true negative rate

responses. While true positives may relate to strategic or non-strategic misreporting, false positives are unambiguously linked to non-strategic misreporting.

When two measures differ in their estimated non-compliance rate, this may therefore reflect differences in strategic or in non-strategic misreporting between them. As a result, it is not possible to conclusively ascribe differences in aggregate estimates to changes in any one of the two types of misreporting. This is important as an increase in the estimated rate of sensitive behavior—typically ascribed to a *reduction* of strategic misreporting—may instead relate to an *increase* in non-strategic misreporting.

Whether non-strategic misreporting leads to an over- or underestimation of sensitive behavior depends on the exact response heuristic employed by non-strategic misreporters and the true rate of sensitive behavior among respondents. For example, if non-strategic misreporters pick a response at random, this biases estimates toward 0.5. In the common scenario of low rates of self-reported sensitive behavior under direct questioning (< 0.5), this may cause an alternative measure to estimate a higher aggregate prevalence without reflecting more accurate measurement. Validation using an individual-level benchmark does not suffer from this limitation and lets us characterize the nature of differences in aggregate measurement error between techniques.

3.2.2 Aggregation across tax levels

For two of our experiments—the double list experiment (LE) and randomized response technique (RRT)—respondents only answered questions with respect to one out of two tax levels (see sections 3.1.2 and 3.1.3 and Table 1). For any of the two experiments, our sample size for each tax level is thus only half the total sample size. For greater statistical power, we pool across tax levels in our main analysis.²⁶ We take care to only ever compare responses to the correct benchmark: the response given by a respondent for a specific tax level to the true tax compliance status for the same tax level.²⁷ Still, we verify the robustness of our results to conducting our analysis separately for each tax level. When comparing the LE or RRT to direct questioning (DQ), we only do so for the BTS control group in order to make a clean comparison.²⁸

3.2.3 Separating knowledge from truthful reporting

A potential threat to our experimental design would be if deviations between self-reported tax compliance and the administrative data benchmark did not reflect untruthful reporting but a lack of knowledge of the tax code or the business’s tax affairs. For example, it could be the case that respondents are simply unaware that their firm had to pay national URA business

²⁶For the list experiment, this exploits the fact that base items were identical between the group randomized into the municipal or national business tax list. See also the approach in Blair and Imai (2012).

²⁷The sensitive behavior prevalence that we estimate is thus an average of tax non-compliance across both tax levels. For the purpose of our analysis, we are interested in the ‘prevalence of sensitive behavior’ more abstractly, rather than a substantive interpretation of the non-compliance rate that we estimate. An exception is the Bayesian hierarchical model in Section 5 for which we never pool across different taxes. The reason is that we elicited priors separately for KCCA and URA here.

²⁸Not imposing this restriction does not alter our results meaningfully.

tax *and* obtain a trade license from the KCCA in every year. Similarly, they might be unaware whether their business filed tax or not. If any of this were the case, a failure of different survey experimental techniques to provide estimates that are closely aligned with the administrative data could be due to the inability of the technique to elicit more truthful responses, or it could be due to people reporting truthfully but based on misconceptions about the firm’s tax obligations and affairs.

We take a multi-pronged approach to mitigating the possible confounding effect of misconceptions and lack of knowledge. First, the respondent for the interview is selected as a person affiliated with the business and knowledgeable about its tax affairs. This safeguards that respondents are familiar with the obligations and conduct of the firm under business tax law.²⁹

Second, our survey experiments are preceded by a survey module on tax knowledge, which tests respondents’ familiarity with the most important cornerstones of business tax regulations in Kampala. Through this, we can directly test for the existence of relevant misconceptions. It also allows us to calculate an index of tax knowledge, which we can control for in a robustness check. Among our respondents, 92% answered at least half of the questions correctly, 64% correctly answered seven or more out of nine questions, and 35% scored eight or nine out of nine points in 2023. We fielded a somewhat shorter module with only five questions in 2024. Among our respondents, 89% answer three or more questions correctly.

Third, enumerators are instructed to read out a standardized information text to all respondents following the tax knowledge module that clarifies the main tax obligations of firms. This ensures that even if misconceptions about tax regulations exist, these are rectified for all respondents in a standardized way before the commencement of the survey experiments.

3.3 Robustness to benchmark issues

3.3.1 Dummy list experiment with non-sensitive benchmark

To probe the robustness of our results to potential issues with the administrative benchmark, we conduct a dummy list experiment with two randomly selected groups. One group received a list of four non-sensitive baseline items (see Section 3.4.1, consistency check) and the other group a list of the same four baseline items *plus* an additional statement. The additional statement was about the division of Kampala that the business is located in. The statement was objectively true for 100 percent of our sample as it was based on the location of the business premises where the interview was taking place.³⁰

²⁹Indeed, in 2023 we interviewed the person reporting to be the most knowledgeable about the firm’s tax affairs in 87% of cases. Only in 8% of the remaining cases did the respondent report having little knowledge of tax affairs. Our results are robust to excluding these respondents. In 2024, we interviewed the most knowledgeable person in the business in 77% of cases. In all remaining cases, respondents report to have at least ‘intermediate’ knowledge of tax affairs.

³⁰Divisions are the coarsest level of administrative and geographic organization in Kampala and boundaries between divisions are unambiguous and well known among the population. The risk of an inadvertently incorrect answer due to a lack of knowledge or a misunderstanding is thus minimal. We exclude 13 observations where the business had moved to a different division and the pre-loaded division was therefore not correct. This does not affect our conclusions.

The item on business location thus acts as the quantity of interest whose prevalence we can estimate using a list experiment setup. However, different from a canonical list experiment, our statement of interest is non-sensitive and based on an exogenous benchmark which we know to be true in 100 percent of cases. It therefore allows us to assess whether the list experiment produces an accurate estimate in a setting where neither item sensitivity nor the validity of our administrative benchmark can drive any observed bias.

Rejecting that the estimated prevalence under the dummy list experiment is equal to 100% would provide evidence that the list experiment format, when implemented in practice, invites non-strategic reporting errors that bias estimates. If so, we further explore the misreporting pattern that gives rise to such error. We test whether the resulting response pattern can be rationalized by a model in which a fraction of respondents non-strategically misreport on the additional fifth item, assuming baseline items are answered the same way in both groups.³¹ The alternative is a pattern in which responses to the baseline items themselves differ between the short and long list groups. In other words, we test whether results are consistent with a simple model of non-strategic misreporting on a trivially true item rather than a more complex mechanism in which respondent behavior changes across all statements once presented with a marginally longer list.

Under the simpler model, the distribution of responses in the long list group is made up of two components. First, responses from those who do not count the fifth item (which mirror the distribution observed in the short list group) and second, responses from those who do count it (which mirror the short list distribution shifted up by one). Given the short list distribution, we estimate the conditional probability of failing to count the trivially true fifth item at each baseline count. We achieve this by matching the model’s predicted long list frequencies to those observed in the data. The test rejects the model if the implied probabilities fall outside the unit interval (since they represent conditional probabilities) or if a separate overidentifying restriction comparing the highest cell to the model’s prediction from the lower cells is violated. Rejecting the simpler model would indicate that adding a single, non-sensitive item to the list changes how respondents process the task as a whole.

3.4 Quantifying non-strategic measurement errors

We conduct two experiments that isolate measurement error originating from non-strategic misreporting at the individual respondent level. We then use these results to explore mechanisms behind measurement uncertainty.

3.4.1 Consistency check

In Module B of the questionnaire, respondents answer questions about their business and its practices. Subsequently in Module D, respondents are presented with a list of items and asked how many apply to their business. None of the items are sensitive and all of the items were

³¹We estimate separate probabilities of failing to count the fifth item according to the respondent’s counterfactual baseline count. This allows failure to be spread flexibly across the item count distribution.

previously included as direct questions in Module B.³²

As all statements had previously been asked to respondents as direct questions, we can construct two item counts covering the same items: (i) from direct questions, (ii) as provided when asked in list format.³³

Comparing the list item count (ii) to the constructed direct question item count (i) provides a measure of reporting consistency at the individual respondent level. Our measure of response quality is a consistent answer (i.e., the same item count) between the list and direct questions.

3.4.2 Recall check

In Module C of the 2024 questionnaire, respondents are quizzed about their tax knowledge. After, enumerators read out a short information text to each respondent that clarifies the main pillars of the tax code. In a question asked at the end of Module D, we include four statements (all of which are false) and ask respondents how many statements are correct. All respondents have been informed about the correct answers to these statements by way of the information text that was read out to them in Module C. Any answers deviating from zero ('no statements are true') therefore indicate that respondents did not correctly recall information provided to them by the enumerator or were inattentive when reading/hearing the statements.

The recall check has three key features compared to the reporting consistency check in Section 3.4.1. First, it should identify not only respondents who do not follow instructions closely or consistently but also those who are inattentive to the exact content of the statements or information provided by the enumerator. Second, while it is possible that respondents report a consistent item count between the direct questions and list by chance in Section 3.4.1, we mitigate this possibility here: Evidence from the 2023 survey showed that only a tiny fraction of respondents answered "0" to list-format questions, suggesting that this is not a common answer option even among those not following instructions. This reduces the probability that some misreporters pass the test by chance. Finally, as the recall check requires enumerators to first read out a text to respondents in Module C, it places a higher emphasis on the role of the enumerator to facilitate accurate responses. We use this to compare between-enumerator variation in respondent misreporting between the consistency check and recall check.

3.5 Mechanisms behind measurement uncertainty

We conduct a series of experiments and tests to investigate the factors that give rise to differential measurement error between survey methods and therefore measurement uncertainty.

³²A randomly selected subsample received five instead of four items. The fifth item was about the administrative division of Kampala that the firm is located in and always true (see Section 3.3.1). This item was pre-loaded and verified by the enumerator at the start of the interview. For respondents who received the fifth item, it is added to the direct question item count.

³³The statements in the list were 1. *I have worked in this business for 10 or more years*, 2. *This business has more than one branch*, 3. *This business has a dedicated bank account*, 4. *This business can be contacted by clients via text or instant message*, 5. *This business is located in [Division of business]* (only long list group and added to direct question count for them). The direct questions asked about identical information but used small variations on the phrasing in order to make the duplication less obvious.

3.5.1 Mental models (phrasing experiment)

In the 2024 survey, a randomly selected subset of 10% of the total sample (stratified by whether the respondent was asked about KCCA trade licenses or URA business tax) received an alternative (non-negated) question format for the list experiment.

The sensitive statement presented was one of

- Tax non-compliance (standard phrasing): *This business did NOT file corporate or presumptive tax to the URA [did NOT pay for a new or renewed KCCA trade license] in 2023*
- Tax compliance (non-negated phrasing): *This business filed corporate or presumptive tax to the URA [paid for a new or renewed KCCA trade license] in 2023*

We conducted a similar experiment when asking about respondents' second-order beliefs:

- Tax non-compliance (standard phrasing): *Imagine 10 businesses similar to yours. How many of the 10 do you think did NOT file corporate or presumptive tax to the URA [did NOT pay for a new or renewed KCCA trade license] in 2023?*
- Tax compliance (non-negated phrasing): *Imagine 10 businesses similar to yours. How many of the 10 do you think did file corporate or presumptive tax to the URA [did pay for a new or renewed KCCA trade license] in 2023?*

To avoid confusing respondents, we conducted the randomization for the phrasing experiment at the interview level such that any one respondent only ever was asked about either tax non-compliance or tax compliance.

3.5.2 Implementation fidelity

Measurement-as-designed may differ from measurement-as-implemented. We exploit variation in the assignment of our 27 enumerators to interviews during the 2024 survey as a window into the survey implementation process. We investigate whether enumerators differ in their implementation of an otherwise identical survey, the extent to which response quality systematically varies between enumerators, and how this affects measurement uncertainty.³⁴ On average, each enumerator completed 71 interviews. Not every enumerator worked every day of the fieldwork such that the number of completed interviews varies between 34 and 98 (Figure B.1).

Differences in enumerator behavior are hard to measure and typically present an unobservable part of survey implementation quality.³⁵ We use three proxies that allow us to investigate the discretionary choices enumerators make when administering an otherwise identical survey. First, we compare enumerators according to their note-taking behavior at the end of the interview. Before the conclusion of each interview, enumerators could note down additional

³⁴We did not pre-specify the enumerator analysis and therefore regard this part as more exploratory.

³⁵One option is random audits of interviews. Conducting such audits without giving enumerators the chance to adjust their behavior accordingly is tricky in in-person surveys but may be more feasible for other survey modes such as over the phone.

information that respondents wanted to pass on. Providing additional information was entirely voluntary for respondents and the final question of the interview. Noting feedback down therefore meant additional effort for enumerators before they could conclude the interview without being able to be held accountable for it. We use this to compare enumerators according to how frequently they took any notes and how long these notes were, on average.³⁶ In addition, we compare enumerators according to the net interview length, i.e., the total length of the interview after removing time taken for the introduction and consent as well as any breaks in between modules.³⁷

We then explore how heterogeneity in survey implementation between enumerators affects data quality as measured through the consistency and recall checks. Crucially, the recall check assigns a much more prominent role to enumerators for safeguarding data quality: they are tasked with reading out the information text based on which respondents could clearly identify the four incorrect statements. This setup mirrors a real survey design choice researchers face. Some measures, such as a direct question, involve enumerators much less heavily in explaining the question format to respondents and administering the question than others, such as a list experiment. It also allows for a clear theoretical prediction: If enumerator discretion gives rise to implementation heterogeneity—and ultimately measurement uncertainty—enumerators should explain a larger share of variation in data quality in the recall check than the consistency check.

We operationalize this analysis through a multilevel mixed-effects model. For binary outcomes, we use multilevel mixed-effects logistic regression with enumerator random effects; for continuous outcomes, we employ mixed-effects linear regression. We estimate the intra-cluster correlation, that is, the share of total outcome variation that is explained by between-enumerator variation (as opposed to variation between different interviews within each enumerator).

A caveat for this analysis is that enumerators in our survey were not explicitly randomized to interviews. Instead, they were split into five teams, each headed by a supervisor. Team composition was such that every team was approximately balanced in terms of gender and the experience of enumerators (e.g., included some enumerators who had already participated in the 2023 survey). Every morning during fieldwork debrief, teams were allocated to a distinct geographic area of Kampala (usually a specific division or list of parishes within a division). To make fieldwork efficient, supervisors then allocated interviews to enumerators such that any enumerator typically covered sampled businesses within a small geographic area (e.g., a specific street or village). In case an enumerator was unable to conduct an interview, they called their supervisor in order to obtain another firm to interview in their area from a randomly drawn replacement list. From conversations with our fieldwork coordinator, the process of allocating enumerators to teams, teams to geographic areas, and specific enumerators to interviews was

³⁶We manually filter out non-substantive notes such as ‘No’, ‘Nothing’, ‘N/A’, etc.

³⁷Doing so is important as respondents regularly had to consult their manager before agreeing to start the interview or had to tend to customers and other business matters in between survey modules. We encountered a handful of broken time stamps in some modules which likely happened when enumerators skipped through the questionnaire and did not trigger a module’s time stamp properly. We replace the module length in these cases with the module’s median length. We also encountered a few large outliers that likely reflect respondents having to take extended breaks during the module (e.g., to tend to urgent business matters). We therefore winsorize module lengths at the 99th percentile.

otherwise as good as random. As the main difference between enumerators in their assignment was therefore geography, we control for location fixed effects in our analysis and conduct a string of robustness checks—controlling for respondent and firm characteristics as well as including high-dimensional random effects in the mixed-effects model—to support the conditional exogeneity assumption.

4 (Mis-)measuring sensitive behavior and the emergence of measurement uncertainty

4.1 Comparative validation of different survey methods

We start by presenting our findings on the accuracy of different measures of tax non-compliance compared to the true non-compliance benchmark. For illustration, we put ourselves in the shoes of a researcher who collected self-reported data on tax non-compliance but suspects that reports from a direct question underreport sensitive behavior.

Indeed, we find a large gap between self-reported and true non-compliance (Table A.3). While just under 10% (5% in 2024) of respondents admit that their firm has failed to comply with its tax duties in the past year, the true non-compliance rate is over four times as high (43%).³⁸ Direct questioning thus results in substantial underreporting of sensitive behavior in line with expectations.

Are survey techniques designed to elicit truthful responses to sensitive questions successful in estimating tax non-compliance rates that are more in line with the true rate? Figure 1 would suggest so. While incentivizing truthful reporting via the *Bayesian Truth Serum* of Prelec (2004) does not result in a different non-compliance estimate, both the list experiment (+10 percentage points relative to direct questioning³⁹) and randomized response technique (+20 percentage points) result in much higher estimated business tax non-compliance.

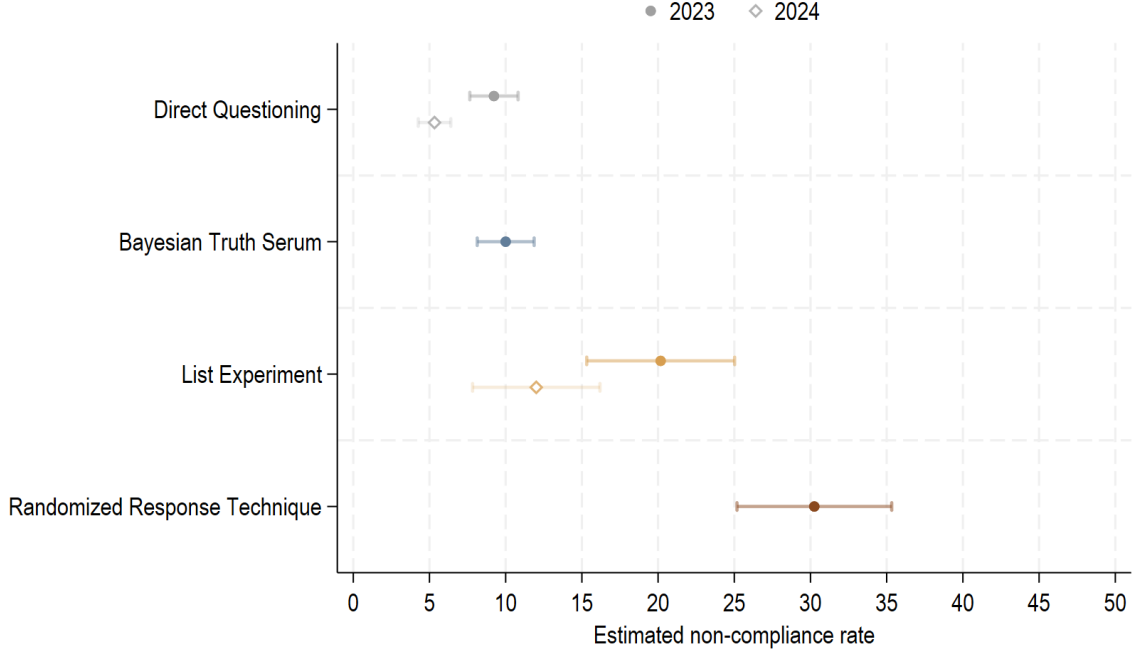
In the absence of any benchmark, it is common research practice to take higher rates of sensitive behavior as reflecting more accurate measurement (Wolter and Diekmann, 2021; Tourangeau and Yan, 2007). Based on this assumption, we would conclude that both the list experiment (LE) and randomized response technique (RRT) reflect more accurate measurement than direct questioning and the BTS and that the RRT provides the most accurate measure.

Two more points are worth noting. First, both the LE and RRT increase the variance of the non-compliance estimate. This bias-variance trade-off is a well-known feature of indirect ques-

³⁸The high rate of tax non-compliance in our sample reflects pervasive informality and very low levels of revenue collection in Sub-Saharan Africa and Uganda in particular (see, e.g., Cohen, 2024; Occhiali and Kalyango, 2023). It is also in line with the observation that many firms *either* pay municipal *or* national business tax, but only a minority pays both (Dietrich et al., 2026). In Dietrich et al. (2026), we extensively discuss the Ugandan context and present estimates of business tax non-compliance in Kampala that are aligned with non-compliance in our survey sample. The numbers are also broadly in line with those reported in Cohen (2024), who studies a sample of small- to medium-sized entrepreneurs in Uganda and finds that only 50% paid presumptive or personal income tax during the 2017–18 fiscal year.

³⁹+7 percentage points in 2024.

FIGURE 1: Relative performance of survey-based measures of sensitive behavior



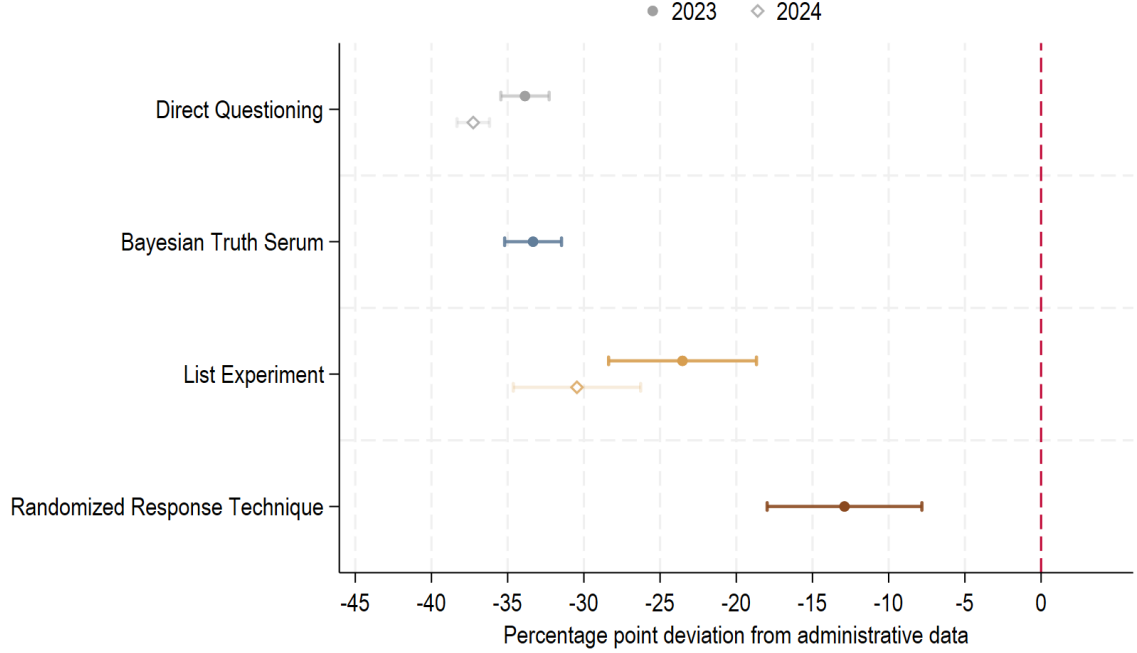
Note: The figure reports the estimated tax non-compliance rate (in percent) across different survey measures. Markers show point estimates while whiskers are 95%-confidence intervals.

tioning techniques (Blair and Imai, 2012; Aronow et al., 2015) and the efficiency cost researchers pay for adding noise to the estimation in order to protect respondents’ privacy. Second, the LE and RRT preclude us from inferring individual compliance status by design.

These conclusions would remain largely unchanged even with access to an aggregate benchmark (Rosenfeld et al., 2016). Comparing each estimate to the correct benchmark rate, it would still appear as if both the LE and RRT reduced measurement error and that the RRT was most successful in doing so (Figure 2). However, Figure 2 also demonstrates that no technique comes close to the true benchmark rate. LE and RRT, respectively, make up 32% and 62% of the initial gap between direct questioning and the benchmark. Even though confidence intervals are large, we can reject the null at $\alpha = 0.05$ that the LE comes within 18 percentage points and the RRT comes within 8 percentage points of the true rate.

In summary, in an applied research or comparative validation setting, Figures 1 and 2 would appear to tell a compelling story: the added anonymity of the LE and RRT seemingly allowed respondents to answer more truthfully, resulting in estimates that are much more closely aligned with the true prevalence of sensitive behavior. When presented with similar findings and in the absence of an aggregate benchmark—by far the most common scenario in applied research—we may even be tempted to conclude that a tripling of reported sensitive behavior, as in the case of the RRT, may mean that bias in reports has been removed altogether. Based on our individual-level benchmark, we now provide evidence that such conclusions would be mistaken.

FIGURE 2: Relative performance of survey-based measures of sensitive behavior, with aggregate benchmark



Note: The figure reports the estimated tax non-compliance rate across different survey measures, relative to a benchmark provided by administrative data. All estimates are normalized such that their administrative benchmark sits at zero and markers show percentage point deviations from the benchmark. Whiskers are 95%-confidence intervals.

4.2 Measurement error by survey method

Our results thus far disregard a key point. If direct questioning underestimates sensitive behavior, higher estimated prevalence of sensitive behavior is not a sufficient condition for misreporting to have been reduced. This is because increases in estimated tax non-compliance can come from two sources: (i) an increase in true positives (that is, more tax non-compliers admitting that their firm did not pay tax—the intended mechanism); and (ii) an increase in *false* positives (that is, more tax compliers are incorrectly estimated to not have complied). As we observe true tax compliance for each firm, we can directly test this using partition validation.

Table A.4 reports the results of our partition validation approach, estimating the true positive and false positive rate for each survey approach. Perfect measurement would imply an estimated non-compliance of 100% for tax non-compliers (true positive rate) and 0% for compliers (false positive rate). However, this is not the case.

Starting with direct questioning, we find that both the true positive and false positive rates are low. About 15% of tax non-compliers (9% in 2024) admit to not having paid their taxes. Among tax compliers, a small but non-negligible group of 5% of respondents (3% in 2024) incorrectly report that their firm did not comply with taxes.⁴⁰ For both the LE and RRT, false

⁴⁰This may have several reasons. For municipal taxes, keep in mind that we adopted a conservative definition of tax compliance that required the firm to have paid for *at least one* trade license. Respondents

positive rates are substantially higher (16% and 27%) while true positive rates remain far from accurate (25% and 35%)⁴¹.

This points to a fundamental pitfall when judging the accuracy of such survey estimates without a benchmark, namely, that their seemingly promising performance observed in aggregate was driven in almost equal parts by increases in true and false positives (Figure 3).⁴² As neither the LE nor RRT asks respondents to openly incriminate themselves or their firm, we interpret these patterns as evidence for increased *non-strategic* misreporting under these techniques. Such misreporting occurs, for example, when respondents do not understand instructions, insufficiently engage with them, or satisfice when giving answers.

As a result, neither technique performs any better than direct questioning in terms of the overall accuracy with which they separate firms into tax compliers and non-compliers (Figure 4 and Table A.7). Put differently, large increases in *non-strategic* misreporting under the LE and RRT negate any accuracy gains from less *strategic* misreporting under these techniques. Neither technique outperforms a random guess by more than 10 percentage points.⁴³

4.3 Detecting internal validity flaws

Our findings create a concern for applied researchers: In the absence of an individual-level benchmark, we would have been unable to detect variation in non-strategic misreporting between survey measures and therefore could not have discerned decreases in strategic misreporting (and therefore improvements in measurement accuracy) from increases in non-strategic misreporting (decreases in the accuracy of measurement). This could have led us to fundamentally flawed conclusions about the accuracy of different measurement techniques and the prevalence of tax non-compliance in Kampala.

Would common internal consistency tests have alerted us to such issues and prevented us from drawing incorrect conclusions about the validity of our findings? In Supplementary Appendix D, we conduct a battery of commonly proposed validity tests for the measures we field. These span tests of internal consistency (Chuang et al., 2021), of core assumptions behind the list experiment (Blair and Imai, 2012), and “controlling for” social desirability bias by way of the 13-item short form of the Marlowe-Crowne social desirability module (see, e.g., Dhar et al., 2022). We demonstrate that none of these tests reliably picks up the subtle but fundamental

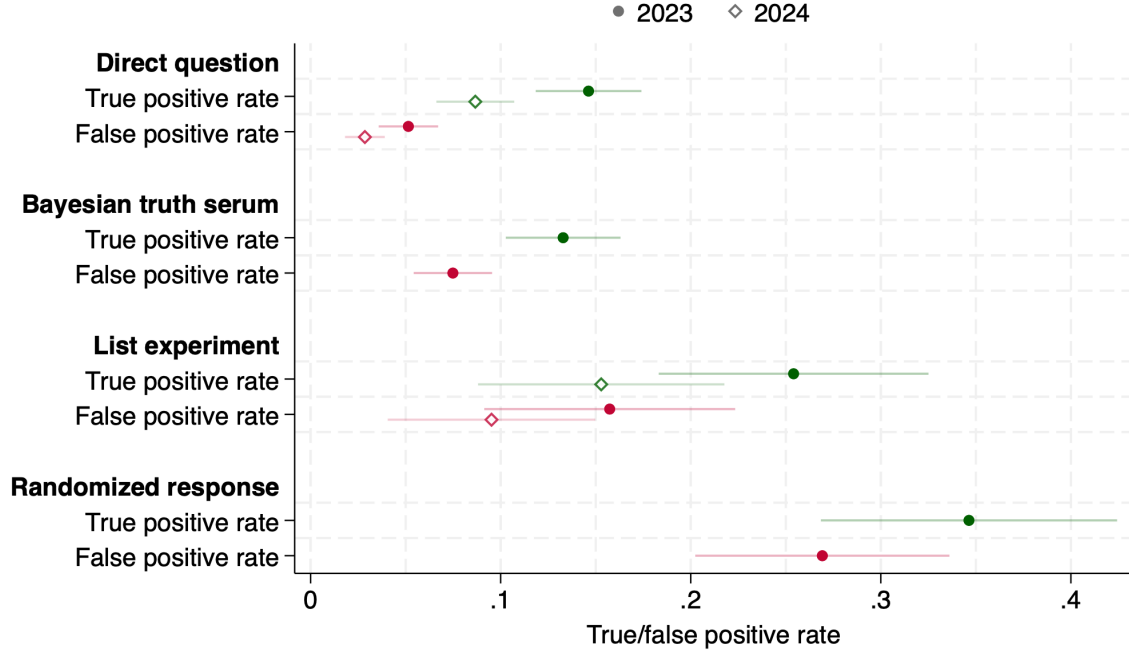
may have been aware of further branches belonging to the same firm that did not obtain a trade license which would make these cases true positives instead of false positives. Another reason may relate to a small share of respondents misunderstanding the question (we asked respondents whether their firm did NOT comply with taxes to be consistent with common survey practice to ask for the sensitive and not the non-sensitive behavior).

⁴¹15% (TPR) and 10% (FPR) for the LE in 2024.

⁴²Further, all techniques—and particularly the list experiment and randomized response technique—do not pick up underlying differences in the true prevalence of sensitive behavior. Estimates of non-compliance with KCCA trade licenses and non-compliance with URA national business tax are very similar and statistically indistinguishable for the LE and RRT, and relatively larger for the direct question (Table A.6). This happens even though the underlying true prevalence of non-compliance is much higher for URA business taxes in our sample (60–63% for URA vs. 24–27% for KCCA).

⁴³In fact, by assuming a tax non-compliance rate of 0% (all firms pay for a trade license and file national business tax), one could have achieved a correct classification rate in the same ballpark (57%).

FIGURE 3: True and false positive rate by survey measure



Note: The figure reports the true positive rate (share of tax non-compliers that are classified as not complying) and false positive rate (share of tax compliers that are classified as non-compliers) by different survey techniques to measure the tax non-compliance rate. Markers show point estimates while whiskers are 95%-confidence intervals.

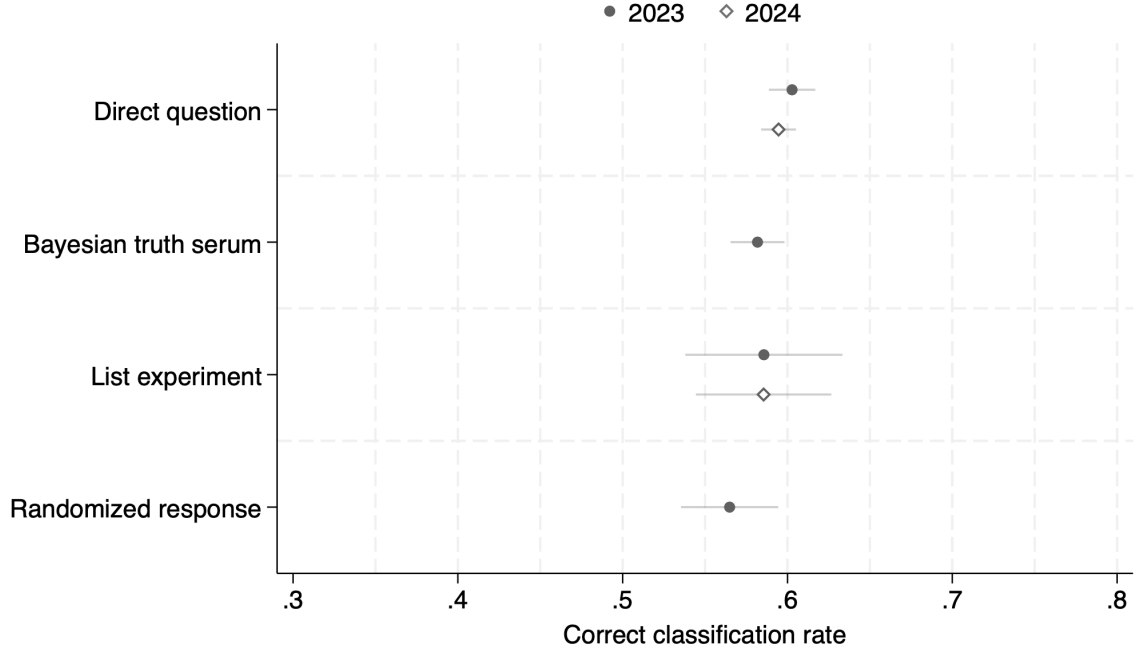
measurement errors driving the breakdown of these techniques in our application. Aggregate heuristics and common tests are therefore no reliable guide to judging (and comparing) the validity of different measurement approaches.

4.4 Robustness to administrative benchmark issues and generalizability

Two potential concerns about our findings may arise. The first is whether issues in the administrative benchmark we use, rather than respondents misreporting, could drive our results. The second is whether the issues we detect are specific to the sensitive topic we study or reflect a more general pattern of misreporting in surveys. We conduct a series of experiments and tests that speak to both.

We implement a dummy list experiment using an exogenous, objective, and non-sensitive benchmark. Instead of asking about tax non-compliance, we ask about the business's location in Kampala. What we estimate is the share of businesses located in the administrative division where the interview is taking place, an observable fact. This is trivially true for 100% of our sample—we are interviewing respondents at their place of business. Yet, taking estimates from this dummy list experiment at face value would suggest that 29% of businesses are *not* located in the geographic division in which the interview took place and where the business is thus demonstrably located (Table A.9). This demonstrates that the list experiment's design itself,

FIGURE 4: Correct classification rate by survey measure



Note: The figure reports the correct classification rate (accuracy) of different survey techniques measuring the tax non-compliance rate. Markers show point estimates while whiskers are 95%-confidence intervals.

not its specific content, gives rise to non-strategic reporting errors. This pattern is consistent with the list experiment’s higher cognitive demands promoting non-strategic errors.⁴⁴

To characterize how this bias arises, we test whether the shift in the distribution of responses between the long and short list groups can be rationalized by a model in which a fraction of respondents fail to count only the additional fifth item (about the business’s location), while baseline responding is unchanged. We allow this failure rate to vary with the respondent’s counterfactual baseline count. The alternative is a more complex pattern in which responses to the baseline items themselves differ between the two groups. The data are consistent with the simpler model (Table A.10). Combined with the properties of the fifth item, which is objectively true for every respondent, non-sensitive, and unambiguous in wording, this result supports interpreting the bias in the dummy list experiment as non-strategic misreporting concentrated on the marginal item.

What do these results tell us about respondent behavior and inference with survey data

⁴⁴Two additional arguments further support the robustness of our findings in Section 4.2. First, they replicate across both types of taxes that businesses are liable for in our context. For each tax, the benchmark comes from separate tax authorities’ registries. In the case of KCCA trade licenses, this also does not involve any cross-linking of tax records as we sample directly from KCCA’s records. Second, neither systematic gaps in the administrative benchmark nor incorrectly logged payments would overturn the conclusions of our paper. Systematic gaps do not affect our main validity test which identifies the false positive rate, i.e., the non-compliance rate among those recorded as tax *compliers* in the benchmark data. Incorrectly logged payments, on the other hand, would mean that the true non-compliance rate is even higher than what is reflected in our benchmark. This would mean that any measurement technique would be even further away from measuring the prevalence of sensitive behavior accurately.

more generally? In two survey tasks, we isolate and quantify non-strategic measurement errors beyond the narrow issue of measuring sensitive behavior.

When comparing item counts asked in list format to the same count constructed from direct questions (the “consistency check”), we find that 30% of respondents report a non-matching count (Table A.11). On average, item counts differ by 0.35 items in absolute terms. In most cases, reported counts differ by one item while they differ by two or more items for 4% of total respondents or 14% of respondents with a non-matching item count (Figure B.2a). As deviations are largely symmetric around zero, they cause ‘classical’ measurement error in an estimate of the mean item count (Figure B.2b). Classical measurement error results in a larger variance of the estimated item count without causing bias at the mean (Table A.11).

Instead of investigating internal consistency in respondents’ reports, a second test investigates whether respondents correctly recall information provided to them by the enumerator (the “recall check”, Table A.13). We find that 50% of respondents fail to provide the correct number of true statements out of four (true answer: zero) even after enumerators had read out a text informing them of the correct answers minutes before.⁴⁵ About one in five respondents (19%) are off by a count of two or more.

4.5 Mechanisms behind measurement uncertainty

Through a range of experiments and tests, we investigate the mechanisms behind measurement uncertainty in survey data. Next to the cognitive complexity channel suggested by the increase in non-strategic misreporting under more complex survey techniques (Section 4.2), we find evidence supporting two more mechanisms.

First, we find a substantial role for mental models that cause the same question to be perceived very differently depending on phrasing. Simply inverting the phrasing of the sensitive statement (asking about tax compliance instead of non-compliance) causes a change in estimated tax non-compliance of 47 percentage points for the list experiment (Table A.14) and of 9 percentage points (25%) for second-order beliefs (Table A.15).⁴⁶ This provides clear evidence that respondents *do* pay attention to the statements they are being presented with but perceive and answer them very differently depending on phrasing.⁴⁷ Figure B.3 demonstrates that higher estimated tax non-compliance under the ‘compliance’ phrasing is driven by excess mass in the upper tail of the distribution. That is, respondents are much more likely to estimate that 20% or fewer businesses comply with tax obligations (alternatively: that 80% or more did *not* comply) when asked about tax compliance instead of non-compliance. This shift comes to a similar extent from the remaining part of the distribution.⁴⁸ While estimates differ substantially

⁴⁵We chose 0 as the correct answer because results from our 2023 survey indicated that few respondents picked extreme values ‘by default’. A true answer of zero should therefore mitigate the probability that a satisficing respondent passes the test by chance.

⁴⁶While we also adjust the direct question phrasing accordingly, answer options remained clearly labeled as ‘My business did pay/file’ and ‘My business did not pay/file’, eliminating the scope for comparisons across phrasings.

⁴⁷If, conversely, respondents did not engage with instructions at all, we would not expect a marginal change in the phrasing of the sensitive statement to result in a different estimate at all.

⁴⁸The exception being more heaping at 0.5 for the non-compliance phrasing.

depending on phrasing, neither phrasing clearly outperforms the other in terms of classification accuracy (Table A.14).⁴⁹

Second, we investigate differences between measurement-as-designed and measurement-as-implemented, offering a window into the survey implementation black box and its role for measurement uncertainty. We provide descriptive evidence that enumerators differ in the way they administer an otherwise identical survey where there is room for discretion, that this matters for data quality, and adds to measurement uncertainty as methods differ in the extent to which they are vulnerable to enumerator effects.

We exploit variation in the assignment of our 27 enumerators to interviews during the 2024 survey and show they differ substantially in their diligence in recording information (proxied by whether optional notes were taken and the length of notes) and effort in eliciting answers (proxied by net interview length⁵⁰) (Figures B.4a and B.4b). Using the consistency and recall checks, we then show that response quality varies between enumerators depending on the discretion enumerators have for deviating from standard protocol (Figures B.4c and B.4d). Using a multilevel mixed-effects model, we find that enumerators account for 50% of variation in data quality for the recall check but less than 10% of variation for the consistency check (Table A.16). These patterns are strongly indicative of differences in enumerator behavior that matter most where enumerators have room for discretionary choices in the way they administer the interview. Accounting for the potential non-random allocation of enumerators to interviews does not change these findings meaningfully (Supplementary Appendix E and Table A.17). We find more experienced enumerators to be associated with higher data quality (but not with fewer false positives or false negatives when measuring sensitive behavior) and indicative evidence for learning effects over the course of the survey (Table A.18). We do not find evidence of a robust association between enumerator fatigue (proxied by the number of previous interviews in a day) and data quality. Supplementary Appendix E provides a more comprehensive discussion of the evidence on enumerator effects.

Our mechanism tests thus reveal that seemingly equivalent survey approaches can yield vastly different estimates of the same outcome due to differences in complexity, differences in framing that activate distinct answer heuristics, and differences in the susceptibility of different measures to enumerator effects—thus causing measurement uncertainty.

⁴⁹For the list experiment, asking about tax non-compliance *underestimates* tax non-compliance by 31 percentage points while asking about compliance *overestimates* tax non-compliance by 17 percentage points. Put differently, asking about non-compliance underestimates non-compliance in the list experiment; and asking about compliance underestimates compliance. This implies that too few respondents (implicitly) answer ‘yes’ to the sensitive statement—irrespective of whether ‘yes’ is the sensitive answer or not. We cannot do the same analysis for second-order beliefs as we lack an equivalent individual-level benchmark against which to compare them. Aggregate comparisons are not reliable indicators of measurement validity as we establish in Section 4.2.

⁵⁰That is, excluding any breaks in between modules that respondents needed to take, for example, to tend to customers

5 Inference under measurement uncertainty

Our results thus far demonstrate a general challenge with survey data: typically, unobservable variation in the sources and size of measurement error can lead to large disparities between different survey measures of the same target. In the case of measuring sensitive behavior through surveys, we find that measures exhibit large heterogeneity in estimates yet near-identical accuracy when validated individually. This makes the choice of measurement approach ambiguous yet highly consequential. As the standard error of any single method only captures within-method (sampling) uncertainty, between-method uncertainty is unaccounted for in standard inference.

When measurement accuracy is largely unobservable, multiple plausible approaches exist, and no single measure can be identified as the most accurate, single-measure inference can therefore become highly unreliable. Instead of betting on a single measure, we propose an alternative approach. We take multiple measures, along with prior information, and use the extent of (dis-)agreement between plausible measures to quantify measurement uncertainty and incorporate it into inference.

The basic idea behind this approach is similar in spirit to already established research practice in causal inference. If there exist multiple plausible ways to estimate the causal effect of interest, researchers are expected to demonstrate that different approaches provide consistent results. Researchers then adjust their confidence in the accuracy of findings depending on observed consistency, though typically only implicitly.

Our approach uses a Bayesian hierarchical model (BHM) to make such heuristics explicit and formally incorporates measurement uncertainty into the inference process. Methodologically, it adapts the framework of Bayesian meta-analysis—which uses hierarchical models to quantify uncertainty around generalizability when pooling effects across studies (Meager, 2019; Bandiera et al., 2021; Vivalt, 2020)—to the problem of pooling multiple measures within a single study. In both cases, modeling heterogeneity guards against false precision: about effects in new contexts for meta-analysis and about the estimate itself in our case. Conceptually, our approach shares the spirit of Bayesian model averaging (Chatfield, 1995; Draper, 1995). Accordingly, uncertainty over the correct specification should be reflected in inferential uncertainty rather than assumed away. It relates to recent work incorporating prior information, such as expert judgments, unstructured data, or qualitative insights, into impact evaluation (Iacovone et al., 2025) and connects to established bounding approaches for survey inference under measurement error (de Quidt et al., 2018; Lee, 2009). Our approach demonstrates how to generalize these insights to observational and descriptive estimates of hard-to-observe behavior and attitudes, quantities that are ubiquitous in economic research and essential inputs to policy.

5.1 A Bayesian hierarchical framework for measurement uncertainty

Our approach is to combine data from multiple independent measures (the ‘likelihood’) with prior information to produce a posterior distribution that quantifies inferential uncertainty about the latent behavior or attitude of interest (tax non-compliance in our case). Unlike standard single-measure inference, which captures only sampling uncertainty, this incorporates disagreement

across measures into the posterior. In doing so, our estimation framework provides a more complete account of what we know and do not know given available evidence. We formalize this in the following Bayesian hierarchical model (BHM) of inference under measurement uncertainty.

5.1.1 Estimation target and parameters

Our ultimate target is the true tax non-compliance rate μ^* . We observe this quantity by virtue of our validation benchmark from administrative records. In typical applications, however, μ^* will be unobserved and we need to estimate it given the available evidence.

The available evidence base comprises any prior information⁵¹ and survey data from K plausible measures of μ^* . Each measure k produces an estimate y_k with measure-specific expected value $E[y_k] = \theta_k$ and standard error se_k . The standard error captures sampling error, that is, the variation of y_k around the expected value in finite samples. As sample size grows arbitrarily large, $se_k \rightarrow 0$ and measure k converges to estimate θ_k .

Even in the limit of large N , any approach may be subject to measurement error and therefore provide only a biased estimate of the true parameter μ^* . Formally:

$$\begin{aligned}\theta_k &= E[y_k] \\ &= \mu^* + b_k\end{aligned}\tag{3}$$

where b_k is the bias for measure k , the systematic difference between what measure k estimates and the true value of interest.

We model the θ_k as draws from the distribution of plausible measurement approaches. That is, we observe a sample of K measures of μ^* that are all part of a possibly larger population of plausible measures.⁵²

$$\theta_k \sim \mathcal{N}(\mu, \tau^2)\tag{4}$$

where μ is the mean across all plausible measures of μ^* , and τ^2 the between-measure variance.

How close μ is to the true value μ^* depends on the average of measure-specific biases:

$$\mu = \mu^* + \bar{b}\tag{5}$$

where $\bar{b} = E[b_k]$. When biases approximately offset across measures ($\bar{b} \approx 0$), μ will be an unbiased estimate of the true parameter of interest. However, when measures share a common

⁵¹The prior flexibly accounts for any extraneous knowledge the researcher may have on the range of values within which the latent parameter likely falls. For instance, such prior information can come from previous studies, qualitative observation, or—as in our study—expert predictions. See [Iacovone et al. \(2025\)](#) for a recent application.

⁵²In practice, there often exist multiple reasonable approaches to measuring a quantity of interest with no clear evidence to favor one approach over all others. Plausible measures are those that are methodologically defensible and among which there exists uncertainty regarding the most accurate approach. Researchers should choose multiple measures—from the list of all plausible measures—so that they together span the space of key measurement choices, without artificial padding through clearly inferior measures or premature narrowing based on untested preferences.

directional bias, $\mu \neq \mu^*$.⁵³

Our goal is to estimate the posterior distribution of μ which quantifies our uncertainty about the location of μ given data and priors. Compared to single-measure inference, this is valuable even in cases where $\mu \neq \mu^*$ for two reasons. First, the posterior reflects the full information available to the researcher rather than just a selected subset. Instead of zeroing in on the information contained in a single measure, the posterior honestly aggregates what the evidence base collectively implies for inference. Second, a single measure’s confidence interval captures only sampling uncertainty, the uncertainty about θ_k that arises because we observe a finite sample. This is only appropriate where there is just one plausible measure for μ^* or when all plausible measurement approaches would yield the same expected value θ_k . However, when multiple plausible measures exist and systematically disagree, the choice of measure itself becomes a source of uncertainty. The hierarchical model acknowledges this by letting the θ_k vary across different measurement approaches ($\tau^2 \neq 0$). The greater τ^2 , the more sensitive estimates are to measurement choices and the greater the role of measurement uncertainty—over and above sampling uncertainty—for reliable inference.

In short, the BHM does not promise unbiasedness. It provides full-information inference and honest uncertainty quantification given the available evidence base.⁵⁴

5.1.2 Model specification

We estimate the following hierarchical model:

$$\begin{aligned}
y_k &\sim \mathcal{N}(\theta_k, \sigma_k^2) && \text{[likelihood]} \\
\theta_k &\sim \mathcal{N}(\mu, \tau^2) && \text{[measurement uncertainty model]} \\
\mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) && \text{[location prior]} \\
\tau^2 &\sim \text{Half-Cauchy}(0, 1) && \text{[heterogeneity prior]}
\end{aligned} \tag{6}$$

The model has four components. The likelihood relates observed estimates from different plausible measures y_k to their individual expected value θ_k , with σ_k^2 capturing each measure’s

⁵³Similarly, if the distribution of all plausible measurement approaches is centered around $\mu = \mu^*$, but we only observe a non-random subset of J measures with a common, systematic bias, our meta-estimate $\hat{\mu}$ will inherit this bias: $\mu^* = \mu = \hat{\mu} - E[b_j]$.

⁵⁴To fix ideas, it may be helpful to compare our hierarchical model for inference under measurement uncertainty to hierarchical models in meta-analysis (Meager, 2019; Bandiera et al., 2021; Vivalt, 2020). In random effects meta-analysis, μ^* is the population average treatment effect (ATE) of an intervention across all relevant contexts. μ represents the average effect across all contexts in which researchers could feasibly conduct an impact evaluation. Each impact evaluation study produces an estimate of the study-specific ATE, y_k , which converges to θ_k with large enough samples. The study-specific ATE θ_k may diverge from the universal average effect μ^* due to the uniqueness of each context. Meta analysis produces $\hat{\mu}$, an estimate of μ formed from the observed studies in the meta-analysis sample. Observing a non-random sample of studies from the set of all feasible impact evaluations will bias $\hat{\mu}$ as an estimate of μ . More generally, the meta-estimate $\hat{\mu}$ will only recover μ^* if study-specific deviations from the universal effect approximately cancel out. As in our case, *systematic* bias in the observed data propagates to bias in the meta-estimate. In this analogy, heterogeneity between studies ($\tau^2 \neq 0$) corresponds to measurement uncertainty in our setting (heterogeneity between plausible survey measures). The larger τ^2 , the more study-specific ATEs diverge from one another, and the greater our uncertainty about the location of μ .

sampling variance. The measurement uncertainty model characterizes the population distribution of these expected values across all plausible measures. The individual θ_k from each measurement approach are part of a distribution with average μ and variance term τ^2 . The latter captures measurement uncertainty, systematic heterogeneity between plausible measurement approaches. The location prior on μ summarizes our ex-ante beliefs on the location of the parameter of interest (the tax non-compliance rate); the prior on τ^2 represents our ex-ante beliefs about the amount of measurement uncertainty. We estimate the posterior distribution of μ which characterizes both our best-guess point estimate, given data and priors, and the uncertainty surrounding it. We now describe each component in turn.

5.1.3 Likelihood

We construct the likelihood from six survey measures in the 2024 survey: the direct question, second-order belief (in ‘non-compliance’ and ‘compliance’ phrasing), list experiment (both phrasing variants), and a post-interview enumerator guess. Each measure k contributes an estimate and standard error, which we transform to the logit scale.⁵⁵

Let y_k denote the transformed estimate and σ_k its standard error. The likelihood specifies

$$y_k \sim \mathcal{N}(\theta_k, \sigma_k^2) \quad (7)$$

where θ_k is the expected value for measure k , the quantity to which that measure would converge with arbitrarily large samples.

5.1.4 Measurement uncertainty model

The key modeling choice is whether observed disagreement between estimates y_k only reflects sampling uncertainty or also genuine differences in θ_k between measures. A fixed-effects specification assumes $\theta_k = \mu$ for all k , that is, all measures converge on the same expected value in the limit of infinite sample sizes.

Our empirical evidence on the size and sources of measurement errors motivates a different assumption. Measure-specific biases, for example arising from differences in question complexity, framing, or susceptibility to enumerator effects, reflect genuine heterogeneity between measures and will persist regardless of sample size.

We therefore model the θ_k in a hierarchical framework as draws from a population distribution of plausible, but heterogeneous, measures:

$$\theta_k \sim \mathcal{N}(\mu, \tau^2) \quad (8)$$

The parameter μ represents the average across different, plausible measurement approaches

⁵⁵Our parameter of interest is a proportion and therefore bounded between $[0,1]$. Since normal distributions have unbounded support, we use the logit transformation $\text{logit}(p) = \log(\frac{p}{1-p})$ to make normal distributions appropriate for both the likelihood and prior. Where values approach the boundaries, we treat them as if coming from a truncated normal distribution on the probability scale, draw simulated values via the inverse CDF method, and transform the resulting draws to logit scale.

for the tax non-compliance rate. Parameter τ^2 captures measurement uncertainty, systematic variation in expected values that reflects the sensitivity of estimates to measurement choices. When τ^2 is large, plausible measurement approaches yield substantially different answers irrespective of sample size. As τ^2 approaches zero, the model reduces to the fixed-effects specification where all that separates measures is sampling variation.

This structure has two implications. First, it induces partial pooling. That is, information is shared across measures but the degree to which each individual measure influences the estimate of μ depends on the ratio of within-measure heterogeneity (σ_k^2) to between-measure heterogeneity (τ^2).⁵⁶ Second, $\tau^2 > 0$ increases posterior uncertainty about μ . Even when all measures are observed precisely ($\sigma_k^2 \rightarrow 0$), uncertainty remains regarding where μ lies relative to the different estimates we observe. Through this mechanism, measurement uncertainty feeds into inferential uncertainty about the quantity of interest (e.g., the tax non-compliance rate).

In our main specification, we treat the θ_k as exchangeable draws from the population of plausible measures given μ and τ^2 , $\theta_k \sim \mathcal{N}(\mu, \tau^2)$, so that $\text{Cov}(\theta_k, \theta_j) = 0$ for $k \neq j$. That is, we assume measures to be conditionally independent.⁵⁷ We later relax this assumption, allowing measures within the same method-family to be correlated through a common method effect.⁵⁸

Specifically, we replace $\theta_k \sim \mathcal{N}(\mu, \tau^2)$ with

$$\theta_k = \mu + \sqrt{\rho} a_{m(k)} + \sqrt{1 - \rho} e_k, \quad a_m \stackrel{iid}{\sim} \mathcal{N}(0, \tau^2), \quad e_k \stackrel{iid}{\sim} \mathcal{N}(0, \tau^2), \quad (9)$$

where $m(k)$ denotes the method-family of measure k and the a_m and e_k are mutually independent. The method effect $a_{m(k)}$ is shared by all measures of the same family, whereas e_k is specific to measure k . This implies

$$\text{Var}(\theta_k) = \tau^2 \quad \text{and} \quad \text{Cov}(\theta_k, \theta_j) = \begin{cases} \rho \tau^2, & m(k) = m(j), \ k \neq j, \\ 0, & m(k) \neq m(j), \end{cases} \quad (10)$$

so that ρ is the correlation between two measures of the same method-family, the variance of each measure around μ remains τ^2 at every ρ , and measures that do not share a family are still assumed to be conditionally independent. Our main specification is the special case $\rho = 0$.

⁵⁶In the hierarchical, random effects model, each measure receives a weight that is proportional to $\frac{1}{\sigma_k^2 + \tau^2}$. In a fixed effects specification, weights heavily favor more precise estimates as they are solely determined by $\frac{1}{\sigma_k^2}$. Accounting for measurement uncertainty therefore has an equalizing effect on weights. As τ^2 gets large, individual differences in the precision of each measure matter less, and weights converge toward equality.

⁵⁷We also assume sampling errors to be uncorrelated across measures for the likelihood, $\text{Cov}(y_k, y_j | \theta) = 0$. Some measures share the same respondents meaning that this assumption may not hold exactly. We assess this by estimating the sampling covariance between measures via a joint bootstrap. Even the largest sampling covariance is under half a percent of the between-measure variance τ^2 such that the effect of assuming $\text{Cov}(y_k, y_j | \theta) = 0$ on the overall BHM results is negligible.

⁵⁸We define four method-families: list experiments (non-compliance and compliance phrasing), second-order beliefs (non-compliance and compliance phrasing), direct question, and enumerator guess.

5.1.5 Location prior

We construct an informative prior for μ , separately for municipal (KCCA) and national taxes (URA), by aggregating predictions from 36 academic researchers which we collected via the *Social Science Prediction Platform (SSPP)* (DellaVigna et al., 2019). Each participant of the SSPP survey provided a point prediction of the non-compliance rate as well as a measure of their uncertainty for each prediction.⁵⁹ We use this information to parametrize each expert’s predictions as a normal distribution on the probability scale, then transform to the logit scale for estimation.

The aggregate prior across experts is

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \quad (11)$$

For the prior mean μ_0 , we take the median of experts’ logit-scale predictions. For the prior variance σ_0^2 , we sum two components:

$$\sigma_0^2 = \underbrace{\text{Var}(m_j)}_{\text{between-expert}} + \underbrace{\frac{1}{J} \sum_{j=1}^J s_j^2}_{\text{avg. within-expert}} \quad (12)$$

where m_j is expert j ’s prediction on the logit scale and s_j^2 their implied uncertainty. The prior uncertainty therefore reflects two conceptually distinct sources of uncertainty: Experts disagree with one another about the likely value (between-expert variance) and each expert is individually uncertain about their own prediction (within-expert variance). The resulting prior incorporates the full information content of the prediction exercise while remaining independent of the survey data or administrative benchmark. Figure B.5 summarizes the expert predictions and aggregate prior.

5.1.6 Heterogeneity prior

Barring strong prior information about the degree of measurement uncertainty, the heterogeneity prior on τ^2 should be sufficiently diffuse to accommodate both small and large values. We therefore choose a weakly informative prior that is in line with the standard recommendation for variance priors (Gelman, 2006).

$$\tau^2 \sim \text{Half-Cauchy}(0, 1) \quad (13)$$

This specification allows the data to determine whether measurement uncertainty is negligible or substantial. To assess sensitivity of our results to this choice, we re-estimate the model with similar results using $\tau^2 \sim \text{Half-Cauchy}(0, 2.5)$ and $\text{Inverse-Gamma}(0.5, 0.5)$ which are yet less informative about τ^2 .

⁵⁹Specifically, participants reported a value v such that they believed with 90% certainty that the true non-compliance rate would fall within $\pm v$ percentage points of their prediction. We interpret this as the half-width of the expert’s subjective 90% credible interval and derive the standard deviation as $\frac{v}{1.645}$. Section 2.3 contains more information on the prediction survey and its participants.

5.1.7 Estimation

We estimate the model via Markov chain Monte Carlo (MCMC) simulation with 20,000 draws from the posterior distribution following 5,000 burn-in iterations.⁶⁰ To ease interpretation, we transform posterior draws of μ to the probability scale and report the posterior median as our point estimate.⁶¹

The result of the BHM is our estimate of the full posterior distribution of μ with the posterior median as its location and the posterior standard deviation (or 95% credible interval) as a measure of spread. Both components are individually informative for our purpose. The location provides a point estimate informed by the full evidence base of data and priors. The spread honestly quantifies inferential uncertainty, taking into account sampling error within each measure and systematic disagreement between measures (measurement uncertainty).

5.2 Results

5.2.1 Main results

We now examine what our framework reveals about inference under measurement uncertainty and how this can be used to improve real-world decision making (Figure 5).

First, neglecting measurement uncertainty grossly understates total uncertainty given the evidence base and compromises the reliability of inference. The BHM, on the other hand, quantifies total uncertainty more honestly, thereby attenuating the risk of overconfident inference. Uncertainty as reflected in the posterior standard deviation is 2.4 times (KCCA) to 2.9 times (URA) larger than what the average standard deviation of any single measure suggests.

This comparison still understates the extent to which accounting for measurement uncertainty matters. Unlike a single measure’s estimate, the BHM combines information from multiple measures and the prior. This information gain partly offsets the increase in uncertainty when accounting for between-measure variation. A more direct comparison contrasts the BHM estimate (a “random effects” (RE) estimate) with a Bayesian fixed effects (FE) model that combines information from the same prior and likelihood but naïvely assumes that measures only differ by chance (Figure B.6).⁶² Comparing posterior standard deviations between the two models isolates the role of measurement uncertainty, increasing total uncertainty by over an order of magnitude (about 15 times the FE uncertainty for both KCCA and URA).⁶³

Second, having just two or three measures already goes much of the way to reducing the overconfidence that is inherent to single-measure inference. The set of plausible measurement variations a researcher could test is potentially (very) large but space in surveys is constrained.

⁶⁰We build and estimate the BHM using Stata’s native `bayesmh` command. Equivalent models can be estimated with other standard software such as R’s `baggr` package (Meager, 2019).

⁶¹The posterior median on the probability scale equals the inverse-logit of the posterior median on the logit scale and closely approximates the inverse-logit of the mean on the logit scale.

⁶²The FE model therefore assumes $\tau^2 = 0$ in the measurement uncertainty model (Section 5.1.4).

⁶³As the BHM can exploit information gains by combining multiple measures and the prior, it may also *reduce* overall uncertainty. In settings where there is little measurement uncertainty (relative agreement between different, plausible measures), the information gain from multiple measures and the prior can dominate the “measurement uncertainty tax” and result in tighter estimates.

We therefore investigate how the posterior uncertainty captured by the BHM changes with the number of available measures. We run the BHM on every possible combination of $k \in \{2, \dots, 6\}$ measures, resulting in a total of 57 specifications (Figure B.7 and Table A.20). Relative to the full set of measures, even the most agreeing pair of measures already closes 73% (KCCA) and 79% (URA) of the gap between the average single measure’s standard error and the full-set posterior standard deviation. There are therefore large returns to making measurement uncertainty estimable by collecting at least one additional independent measure of the outcome of interest. However, collecting even more independent measures can still have direct inferential payoffs: With large measurement uncertainty, as in our case, the resulting posterior from few measures is noisy and depends on the specific combination of measures included. Some combinations of measures therefore also result in a posterior that is too conservative. Including more independent measures tends to reduce noise and tighten the posterior standard deviation, resulting in sharper inference.

Third, the comparison of posteriors against the administrative benchmark reveals both the value and limits of the Bayesian hierarchical framework we propose. For non-compliance with KCCA trade licenses, the BHM posterior aligns well with the benchmark. It assigns a 33% probability to values within 5 percentage points of the benchmark, 66% to values within 10 percentage points, and nearly all probability mass (90%) to within 15 percentage points. For URA national business income tax, however, the posterior underestimates the benchmark, with only 1–2% probability mass around it.

This contrast between the KCCA and URA posteriors is instructive for what researchers can expect from adopting our estimation framework. The BHM represents total uncertainty given its inputs, but it does not by itself turn poor measurement into good measurement. When measure biases approximately offset (as in Figure 5a), the posterior median falls near the benchmark; when biases are directionally aligned (as in Figure 5b), the posterior inherits this bias. This is a limitation shared by any aggregation procedure and we are explicit that unbiasedness is not the framework’s purpose. Even so, the Bayesian hierarchical framework maintains the distinct advantage it was conceived to deliver. Single-measure inference compounds the problem by overstating precision around a potentially heavily biased point estimate. The BHM, on the other hand, mitigates this risk via a wide posterior that acknowledges significant uncertainty around the true value’s location given the evidence base. Section 5.1.1 formalizes this point.

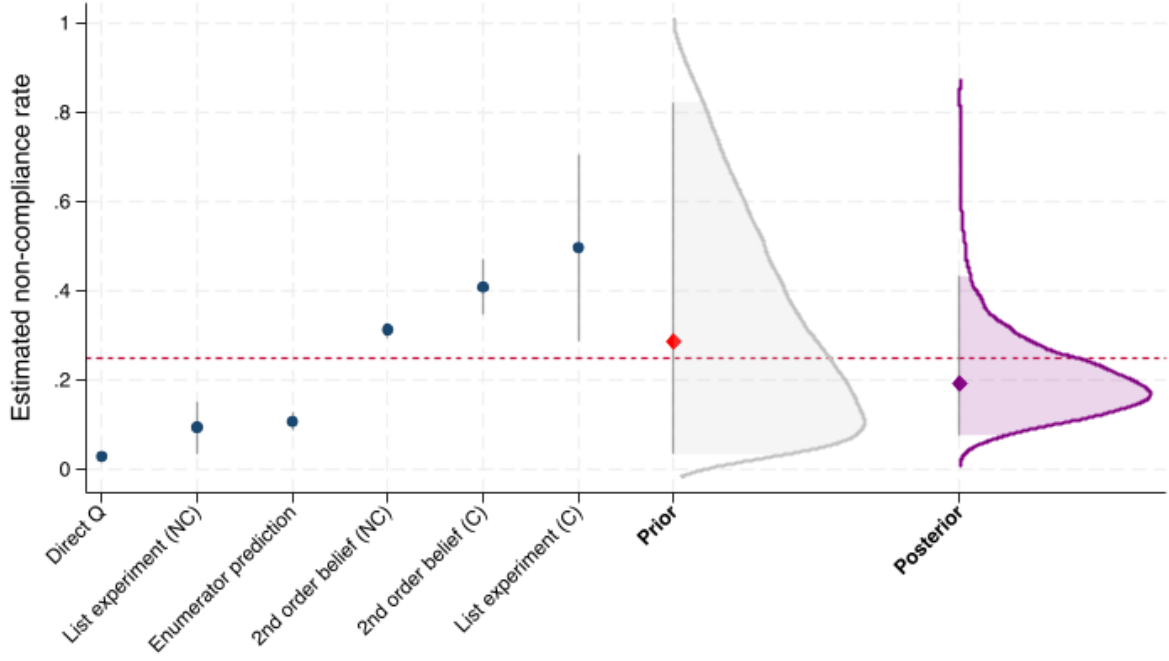
Fourth, the BHM approach communicates inferential uncertainty in a form directly usable for decision problems. Consider a policymaker weighing whether to allocate additional resources to tax enforcement. The policymaker may only consider this an efficient use of scarce resources if tax non-compliance θ exceeds a critical threshold κ . The posterior distribution allows us to directly calculate the probability, given prior and data, that this threshold is reached. That is

$$P(\theta > \kappa \mid \text{data}) > p^*$$

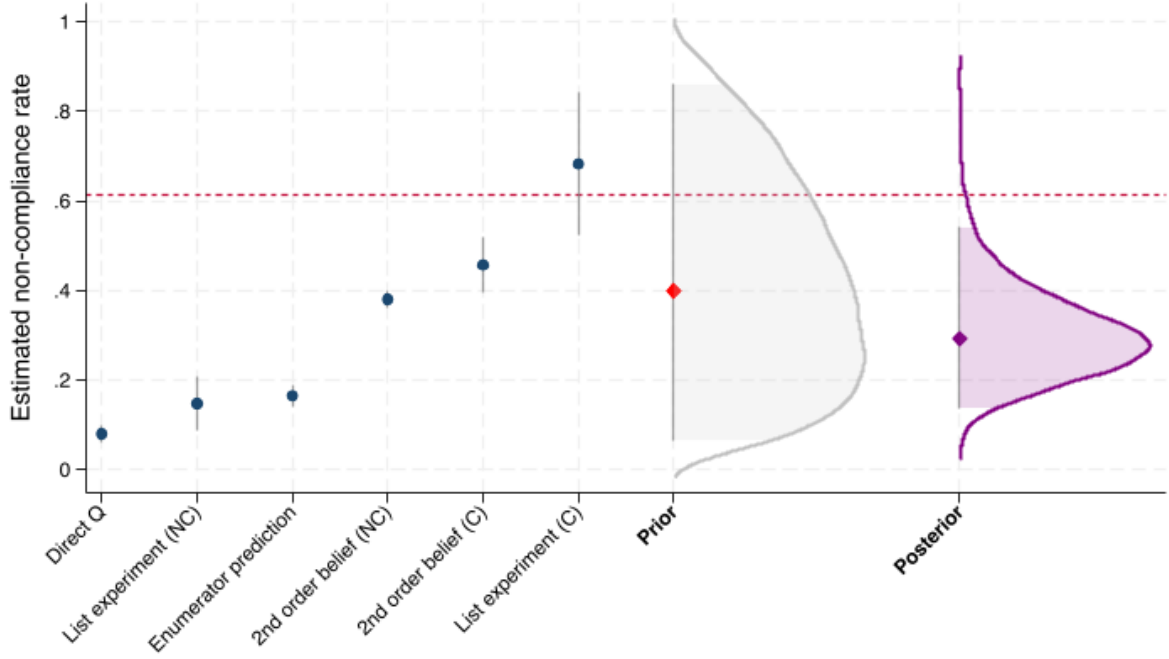
where p^* is the policymaker’s required probability before acting and both p^* and κ are chosen according to the policymaker’s preferences.

Figure B.8 charts out $P(\theta > \kappa \mid \text{data})$ for different thresholds κ . For instance, we estimate a

FIGURE 5: Bayesian hierarchical model



(A) KCCA



(B) URA

Note: The figure visualizes the inputs and outputs of the Bayesian hierarchical model we estimate: six survey estimates of tax non-compliance (the ‘likelihood’), the prior distribution derived from 36 ‘expert predictions’, and the posterior distribution resulting from updating the prior with the likelihood. Markers and whiskers are the estimated means and 95%-confidence intervals in the case of likelihood and medians with 95%-credible intervals in the case of the prior and posterior distribution. Where we randomized question phrasing, we denote this by (NC) for the non-compliance phrasing and (C) for the tax compliance phrasing. All values are shown on the probability scale to ease interpretation. The actual model is fit on the logit scale.

73% probability that non-compliance with KCCA trade licenses is at least 15 percent but only an 11% chance that it is above 30 percent; or an 88% probability that more than 20% of firms do not file URA tax but only 3% that over half do not file. Similarly, suppose a policymaker requires 80% certainty that enforcement will be profitable before allocating additional resources. This condition is only met if the break-even non-compliance rate—the rate above which enforcement pays for itself—is sufficiently low. For KCCA, the posterior provides 80% certainty only for break-even rates below 14 percent; for URA, it does so for break-even rates up to 22 percent. In other words, a cautious policymaker would learn that they can allocate extra resources to URA enforcement even if it requires, relatively, higher non-compliance to be cost-effective; but that they should only pursue increased KCCA enforcement if it already pays off at low non-compliance rates.

5.2.2 Robustness

We confirm the robustness of our results to three key deviations from the main analysis.

First, we use different prior distributions for the measurement uncertainty parameter τ^2 . This could matter as we *assume* $\tau^2 \sim \text{Half-Cauchy}(0, 1)$ but lack empirical data that ties our hands as with the location prior. We therefore confirm the robustness of our headline findings to using $\tau^2 \sim \text{Half-Cauchy}(0, 2.5)$ instead, which puts more mass on large measurement uncertainty, and $\tau^2 \sim \text{Inverse-Gamma}(0.5, 0.5)$ which does not let $\tau^2 \rightarrow 0$. Our headline results are very stable across these different specifications of the heterogeneity prior (Table A.21).

Second, we investigate the sensitivity of our results to dropping expert predictions from the location prior that imply substantial probability mass outside the unit interval. In our survey on the SSPP, some participants reported combinations of point estimates with uncertainty ranges that put probability mass outside the $[0, 1]$ unit interval for the tax non-compliance rate. We drop three experts (two for URA) whose predictions imply 20% or more probability mass outside this interval and recalibrate the prior without the egregious predictions. Our headline results barely change (Table A.21).

Third, we relax the assumption of conditional independence between measures, allowing measures that share the same methodological machinery to be correlated. We group the six measures into four method-families: (i) the list experiment and (ii) the second-order belief, each fielded with two phrasing variants, and (iii) the direct question and (iv) enumerator guess, which contribute one measure each.⁶⁴ Correlations are plausible if, for example, all list experiments induce a common bias. We therefore ascribe part of each measure’s variance to a component shared with other measures of the same method-family. Relative to our main specification, such correlation reduces the effective number of measures informing the pooled estimate and should therefore widen its posterior.

Because each measure yields a single estimate, the structural correlation between measures of the same family cannot be credibly estimated from the data. We instead bound the consequences of correlation between measures by varying the assumed within-method correlation

⁶⁴Within-family correlation is therefore relevant for the list experiment and second-order beliefs which each contribute two variants.

ρ over a grid (Table A.22). This decomposes a measure’s overall variance τ^2 into a component shared by measures of the same method-family and an idiosyncratic component specific to each measure. The posterior standard deviation of the estimated tax non-compliance rate rises steadily as ρ moves from 0 to 0.8 with the 95%-credible interval widening from 36 to 43 percentage points (KCCA) and 40 to 52 percentage points (URA). In other words, once two variants of the same method are treated as largely redundant rather than two independent pieces of evidence, the uncertainty we are left with after seeing the data increases.⁶⁵ Provided positive within-method correlations, this suggests that our headline uncertainty estimates represent, if anything, a lower bound on the total uncertainty we should have given the evidence base and, therefore, on the extent to which single-measure inference misleads.

6 Discussion and conclusion

How can researchers reliably draw inference when outcome measurement is subject to researcher discretion and there is no benchmark to adjudicate between different approaches?

Economics and the social sciences are replete with questions that concern inherently unobservable behavior, attitudes, and preferences. Their correct identification, and the insights drawn from them, are the subject of frequent debate among scientists and sensitive to hard-to-observe measurement errors. This often makes multiple approaches plausible but disagreeing in the conclusions they purport.

In this paper, we demonstrate that by disregarding the above, standard empirical practice leaves a source of inferential uncertainty unaccounted for that can prove exceedingly consequential for reliable research. We denote this source “measurement uncertainty”, formally distinguish it from sampling uncertainty captured in conventional standard errors, and show how to account for it in empirical research.

Factoring in measurement uncertainty—and thereby honestly quantifying what the available evidence collectively implies—more than doubles accounted-for uncertainty in an empirical application estimating business tax non-compliance. Importantly, measurement uncertainty does not decline with larger sample sizes but only with more consistent measurement between plausible alternatives. For example, this is the case when results are quantitatively robust to alternative measures of the same outcome; or when validated best-practice tools exist and the set of plausible alternatives is therefore small.

In the face of measurement uncertainty, committing to one measurement approach reveals only a single point in the range of results the study could have plausibly produced. A close analogue is the literature on the generalizability of impact evaluation results (Vivalt, 2020) where an experimental site reveals one point in the distribution of program impacts found elsewhere. The consequences are unduly confident research conclusions (cf. Stevenson and Fischman, 2026) and disagreements that harden into “dueling certitudes” (Manski, 2011) rather than recognizing

⁶⁵We note that, given the large differences between phrasing variants of the same method (47 percentage points for the list experiment and 9 percentage points for second-order beliefs), assuming high ρ between them represents an extreme check for expositional purposes rather than a scenario that seems supported by the data.

that they are a symptom of an understated uncertainty space. Our study provides researchers with a principled empirical framework through which to address these issues.

How should economists deal with this challenge in practice? Our study has implications for survey design, pre-registration practice, and econometric convention. We briefly summarize some highlights here and provide comprehensive practical guidelines for applied researchers in Supplementary Appendix F.

For survey design, the returns to better measurement are larger than what is commonly recognized. Standard inference quantifies only a fraction of what poor measurement costs, that is, the part that leads to noise in a single measure. As is well known, classical measurement error approaches zero in the limit of large sample sizes. Conversely, reducing non-classical measurement error—systematic bias in a single measure—carries a quantified inferential payoff only once measurement uncertainty is factored in.

For example, this payoff is realized by mitigating known error sources such as complex survey questions, respondent fatigue, or demand effects that plausibly differ between measures of the same outcome. This can reduce between-measure variation and result in tighter inference over and above what increasing sample size can deliver. Our study also draws attention to error modes that only recently started to be more systematically explored by economists such as enumerator effects.⁶⁶ Measurement approaches commonly differ in their susceptibility to discretionary choices by enumerators. Mitigating such potential through intensive enumerator training and oversight may compress between-measure variation and carry large inferential returns. In-built data quality checks and at-scale audio audits may further help unpack the implementation fidelity black box. Still, judging the existence of measurement errors, or lack thereof, from aggregates can be highly problematic as we demonstrate using the example of measuring sensitive behavior. This highlights an important role for validated, best-practice measures that shrink the choice set of plausible measurement options, thereby providing a public good for inference.

For pre-registration, our results suggest a gap in current practice. Large variation between plausible measurement approaches showcases the scope for specification search that pre-registration is meant to eliminate. However, handpicking a single measure ex-ante is what keeps measurement uncertainty unidentified in the first place. Where scope for measurement uncertainty exists, a better approach is therefore to pre-register plausible alternative measures as robustness checks and the rule for aggregating them such that measurement uncertainty is accounted for.

How should these measures be chosen? To efficiently map the uncertainty space, each measure should be methodologically defensible and represent a significant departure in design from all other measures such that they together span the key measurement choices. This neither involves artificial padding through clearly inferior or highly correlated measures, nor premature narrowing based on untested preferences. In practice, our results suggest that even a small selection of one or two additional measures may significantly improve the reliability of inference. As such, the demands that accounting for measurement uncertainty places on survey space can

⁶⁶Recent examples include [Figueiredo Walter et al. \(2026\)](#), [Figueiredo Walter and Moneke \(2026\)](#), [Baseler et al. \(2027\)](#), and [Di Maio and Fiala \(2020\)](#).

be comparable to those of other influential contributions to the measurement error literature. For example, [Gillen et al. \(2019\)](#) propose collecting duplicate elicitations of the same measure from respondents in order to correct for classical measurement error in survey data. Similarly, [de Quidt et al. \(2018\)](#) recommend repeating tasks or questions at the end of a survey with an induced “demand treatment” in order to bound experimenter demand effects. Ultimately, however, the number of measurement variations to test will depend on the size of the existing measurement uncertainty space.

Even though our paper is mainly concerned with descriptive estimates of behavior and attitudes in surveys, its implications extend beyond. For example, measurement uncertainty affects experimental research if treatment interacts differentially with different approaches to measuring the outcome of interest. Given the sensitivity of respondent behavior to small design variations that we and others have documented, assuming that such interactions exist is the reasonable default and should be falsified not by assumption but by explicitly testing for it in the spirit of our framework. Further, the non-strategic measurement errors that we show go easily undetected in survey data will reduce statistical power even when randomization cancels out bias. Our framework also extends naturally to other data sources and situations where researchers make discretionary measurement choices. For example, this includes the choice between different remotely sensed earth observation products which have been found to support diverging conclusions ([Josephson et al., 2026](#)).

We are candid that our framework is no panacea for resolving measurement issues and neither do we advocate an “anything-goes” approach in which pooling estimates turns poor measurement into good measurement. Quite the contrary: we argue that by forcing measurement uncertainty to the fore and into inference, our framework compels analysts to think carefully about measurement as poor choices carry quantified inferential costs. Narrowing the space of such choices is the role of high-quality methodological research that establishes best practices.

Importantly, our framework does not promise unbiasedness but is an uncertainty quantification tool. Where the structure of measurement error can be credibly identified, correcting it is still the first choice. An example is the obviously related instrumental variables (ORIV) estimator of [Gillen et al. \(2019\)](#) that recovers the underlying parameter via duplicate elicitations of the same measure as long as measurement error is classical. Our results show why this route is frequently closed for the behaviors economists study. The errors we document are systematic rather than classical, specific to each method, and capable of moving in the same direction, so that averaging or instrumenting across measures need not converge on the truth. A correction premised on classical error does not remove the bias in this setting but hides it inside a point estimate reported with misleading precision.

References

- Aronow, P. m., Coppock, A., Crawford, F. W., and Green, D. P. (2015). Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence. *Journal of Survey Statistics and Methodology*, 3(1):43–66.
- Artinger, F. M., Gigerenzer, G., and Jacobs, P. (2022). Satisficing: Integrating Two Traditions. *Journal of Economic Literature*, 60(2):598–635.
- Bandiera, O., Fischer, G., Prat, A., and Ytsma, E. (2021). Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments. *American Economic Review: Insights*, 3(4):435–454.
- Baseler, T., Chen, L., and Ginn, T. (2027). Familiar faces, honest answers? the impacts of enumerator familiarity and modality on survey responses. *Journal of Development Economics*, 185:103914.
- Bertelli, O., Calvo, T., Lavallée, E., Mercier, M., and Mesplé-Somps, S. (2025). What one thinks, what one says and what one does: Male justifications and practices of gender-based violence in Mali. *Journal of Development Economics*, page 103479.
- Bertrand, M. and Mullainathan, S. (2001). Do People Mean What They Say? Implications for Subjective Survey Data. *American Economic Review*, 91(2):67–72.
- Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, 74(5):817–848.
- Blair, G., Chou, W., and Imai, K. (2019). List Experiments with Measurement Error. *Political Analysis*, 27(4):455–480.
- Blair, G. and Imai, K. (2012). Statistical Analysis of List Experiments. *Political Analysis*, 20(1):47–77.
- Blair, G., Imai, K., and Lyall, J. (2014). Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan. *American Journal of Political Science*, 58(4):1043–1063.
- Blair, G., Imai, K., and Zhou, Y.-Y. (2015). Design and Analysis of the Randomized Response Technique. *Journal of the American Statistical Association*, 110(511):1304–1319.
- Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement Error in Survey Data. In *Handbook of Econometrics*, volume 5, pages 3705–3843. Elsevier.
- Bursztyn, L., Haaland, I., Roever, N., and Roth, C. (2025). The Social Desirability Atlas. *Mimeo*.

- Celhay, P., Meyer, B. D., and Mittag, N. (2024). What leads to measurement errors? Evidence from reports of program participation in three surveys. *Journal of Econometrics*, 238(2):105581.
- Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419–466. Publisher: [Oxford University Press, Royal Statistical Society].
- Christensen, G. and Miguel, E. (2018). Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature*, 56(3):920–980.
- Chuang, E., Dupas, P., Huillery, E., and Seban, J. (2021). Sex, lies, and measurement: Consistency tests for indirect response survey methods. *Journal of Development Economics*, 148:102582.
- Cohen, I. (2024). Technology and the state: Building capacity to tax via text. *Journal of Public Economics*, 236:105154.
- Crépon, B., Elsayed, A., and Gazeaud, J. (2025). Unbiased and Accurate: Measuring Sensitive Outcomes Through Ballot-Bag Surveys. *Forthcoming: American Economics Review: Insights*.
- de Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and Bounding Experimenter Demand. *American Economic Review*, 108(11):3266–3302.
- De Weerdt, J., Gibson, J., and Beegle, K. (2020). What Can We Learn from Experimenting with Survey Methods? *Annual Review of Resource Economics*, 12(1):431–447.
- DellaVigna, S., Pope, D., and Vivalt, E. (2019). Predict science to improve science. *Science*, 366(6464):428–429. Publisher: American Association for the Advancement of Science.
- Dhar, D., Jain, T., and Jayachandran, S. (2022). Reshaping Adolescents’ Gender Attitudes: Evidence from a School-Based Experiment in India. *American Economic Review*, 112(3):899–927.
- Di Maio, M. and Fiala, N. (2020). Be wary of those who ask: A randomized experiment on the size and determinants of the enumerator effect. *The World Bank Economic Review*, 34(3):654–669.
- Dietrich, S., Markhof, Y., and Vincent, R. C. (2026). The cost of bureaucratic fragmentation: business tax evasion and revenue mobilization in a low-income country. *IFS Working Paper*, (26/39).
- Dillon, A., Karlan, D., Udry, C., and Zinman, J. (2020). Good identification, meet good data. *World Development*, 127:104796.
- Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):45–70.

- Figueiredo Walter, T. and Moneke, N. (2026). Endogenous Sample Selection.
- Figueiredo Walter, T., Moneke, N., and Radu, A. (2026). The Missing Poor. *American Economic Review: Insights*. Forthcoming.
- Frank, M. R., Cebrian, M., Pickard, G., and Rahwan, I. (2017). Validating Bayesian truth serum in large-scale online human experiments. *PLOS ONE*, 12(5):e0177385.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3).
- Genest-Grégoire, A., Godbout, L., and Guay, J.-H. (2022). Lists: A Novel Experimental Method to Measure Tax Evasion. *National Tax Journal*, 75(3):517–537.
- Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study. *Journal of Political Economy*, 127(4):1826–1863.
- Goldsmith-Pinkham, P. (2026). Tracking the Credibility Revolution across Fields.
- Heck, D. W., Hoffmann, A., and Moshagen, M. (2018). Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behavior Research Methods*, 50(5):1895–1905.
- Höglinger, M. and Diekmann, A. (2017). Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis*, 25(1):131–137.
- Höglinger, M. and Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PLOS ONE*, 13(8):e0201770.
- Hoy, C., Jolevski, F., and Obeyesekere, A. (2024). *Revealing Tax Evasion: Experimental Evidence from a Representative Survey of Indonesian Firms*. Washington, DC: World Bank.
- Iacovone, L., McKenzie, D., and Meager, R. (2025). Bayesian Impact Evaluation with Informative Priors: An Application to a Colombian Management and Export Improvement Program. *Econometrica*, (May).
- Iraman, E., Ono, Y., and Kakinaka, M. (2022). Tax compliance and social desirability bias of taxpayers: experimental evidence from Indonesia. *Journal of Public Policy*, 42(1):92–109.
- Josephson, A., Michler, J. D., Kilic, T., and Murray, S. (2026). The mismeasure of weather: Using earth observation data for estimation of socioeconomic outcomes. *Journal of Development Economics*, 178:103553.
- Kuhn, P. M. and Vivyan, N. (2022). The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries. *Political Analysis*, 30(3):381–402.

- Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, 76(3):1071–1102.
- Manski, C. F. (2011). Policy Analysis with Incredible Certitude. *The Economic Journal*, 121(554):F261–F289.
- Manski, C. F. (2019). Communicating uncertainty in policy analysis. *Proceedings of the National Academy of Sciences*, 116(16):7634–7641.
- Meager, R. (2019). Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.
- Menkveld, A. J., Dreber, A., Holzmeister, F., et al. (2024). Nonstandard Errors. *The Journal of Finance*, 79(3):2339–2390.
- Meyer, B. D., Mok, W. K. C., and Sullivan, J. X. (2015). Household Surveys in Crisis. *Journal of Economic Perspectives*, 29(4):199–226.
- Miller, J. D. (1984). *A new survey technique for studying deviant behavior*. Univ. Microfilms International, Ann Arbor, Mich.
- Occhiali, G. and Kalyango, F. (2023). Can tax agents support tax compliance in low-income countries? Evidence from Uganda. *Public Administration and Development*, 43(4):269–279.
- Oprea, R. (2024). Complexity and Its Measurement. In *prepared for the Handbook of Experimental Methods in the Social Sciences*.
- Prelec, D. (2004). A Bayesian Truth Serum for Subjective Data. *Science*, 306(5695):462–466.
- Raghavarao, D. and Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):40–45.
- Rosenfeld, B., Imai, K., and Shapiro, J. N. (2016). An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science*, 60(3):783–802.
- Stevenson, M. T. and Fischman, J. B. (2026). Humans in the Loop: The Next Frontier in the Credibility Revolution.
- Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5):859–883.
- Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, 18(6):3045–3089.

- Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309):63–69.
- Weaver, R. and Prelec, D. (2013). Creating Truth-Telling Incentives with the Bayesian Truth Serum. *Journal of Marketing Research*, 50(3):289–302.
- West, B. T. and Blom, A. G. (2016). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*, page smw024.
- Wolter, F. and Diekmann, A. (2021). False Positives and the “More-is-Better” Assumption in Sensitive Question Research. *Public Opinion Quarterly*, 85(3):836–863.
- World Bank (2013). The Uganda 2013 Enterprise Surveys Data Set - Implementation Report. Technical report, World Bank.

Appendix A: Supplementary tables

TABLE A.1: Summary statistics (2023 and 2024)

Survey round	2023	2024
Non-compliance rates		
Non-compliance with KCCA ($t - 1$)	9.13	11.13
Non-compliance with URA ($t - 1$)	64.84	57.49
Non-compliance with KCCA (t)	25.76	24.80
Non-compliance with URA (t)	61.39	61.45
Business characteristics		
Industry		
Manufacturing	2.50	5.38
Construction	2.43	2.74
Retail	47.94	52.84
Wholesale trade	11.16	8.74
Transport and storage	2.30	2.43
Accommodation and food	6.09	7.45
Information and communication	2.84	1.76
Finance and insurance	2.84	1.96
Real estate	0.81	0.31
Professional, scientific and technical	6.15	13.44
Education	0.88	1.50
Other - misc.	14.06	1.45
Kampala Division (location)		
Central	62.95	65.93
Kawempe	5.68	6.31
Makindye	10.01	9.62
Nakawa	11.97	11.63
Rubaga	9.40	6.51
Respondent: Owner	57.74	51.65
Respondent: Others	42.26	48.35
Number of employees	6.60	5.53
Number of branches	7.30	9.46
Self-reported tax knowledge (among non-owners)	6.79	3.38
Female owner	36.04	40.54

TABLE A.2: Descriptive statistics from prediction survey

	Mean	1 st Quartile	Median	3 rd Quartile	SD	Min	Max
Prediction: KCCA non-compliance	0.339	0.200	0.300	0.475	0.198	0.030	0.760
Uncertainty range: KCCA prediction	0.190	0.080	0.150	0.225	0.170	0.040	0.900
Prediction: URA non-compliance	0.415	0.275	0.400	0.500	0.193	0.100	0.900
Uncertainty range: URA prediction	0.189	0.100	0.115	0.250	0.163	0.030	0.900
Observations	36						

Note: The table shows predictions as collected on the probability scale. For estimation of the Bayesian hierarchical model, we transform them to the logit scale (Section 5.1.2). ‘Prediction’ refers to experts’ point predictions of the tax non-compliance rate. ‘Uncertainty range’ refers to the half-width of their stated 90% credible interval (the value v for which they believed with 90% certainty that the true non-compliance rate would be within $\pm v$ percentage points of their prediction).

TABLE A.3: Estimated tax non-compliance rate with different measurement approaches

	(1) Direct Questioning	(2) List Experiment	(3) Randomized Response	(4) Bayesian Truth Serum
Panel A: Survey 2023				
Estimated non-compliance	0.092 (0.0081) [0.076,0.108]	0.202 (0.0247) [0.153,0.250]	0.303 (0.0259) [0.252,0.353]	0.100 (0.0095) [0.081,0.119]
Observations	1571	1459	1460	1320
Benchmark rate	0.431	0.437	0.432	0.433
Panel B: Survey 2024				
Estimated non-compliance	0.053 (0.0054) [0.043,0.064]	0.120 (0.0213) [0.078,0.162]		
Observations	1707	1731		
Benchmark rate	0.426	0.425		

Note: Estimated tax non-compliance using different measurement approaches. The benchmark non-compliance rate from administrative data for each sample is included at the bottom. Standard errors are in parentheses, 95% confidence intervals are in brackets.

TABLE A.4: False and true positive rate

	(1) Direct Questioning	(2) List Experiment	(3) Randomized Response	(4) Bayesian Truth Serum
Panel A: Survey 2023				
False positive rate	0.051 (0.0080)	0.157 (0.0336) [0.013]	0.269 (0.0341) [0.000]	0.075 (0.0105) [0.076]
True positive rate	0.146 (0.0142)	0.254 (0.0362) [0.001]	0.346 (0.0397) [0.000]	0.133 (0.0154) [0.522]
Observations	1571	1459	1460	1320
Panel B: Survey 2024				
False positive rate	0.029 (0.0053)	0.095 (0.0278) [0.415]		
True positive rate	0.087 (0.0104)	0.153 (0.0330) [0.242]		
Observations	1707	1731		

Note: Estimated true positive rate (share of non-compliant firms classified as non-compliant) and false positive rate (share of compliant firms classified as non-compliant). Standard errors are in parentheses, p-values for testing each estimate against direct questioning are in brackets.

TABLE A.5: False and true positive rate, by tax

	Direct Questioning		List Experiment		Randomized Response		Bayesian Truth Serum	
	(1) KCCA	(2) URA	(3) KCCA	(4) URA	(5) KCCA	(6) URA	(7) KCCA	(8) URA
Panel A: Survey 2023								
False positive rate	0.039 (0.0080)	0.075 (0.0151)	0.145 (0.0408) [0.011]	0.186 (0.0595) [0.059]	0.263 (0.0430) [0.000]	0.279 (0.0559) [0.000]	0.053 (0.0101) [0.290]	0.118 (0.0203) [0.091]
True positive rate	0.085 (0.0197)	0.172 (0.0173)	0.301 (0.0668) [0.001]	0.230 (0.0434) [0.223]	0.340 (0.0778) [0.000]	0.349 (0.0462) [0.000]	0.083 (0.0212) [0.952]	0.154 (0.0180) [0.460]
Observations	790	781	775	684	684	776	663	657
Panel B: Survey 2024								
False positive rate	0.022 (0.0058)	0.042 (0.0109)	0.089 (0.0340) [0.055]	0.106 (0.0484) [0.179]				
True positive rate	0.036 (0.0133)	0.105 (0.0133)	0.109 (0.0637) [0.238]	0.174 (0.0386) [0.085]				
Observations	838	869	862	869				

Note: Estimated true positive rate (share of non-compliant firms classified as non-compliant) and false positive rate (share of compliant firms classified as non-compliant). Standard errors are in parentheses, p-values for testing each estimate against direct questioning are in brackets.

TABLE A.6: Estimated tax non-compliance rate, by tax

	Direct Questioning		List Experiment		Randomized Response		Bayesian Truth Serum	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	KCCA	URA	KCCA	URA	KCCA	URA	KCCA	URA
Panel A: Survey 2023								
Estimated non-compliance	0.051 (0.0078) [0.035,0.066]	0.134 (0.0122) [0.110,0.158]	0.190 (0.0347) [0.122,0.258]	0.214 (0.0352) [0.145,0.284]	0.282 (0.0377) [0.208,0.356]	0.321 (0.0356) [0.251,0.391]	0.060 (0.0093) [0.042,0.079]	0.140 (0.0135) [0.113,0.167]
Observations	790	781	775	684	684	776	663	657
Benchmark rate	0.254	0.609	0.268	0.629	0.240	0.601	0.255	0.613
Panel B: Survey 2024								
Estimated non-compliance	0.025 (0.0054) [0.014,0.036]	0.081 (0.0092) [0.062,0.099]	0.093 (0.0300) [0.034,0.152]	0.147 (0.0302) [0.088,0.207]				
Observations	838	869	862	869				
Benchmark rate	0.233	0.612	0.235	0.612				

Note: Estimated tax non-compliance using different measurement approaches, by tax level. The benchmark non-compliance rate from administrative data for each sample is included at the bottom. Standard errors are in parentheses, 95% confidence intervals are in brackets.

TABLE A.7: Correct classification rate

	(1)	(2)	(3)	(4)
	Direct	List	Randomized	Bayesian
	Questioning	Experiment	Response	Truth Serum
Panel A: Survey 2023				
Correct classification rate	0.603 (0.0072)	0.586 (0.0243) [0.867]	0.565 (0.0151) [0.583]	0.582 (0.0083) [0.057]
Observations	1571	1459	1460	1320
Panel B: Survey 2024				
Correct classification rate	0.595 (0.0054)	0.586 (0.0210) [0.000]		
Observations	1707	1731		

Note: Estimated correct classification rate (accuracy; the share of firms reporting their true tax compliance status). Standard errors are in parentheses, p-values for testing each estimate against direct questioning are in brackets.

TABLE A.8: Correct classification rate, by tax

	Direct Questioning		List Experiment		Randomized Response		Bayesian Truth Serum	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	KCCA	URA	KCCA	URA	KCCA	URA	KCCA	URA
Panel A: Survey 2023								
Correct classification rate	0.738 (0.0078)	0.466 (0.0121)	0.707 (0.0337) [0.375]	0.447 (0.0351) [0.611]	0.641 (0.0225) [0.000]	0.497 (0.0202) [0.194]	0.727 (0.0092) [0.363]	0.435 (0.0135) [0.090]
Observations	790	781	775	684	684	776	663	657
Panel B: Survey 2024								
Correct classification rate	0.759 (0.0054)	0.436 (0.0092)	0.722 (0.0299) [0.226]	0.453 (0.0293) [0.580]				
Observations	838	869	862	869				

Note: Estimated correct classification rate (accuracy). Standard errors are in parentheses, p-values for testing each estimate against direct questioning are in brackets.

TABLE A.9: Dummy list experiment with non-sensitive benchmark

Survey 2024	
Estimated prevalence	0.710 (0.0457) [0.620,0.800]
Observations	1909
Benchmark prevalence	1.000

Note: Dummy list experiment with exogenous, non-sensitive benchmark. The additional statement reads ‘*This business is located in [division]*’ wherein the business’ division is pre-filled by the survey tool according to the true business location. It is therefore always true and should yield a prevalence estimate of 1 if respondents follow instructions. Robust standard errors are in parentheses and 95% confidence intervals in brackets.

TABLE A.10: Test for misreporting patterns in the dummy list experiment

Panel A: Estimated probability of failing the fifth statement	
Baseline count = 0	0.065 [0.031, 0.107]
Baseline count = 1	0.307 [0.198, 0.414]
Baseline count = 2	0.331 [0.201, 0.451]
Baseline count = 3	0.453 [0.257, 0.619]
Baseline count = 4	0.323 [-0.370, 0.688]
Observations	1910
Panel B: Overidentification test	
Baseline count = 4 (Bottom-up)	0.323 [-0.370, 0.688]
Baseline count = 4 (Top-down)	0.375 [-0.272, 0.728]
Gap: Bottom-up vs. top-down	-0.052 [-0.198, 0.000]
Observations	1910

Note: Panel A reports estimated probabilities of failing to count the trivially true fifth item, by counterfactual baseline count. Estimates are obtained by matching the model-predicted distribution of long list responses to the observed distribution, under the assumption that long list responses differ from short list responses only through the fifth item. Estimates outside the unit interval indicate that the observed response pattern cannot be rationalized under this assumption. Panel B reports an overidentifying restriction: the failure rate at baseline count 4 implied by the lowest cells of the long list distribution should equal that implied by the highest cell. A non-zero gap indicates the same. 95% percentile bootstrap confidence intervals in brackets.

TABLE A.11: Reporting consistency

Survey 2024	
Share reporting non-matching item counts	0.303 (0.0106)
Mean <i>absolute</i> difference in item counts	0.352 (0.0136)
Mean difference in item counts	0.001 (0.0158)
Observations	1864

Note: The table compares respondents' reported item counts when asked in list format to the same item count constructed from direct questions to the respondent.

TABLE A.12: Complexity experiment

	(1) Item counts don't match	(2) Abs. Δ item counts	(3) List count - DQ count
Survey 2024			
Longer list	0.046** (0.0213)	0.072*** (0.0271)	-0.231*** (0.0311)
Constant	0.276*** (0.0147)	0.312*** (0.0179)	0.112*** (0.0203)
Observations	1851	1851	1851

Note: OLS regressions with the following dependent variables: a dummy variable denoting a non-matching item count between the list and when constructed from direct questions (column 1); the absolute difference in item counts (column 2); and the simple difference between the list item count and direct question count (column 3). The coefficient of interest is for a dummy variable denoting whether the respondent randomly received a list with one additional statement (five statements instead of four). Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.13: Recall experiment

	Percent
Survey 2024	
Item count	
0	50.1
1	30.6
2	14.6
3	3.8
4	1.0
Observations	1923
Note: Tabulation of item count responses. All values in percent.	

TABLE A.14: Statement phrasing experiment: List experiments

	Pooled		KCCA		URA	
	(1)	(2)	(3)	(4)	(5)	(6)
	Non-compl.	Compl.	Non-compl.	Compl.	Non-compl.	Compl.
Survey 2024						
Estimated non-compliance rate	0.120 (0.0213)	0.590 (0.0672) [0.000]	0.093 (0.0300)	0.496 (0.1072) [0.000]	0.147 (0.0302)	0.683 (0.0822) [0.000]
True positive rate	0.153 (0.0330)	0.616 (0.1155)	0.109 (0.0637)	0.287 (0.2429)	0.174 (0.0386)	0.781 (0.1207)
False positive rate	0.095 (0.0278)	0.537 (0.0859)	0.089 (0.0340)	0.563 (0.1142)	0.106 (0.0484)	0.480 (0.1320)
Correct classification rate	0.586 (0.0210)	0.527 (0.0673)	0.722 (0.0299)	0.396 (0.0984)	0.453 (0.0293)	0.668 (0.0901)
Estimate - Benchmark	-0.305***	0.171**	-0.142***	0.225**	-0.465***	0.114
N	1731	191	862	96	869	95

Note: Estimated tax non-compliance, true and false positive rates, and correct classification rate using different statement phrasings. Columns (1), (3), and (5) ask about tax non-compliance as is our standard phrasing. Columns (2), (4), and (6) instead ask a random sample about tax compliance. All estimates are presented in terms of tax non-compliance so that they are directly comparable. Standard errors are in parentheses. In brackets are p-values from a test of column (1) vs. (2), (3) vs. (4), and (5) vs. (6). The difference between the estimated non-compliance rate and the benchmark rate from administrative data for each column is included at the bottom. Stars denote statistical significance of the estimated difference to the benchmark rate.

TABLE A.15: Statement phrasing experiment: Second-order beliefs

	(1) Pooled	(2) KCCA	(3) URA
Survey 2024			
DV: Est. non-compliance rate			
Compliance-phrasing (dummy)	0.086*** (0.0234)	0.096*** (0.0329)	0.077** (0.0331)
Constant	0.345*** (0.0062)	0.312*** (0.0085)	0.380*** (0.0089)
Observations	1708	871	837

Note: OLS regressions of respondents' second-order beliefs of the tax non-compliance rate among businesses similar to theirs on a dummy denoting whether respondents were randomly asked about tax non-compliance (dummy = 0) or tax compliance (dummy = 1). In the latter case, we invert answers for this table so that both estimates are expressed in terms of estimated tax non-compliance and directly comparable. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.16: Variation in data quality and survey implementation between enumerators

	Reporting consistency	Recall check	Any note taking	Note length	Net interview length
Survey 2024					
ICC Enumerators	0.093 (0.030)	0.503 (0.075)	0.628 (0.092)	0.238 (0.026)	0.162 (0.037)
Observations	1864	1923	1923	502	1923

Note: Intraclass correlations from multilevel models with enumerators as the only specified level. Estimates represent the share of total outcome variation explained by variation *between* enumerators (as opposed to variation between different respondents *within* each enumerator). Columns 1–3 use multilevel mixed-effects logistic regression whereas columns 4–5 use multilevel mixed-effects linear regression. Robust standard errors in parentheses.

TABLE A.17: Between-enumerator variation with controls and parish random effects

	Reporting consistency	Recall check	Any note taking	Note length	Net interview length
Panel A					
ICC Enumerators	0.084 (0.030)	0.481 (0.075)	0.635 (0.092)	0.263 (0.035)	0.159 (0.038)
Observations	1795	1839	1839	484	1839
Controls	YES	YES	YES	YES	YES
Panel B					
ICC Parish	0.011 (0.022)	0.000 (0.000)	0.000 (0.000)		0.049 (0.042)
ICC Parish - Enumerators	0.152 (0.032)	0.487 (0.046)	0.595 (0.057)		0.210 (0.060)
Observations	1795	1839	1839		1839
Controls	YES	YES	YES		YES

Note: Intraclass correlations from multilevel models. In Panel A, estimates represent the share of total outcome variation explained by variation *between* enumerators (as opposed to variation between different interviews *within* each enumerator). In Panel B, we estimate the share of variation explained by between-parish variation and within-parish-between-enumerator variation. Columns 1–3 use multilevel mixed-effects logistic regression whereas columns 4–5 use multilevel mixed-effects linear regression. Convergence issues due to the lower sample size prevent the estimation of column 4 in Panel B. All models control for respondent characteristics (gender, education, business ownership, tax knowledge, Marlowe-Crowne social desirability score) and firm characteristics (log number of staff, a retail dummy, and division). Robust standard errors in parentheses.

TABLE A.18: Enumerator-level correlates of data quality and survey implementation

	(1) Fail consistency check	(2) Fail recall check	(3) Took any notes	(4) Note length	(5) Net interview length
Survey 2024					
Enumerator participated in 2023 survey	0.021 (0.0224)	-0.235*** (0.0228)	0.063*** (0.0219)	74.377*** (13.8787)	5.086*** (0.7886)
Enumerator nbr of previous interviews in 2024 survey	-0.001** (0.0006)	-0.004*** (0.0006)	-0.003*** (0.0005)	-0.151 (0.3362)	-0.172*** (0.0223)
Enumerator nbr of previous interviews on same day	0.001 (0.0061)	0.005 (0.0063)	-0.009 (0.0057)	-4.955 (3.0529)	-0.353 (0.2249)
Observations	1864	1923	1923	502	1923
Division FEs	YES	YES	YES	YES	YES
Enumerator FEs	NO	NO	NO	NO	NO
Mean dep. var.	0.303	0.499	0.261	113.544	24.842

Note: Results from OLS regressions (columns 1, 2, 3 and 5) and from a two-part model with a logit as a first part (any note taken) and a linear regression model as second part (note length; column 4). The dependent variables are a dummy for whether the respondent failed the consistency check (column 1), a dummy for whether the respondent failed the recall check (column 2), a dummy for whether the enumerator took notes at the end of the interview (column 3), the length of the note (column 4), and the net interview length, i.e., excluding survey introduction and consent as well as breaks between modules (column 5). All regressions control for the geographic division of the business. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

TABLE A.19: Enumerator-level correlates of data quality and survey implementation with enumerator fixed effects

	(1) Fail consistency check	(2) Fail recall check	(3) Took any notes	(4) Note length	(5) Net interview length
Survey 2024					
Enumerator nbr of previous interviews in 2024 survey	-0.002*** (0.0006)	-0.001** (0.0006)	-0.002*** (0.0004)	0.152 (0.3442)	-0.121*** (0.0234)
Enumerator nbr of previous interviews on same day	0.000 (0.0061)	0.010* (0.0055)	-0.005 (0.0050)	0.155 (2.4514)	-0.074 (0.2125)
Observations	1864	1923	1923	502	1923
Division FEs	YES	YES	YES	YES	YES
Enumerator FEs	YES	YES	YES	YES	YES
Mean dep. var.	0.303	0.499	0.261	113.544	24.842

Note: Results from OLS regressions (columns 1, 2, 3 and 5) and from a two-part model with a logit as a first part (any note taken) and a linear regression model as second part (note length; column 4). All models contain enumerator fixed effects. The dependent variables are a dummy for whether the respondent failed the consistency check (column 1), a dummy for whether the respondent failed the recall check (column 2), a dummy for whether the enumerator took notes at the end of the interview (column 3), the length of the note (column 4), and the net interview length, i.e., excluding survey introduction and consent as well as breaks between modules (column 5). Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

TABLE A.20: Median posterior SD relative to the average single measure, by number of measures

k	KCCA			URA		
	Median	Min	Max	Median	Min	Max
2	3.40	2.05	4.30	3.99	2.47	5.12
3	2.98	1.60	3.71	3.76	2.59	4.37
4	2.97	2.18	3.20	3.28	2.66	3.83
5	2.70	2.16	2.82	3.17	2.76	3.26
6	2.44	2.44	2.44	2.87	2.87	2.87

Note: For each number of measures k , every combination of k of the six measures is used to refit the hierarchical model, and the resulting posterior standard deviation of the cross-measure non-compliance rate is expressed as a ratio to the average single-measure standard error. Columns report the median, minimum and maximum of that ratio across all such combinations: 15, 20, 15, 6 and 1 combinations for k running from two to six. At $k = 6$ there is only one combination, so the median, minimum and maximum coincide. A value of one would mean the pooled estimate across measures is no more uncertain than a typical single measure. The heterogeneity prior is Half-Cauchy(0, 1), as in the main specification.

TABLE A.21: Robustness of the BHM results to the heterogeneity prior

	Median	Inflation	Ratio RE/FE	Reduction	Mass 10 pp	Break-even
KCCA						
Main: HC(0,1)	0.191	2.442	15.699	0.593	0.660	0.134
Alternative 1: HC(0,2.5)	0.193	2.587	16.630	0.565	0.640	0.133
Alternative 2: IG(0.5,0.5)	0.191	2.407	15.474	0.597	0.664	0.135
Excl. egregious experts	0.193	2.406	15.394	0.581	0.667	0.136
URA						
Main: HC(0,1)	0.292	2.873	15.411	0.549	0.034	0.221
Alternative 1: HC(0,2.5)	0.294	3.000	16.095	0.525	0.040	0.220
Alternative 2: IG(0.5,0.5)	0.290	2.837	15.217	0.552	0.032	0.220
Excl. egregious experts	0.292	2.803	15.245	0.549	0.031	0.222

Note: The table repeats our main analysis (first row) using alternative heterogeneity (between-measure variance) priors: $\tau^2 \sim \text{Half-Cauchy}(0, 2.5)$, $\tau^2 \sim \text{Inverse-Gamma}(0.5, 0.5)$, and the main prior but dropping egregious expert predictions that put more than 20% of probability mass outside the unit interval. Columns summarize our headline results: The posterior median, uncertainty inflation factor (vs avg single measure), RE/FE total-uncertainty ratio, prior-to-posterior SD reduction (total prior), posterior mass within 10pp of the benchmark, and the 80%-certainty break-even rate.

TABLE A.22: Change in posterior uncertainty when allowing between-measure correlations

ρ	KCCA				URA			
	SD(μ)	Infl.	95% CI, pp	Med(μ), %	SD(μ)	Infl.	95% CI, pp	Med(μ), %
0.0	0.555	2.457	35.6	19.3	0.489	2.832	39.6	29.0
0.1	0.569	2.516	36.3	18.7	0.503	2.914	40.6	28.6
0.2	0.577	2.555	36.4	18.3	0.516	2.987	41.0	28.2
0.3	0.594	2.630	37.4	18.1	0.539	3.123	42.6	28.1
0.4	0.600	2.657	36.9	17.8	0.551	3.193	43.5	27.6
0.5	0.616	2.726	37.8	17.5	0.573	3.318	44.8	27.3
0.6	0.639	2.829	39.0	17.4	0.608	3.522	47.0	27.1
0.7	0.667	2.954	40.8	17.1	0.639	3.702	49.5	27.5
0.8	0.701	3.105	42.8	17.3	0.689	3.987	52.3	27.5

Note: Each row refits the hierarchical model with $\theta = \mu + \sqrt{\rho}a + \sqrt{1-\rho}e$, where a is a method-level effect shared by the two phrasing variants of the same elicitation method, e is a measure-specific effect, and both are $N(0, \tau^2)$. The four method-families are the second-order belief pair, the list-experiment pair, the direct question, and the enumerator guess, the last two contributing a single measure each. The weights decompose a measure's total variance around μ into a method-level and a measure-specific component with shares ρ and $1-\rho$, so that τ^2 retains its interpretation as the variance of a single measure around μ at every ρ , the Half-Cauchy(0, 1) prior on τ^2 applies to the same quantity throughout, and ρ is exactly the correlation between two variants of the same method. τ^2 itself is freely estimated at each grid point. $\rho = 0$ is the main specification. The measure-level effects are integrated out analytically, so each row is fitted on the six measure estimates with only μ and τ free. Chains are run for 200,000 draws rather than the 20,000 of the main specification, because the credible-interval endpoints are tail quantiles and require more draws than a posterior standard deviation to be estimated precisely. Columns report the posterior standard deviation of μ on the logit scale on which the model is fitted, that standard deviation divided by the average single-measure standard error (the uncertainty inflation factor), the width of the 95 percent equal-tailed credible interval for μ in percentage points, and the posterior median of μ , both on the probability scale. The percentage increase along the grid is smaller in percentage points than in logit units because the inverse-logit transformation compresses the logit scale unevenly.

TABLE A.23: [Chuang et al. \(2021\)](#) test of list experiment internal validity

	(1) Pooled	(2) KCCA	(3) URA
Panel A: Survey 2023			
List 2 (dummy)	-0.127 (0.0785)	-0.057 (0.1058)	-0.205* (0.1170)
Constant	0.266*** (0.0486)	0.219*** (0.0664)	0.318*** (0.0710)
Observations	1459	775	684
Panel B: Survey 2024			
List 2 (dummy)	0.055 (0.0709)	0.100 (0.1014)	0.012 (0.0990)
Constant	0.093** (0.0432)	0.044 (0.0627)	0.142** (0.0596)
Observations	1731	862	869

Note: Difference in estimated non-compliance rate in list 1 vs. list 2 of double list experiment. The constant captures the estimated non-compliance rate in list 1 whereas the list 2 dummy captures the difference between list 1 and list 2. Clustered standard errors at the firm level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.24: Tabulation of item counts by list

	Pooled		KCCA		URA	
	Short	Long	Short	Long	Short	Long
Panel A: Survey 2023						
List 1						
0	1.4	1.5	0.5	1.5	2.4	1.5
1	22.1	18.7	24.7	17.2	19.1	20.3
2	48.1	44.4	46.1	46.1	50.3	42.4
3	25.3	28.7	25.8	28.4	24.7	29.1
4	3.2	6.6	2.9	6.6	3.5	6.7
5	0.0	0.1	0.0	0.3	0.0	0.0
List 2						
0	2.3	2.5	2.0	2.4	2.6	2.6
1	28.1	19.6	28.6	23.9	27.6	14.7
2	43.7	39.9	44.6	38.4	42.7	41.5
3	21.2	27.5	21.5	25.0	20.9	30.3
4	4.6	10.3	3.3	9.7	6.1	10.9
5	0.0	0.3	0.0	0.5	0.0	0.0
Panel B: Survey 2024						
List 1						
0	3.4	2.7	3.5	2.6	3.4	2.8
1	24.7	18.9	25.8	20.3	23.6	17.5
2	49.8	49.3	51.2	51.4	48.5	47.2
3	19.8	26.4	17.7	23.1	21.8	29.7
4	2.3	2.6	1.8	2.3	2.7	2.8
5	0.0	0.1	0.0	0.2	0.0	0.0
List 2						
0	4.3	3.9	4.9	4.4	3.7	3.4
1	32.9	30.3	31.3	32.5	34.6	28.1
2	40.2	39.9	40.0	36.6	40.4	43.1
3	20.3	21.5	21.3	22.4	19.4	20.6
4	2.2	4.3	2.6	3.9	1.9	4.8
5	0.0	0.1	0.0	0.2	0.0	0.0

Note: Tabulation of item count responses. All values in percent.

TABLE A.25: Blair and Imai (2012) test of list experiment design effects

	P>Lambda	P _{Bonferroni} >Lambda
Panel A: Survey 2023		
List 1 (Pooled) - Pr(R ,S=0)	1	1
List 1 (Pooled) - Pr(R ,S=1)	.4365272	.8730543
List 2 (Pooled) - Pr(R ,S=0)	1	1
List 2 (Pooled) - Pr(R ,S=1)	.4016903	.8033807
List 1 (KCCA) - Pr(R ,S=0)	1	1
List 1 (KCCA) - Pr(R ,S=1)	.0835994	.1671989
List 2 (KCCA) - Pr(R ,S=0)	1	1
List 2 (KCCA) - Pr(R ,S=1)	.3723884	.7447768
Panel B: Survey 2024		
List 1 (Pooled) - Pr(R ,S=0)	1	1
List 1 (Pooled) - Pr(R ,S=1)	1	1
List 2 (Pooled) - Pr(R ,S=0)	1	1
List 2 (Pooled) - Pr(R ,S=1)	1	1
List 1 (KCCA) - Pr(R ,S=0)	1	1
List 1 (KCCA) - Pr(R ,S=1)	1	1
List 2 (KCCA) - Pr(R ,S=0)	1	1
List 2 (KCCA) - Pr(R ,S=1)	.421294	.842588

Note: Test for design effects as in Blair and Imai (2012). The test estimates the probability of each respondent type with the number of types corresponding to twice the number of items in the list (since those who carry the sensitive trait of interest are estimated as a separate type from those who do not). It then tests for trait carriers (S=0) and trait non-carriers (S=1) whether the probability to belong to any type is negative. Rejecting the null of no negative probabilities would constitute a violation of the no design effects assumption. We present simple, unadjusted p-values and the Bonferroni-corrected p-values suggested by Blair and Imai (2012). We cannot perform this test for the URA group as no respondents reported an item count of five.

TABLE A.26: Association of Marlowe-Crowne social desirability scores with self-reported tax non-compliance

	(1)	(2)
	Direct questioning	List experiment
Coefficients on social desirability score		
Non-compliance rate	0.000 (0.0022)	0.013 (0.0082)
True positive rate	0.002 (0.0042)	0.013 (0.0132)
False positive rate	-0.001 (0.0020)	0.014 (0.0105)
Observations	1,707	1,731

Note: Association of respondents' Marlowe-Crowne social desirability score (MC score) with self-reported tax non-compliance. The MC score is a numeric score between 0 and 13 and reflects the number of 'socially desirable' answers to thirteen questions on 'too-good-to-be-true' behaviors and attitudes. Higher MC scores are meant to proxy for a higher propensity of the respondent to answer questions in a socially desirable manner. We regress self-reported tax non-compliance on MC scores and report the coefficient on the score. The first row captures the association of social desirability scores with self-reported tax non-compliance overall, the second row captures the association with the true positive rate and the final row captures the association with the false positive rate. We conduct this exercise for direct questioning (column 1) and for the list experiment (column 2). Clustered standard errors at the firm level are in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

TABLE A.27: Difference in true and false positives by enumerator experience

	Direct			List		
	Questioning			Experiment		
	Pooled	URA	KCCA	Pooled	URA	KCCA
Survey 2024						
Enumerator participated in 2023 survey	0.006 (0.0113)	0.012 (0.0128)	-0.005 (0.0221)	0.056 (0.0572)	0.056 (0.0695)	0.061 (0.1003)
Paid or filed * Participated in 2023 survey	-0.031 (0.0237)	-0.023 (0.0296)	-0.027 (0.0346)	0.037 (0.0881)	-0.013 (0.1573)	0.037 (0.1262)
Paid or filed tax (admin)	0.070*** (0.0152)	0.022 (0.0186)	0.074*** (0.0227)	0.044 (0.0556)	0.034 (0.0917)	0.053 (0.0805)
Constant	0.026*** (0.0065)	0.017*** (0.0065)	0.043*** (0.0142)	0.075** (0.0356)	0.068 (0.0437)	0.085 (0.0622)
Observations	1707	838	869	1731	862	869

Note: Difference in estimated false and true positive rates by enumerator experience. The uninteracted dummy for enumerator participation in the 2023 survey tests whether the false positive rate differs between these enumerators and those who only participated in 2024. The interaction term tests whether the true positive rate differs between both groups. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Appendix B: Supplementary figures

FIGURE B.1: Number of interviews by enumerator

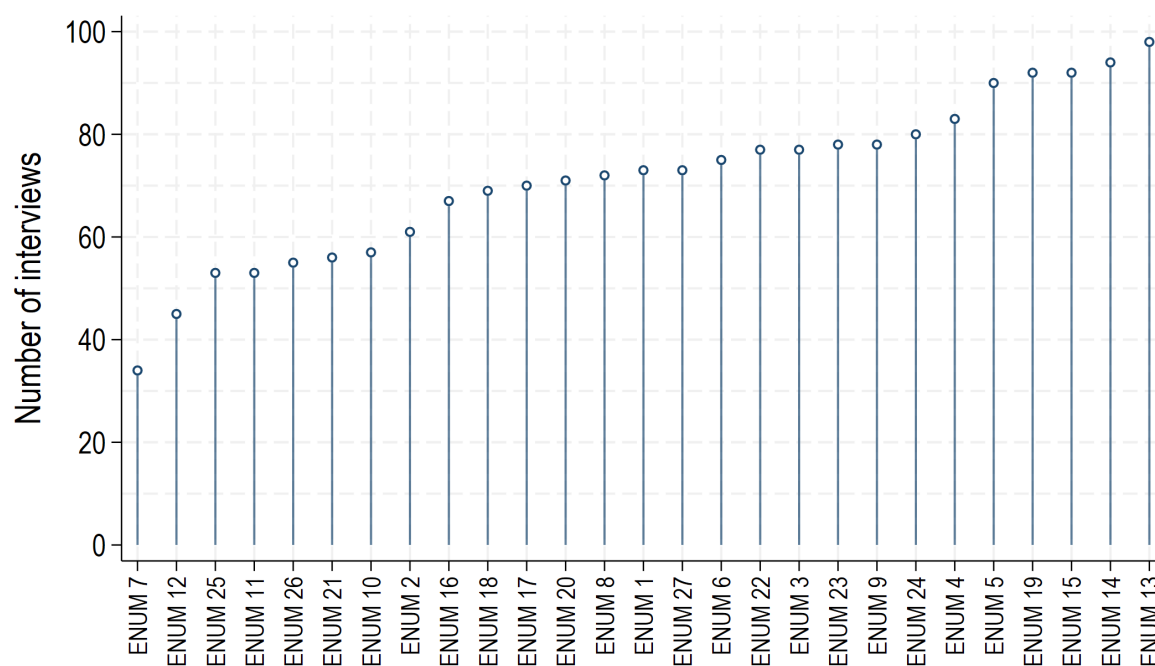
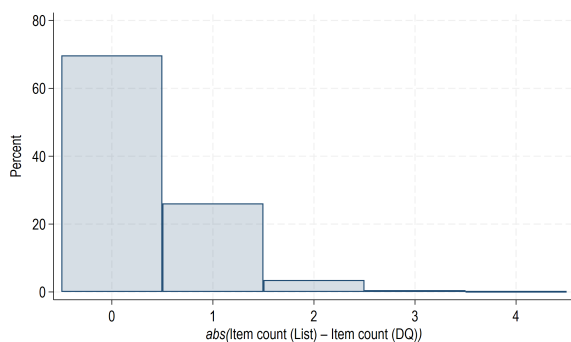
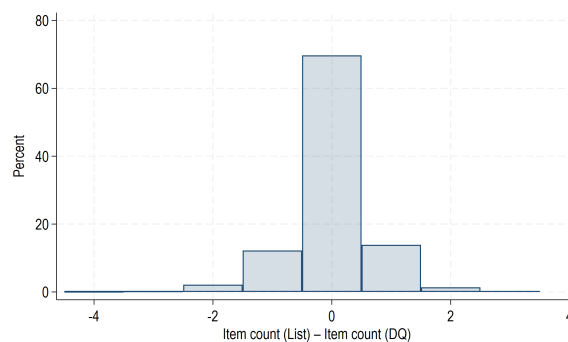


FIGURE B.2: Reporting consistency: Difference in item counts



(A) Absolute difference



(B) List count – direct question count

FIGURE B.3: Phrasing experiment: Histograms of second-order beliefs

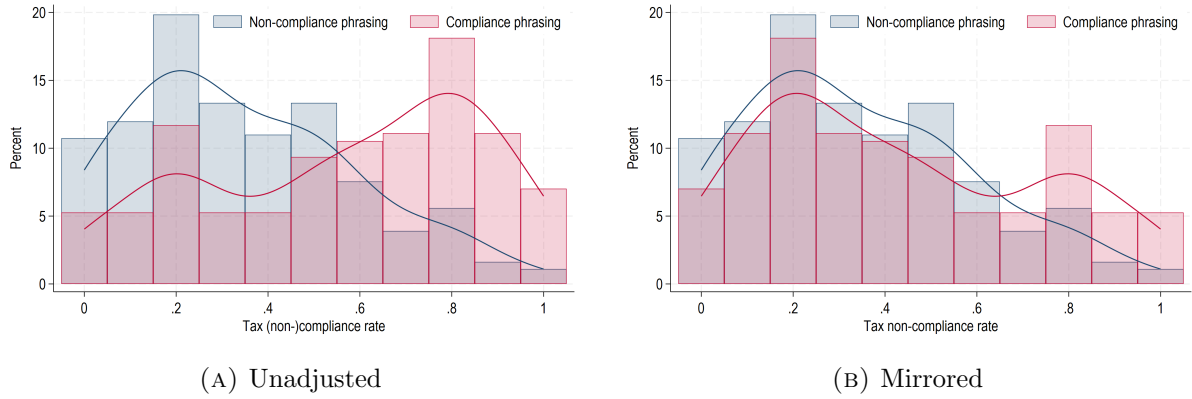


FIGURE B.4: Enumerators and implementation heterogeneity

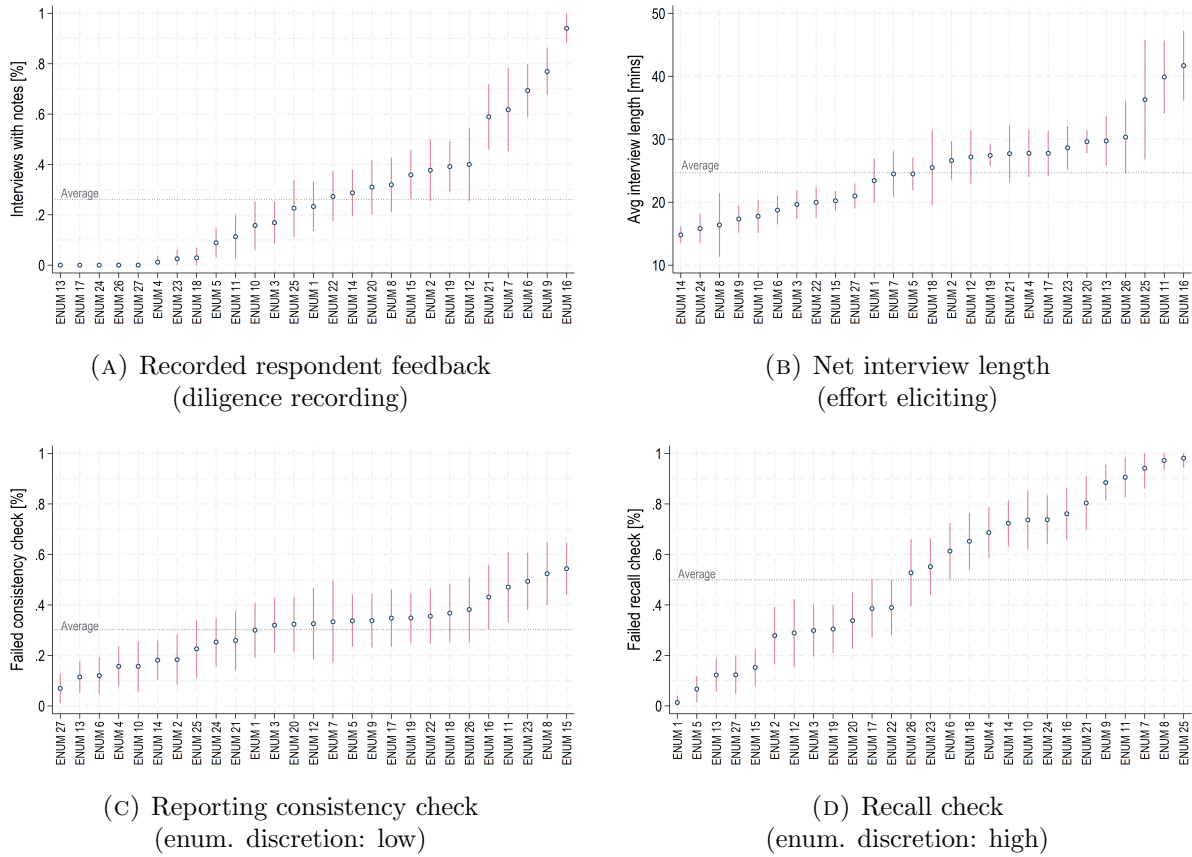
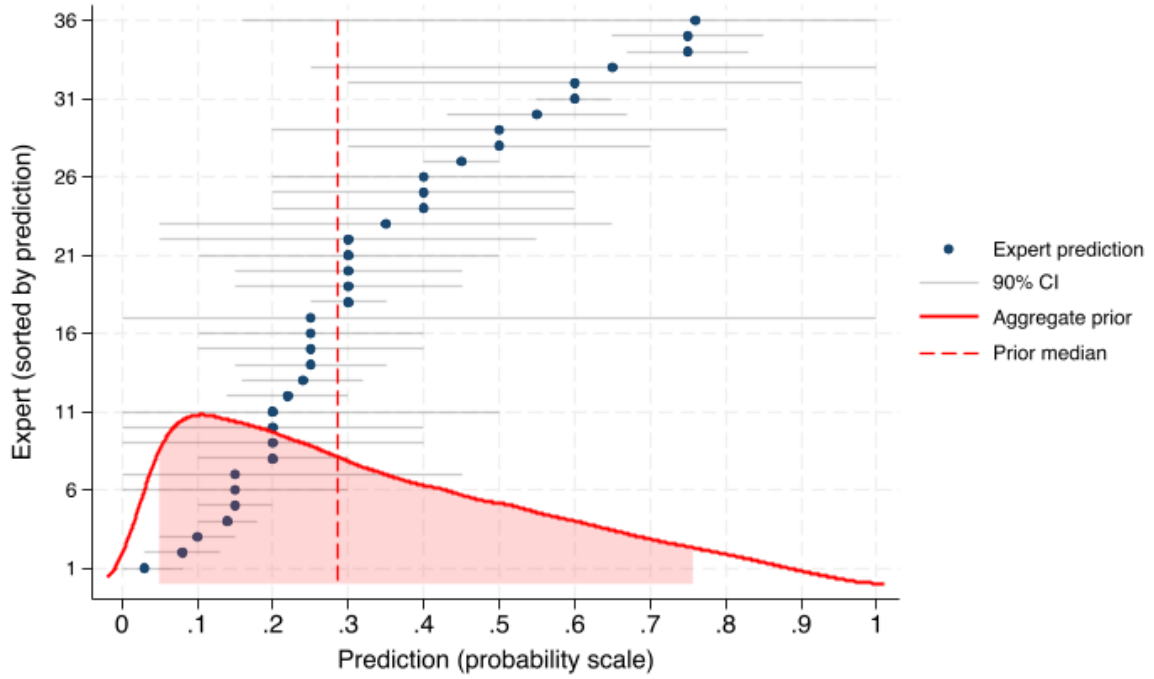
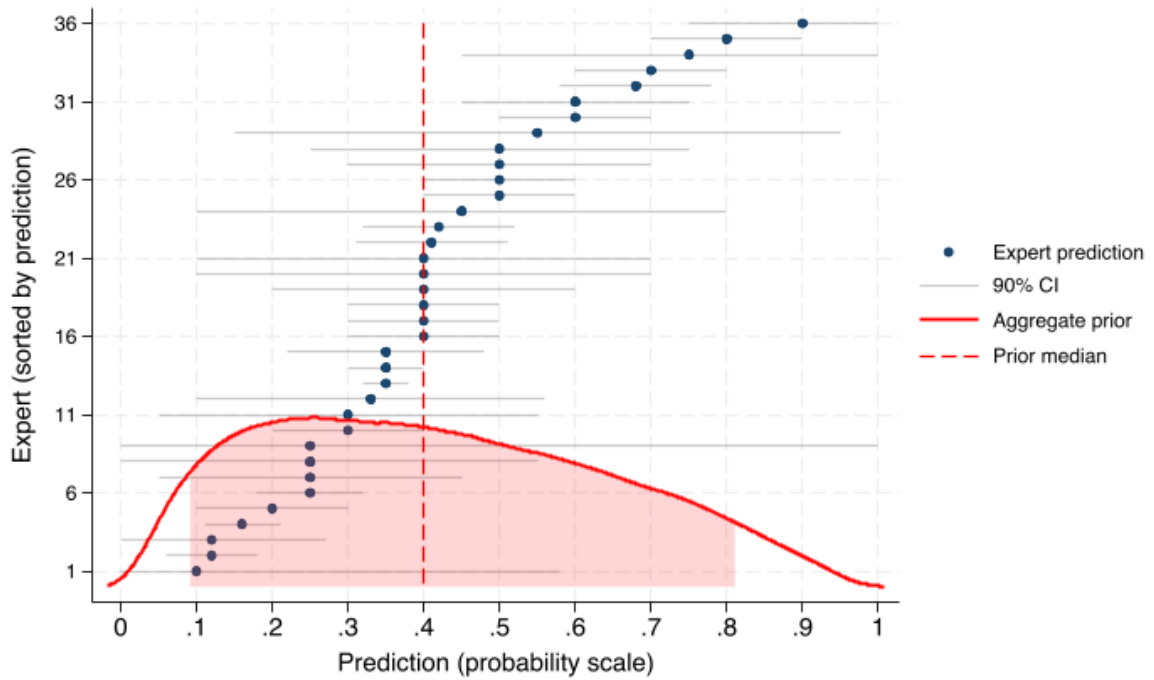


FIGURE B.5: BHM: Expert predictions and aggregate prior



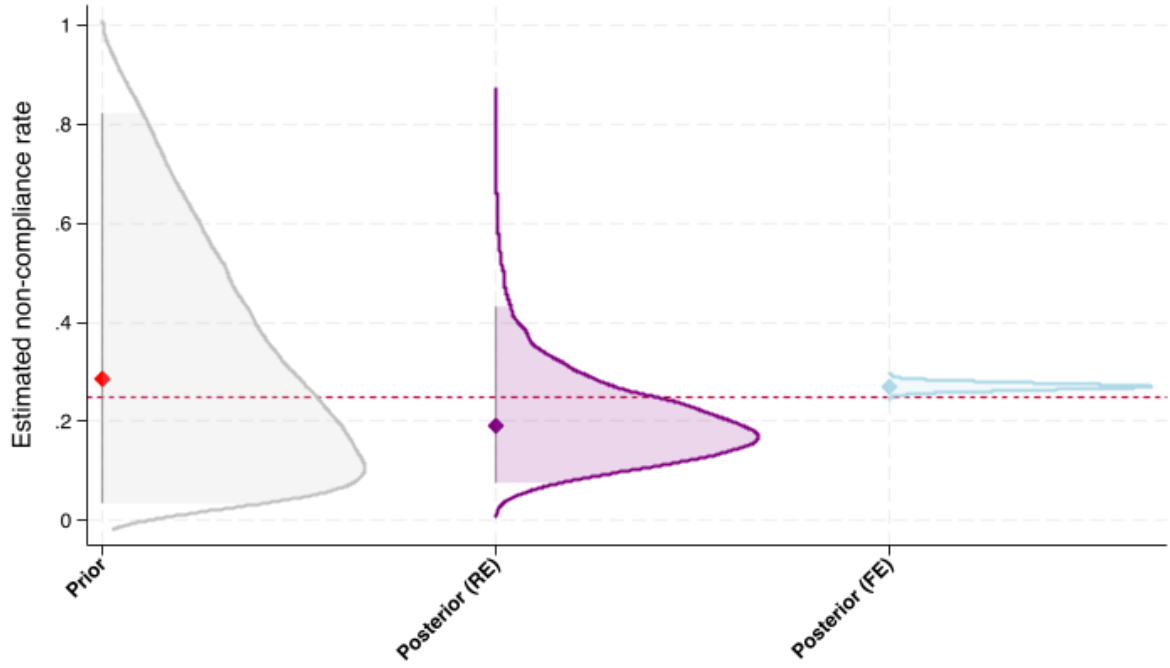
(A) KCCA



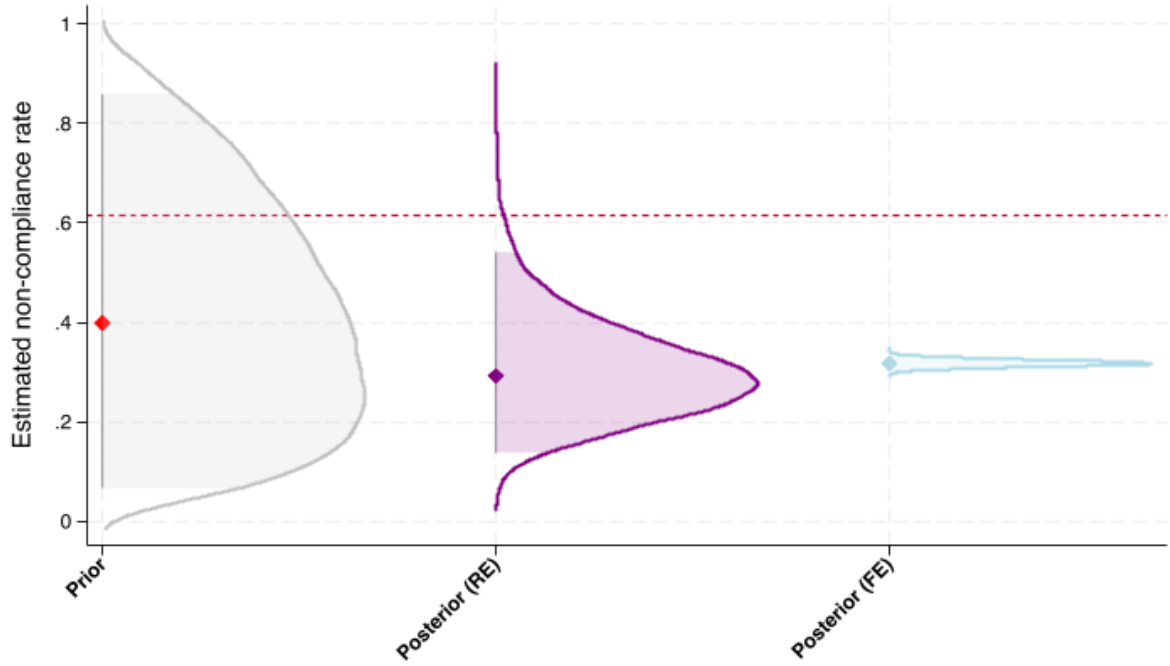
(B) URA

Note: The figure visualizes the 36 ‘expert predictions’ with their associated reported 90%-credible interval and the aggregate prior derived from them. The shaded area demarcates the 90%-credible interval of the prior distribution. All values are shown on the probability scale to ease interpretation. The actual model is fit on the logit scale.

FIGURE B.6: BHM: Prior and posteriors for RE and FE models



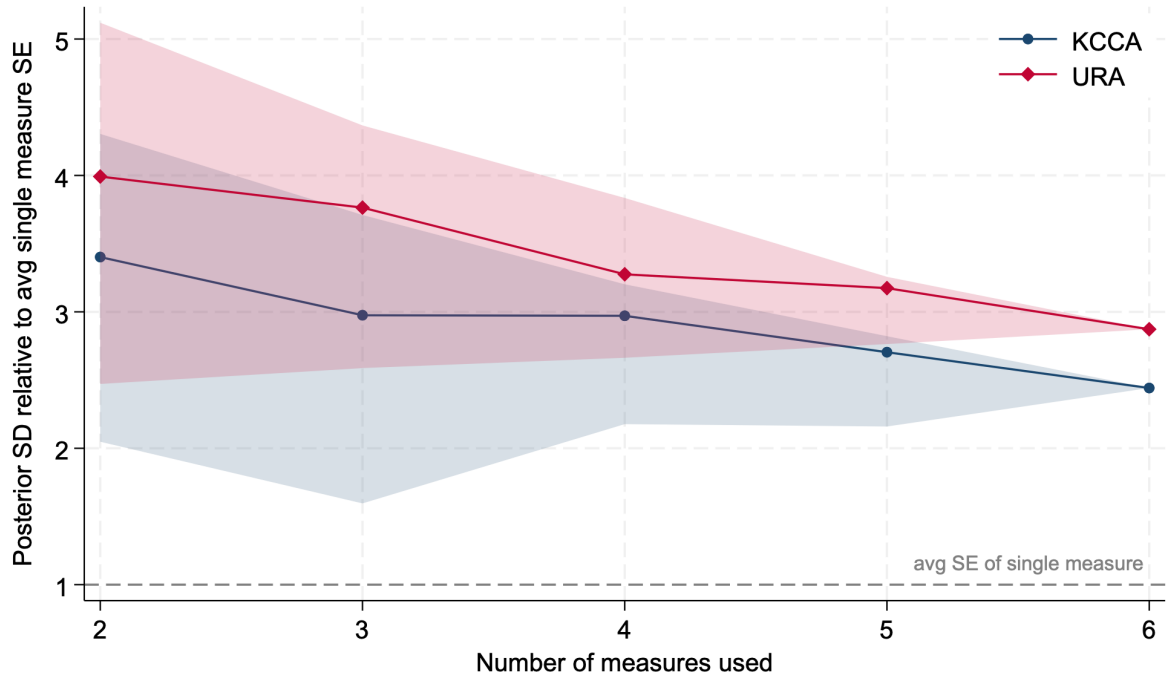
(A) KCCA



(B) URA

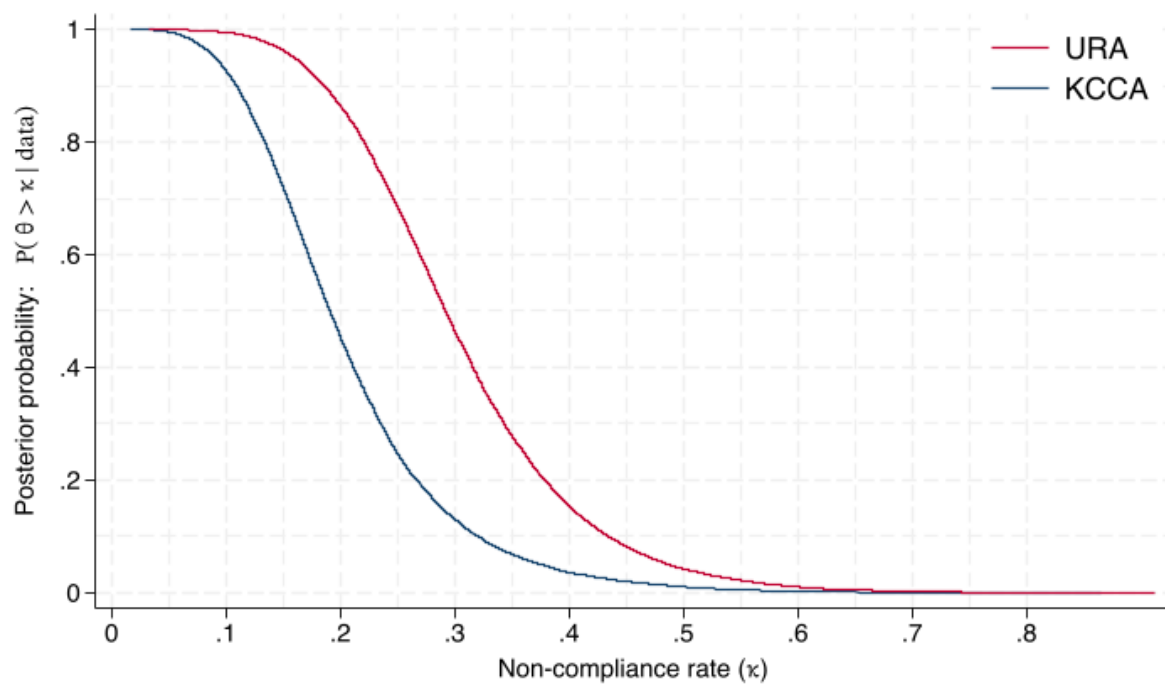
Note: The figure visualizes the prior distribution as well as estimated posterior distributions for two models: our preferred random effects model (RE) which allows for between-measure uncertainty and a fixed effects (FE) model treating all variation between measures as noise, therefore not incorporating measurement uncertainty. All values are shown on the probability scale to ease interpretation. The actual model is fit on the logit scale.

FIGURE B.7: Posterior standard deviation by number of included measures



Note: The figure shows the posterior standard deviation relative to the average standard error of a single measure for different sets of k measures included in the BHM. We create all possible combinations of $k \in \{2, \dots, 6\}$ measures and re-run the BHM with each set of measures. Solid markers show the median posterior standard deviation for combinations of k measures and the shaded region visualizes the minimum and maximum bound.

FIGURE B.8: Posterior probability of exceeding threshold



Note: The figure reports the posterior probability of the non-compliance rate exceeding a threshold κ for different κ .

Appendix C: Additional information on experiments

C.1 List experiment

We design our list experiments in line with established recommendations (Blair et al., 2014; Blair and Imai, 2012). The list of items is chosen such that the items are related to business practices and similar topics and do not make the sensitive statement stand out too much. To avoid corner cases (no or all items are true), which would effectively reveal individual behavior, one item is chosen such that it is expected to have a high prevalence, and one is chosen such that it has a low prevalence among the sample. The remaining two items are chosen such that they are expected to have neither high nor low prevalence. In expectation, this setup should yield a modal answer of two, which safeguards anonymity with respect to individual behavior and maximizes the precision of the estimate of interest.

Under the assumptions that businesses were successfully randomized between the long and short list; that list length (short or long) does not affect answers to the baseline statements (‘no design effect’); and that respondents comply with instructions (‘no liars’), we can arrive at a simple estimate of the aggregate tax non-compliance rate: it is simply the difference in means between the item count in the long list group and the item count in the short list group.

In a double list experiment, this process is performed twice, each time for a different list of base items and switching which respondents get the short list and which get the long list.¹ The computed non-compliance rate is the average between both iterations, which further reduces the variance of estimates. To practically implement this and estimate the TPR, FPR, and CCR, we employ the linear regression estimator developed by Blair and Imai (2012).

In order to reduce respondent burden and limit the scope for strategic behavior in response to the experiment, we only implement the double list experiment for one of two tax levels for each respondent. Therefore, each respondent is only presented with two lists (instead of four had we asked each respondent to complete the double list experiment for each tax level): a short list with four non-sensitive items and a long list with four (new) non-sensitive items *plus* the sensitive item about tax non-compliance with either municipal or national business taxes. We randomize whether respondents start with the short list or long list and whether respondents are asked about municipal or national business tax non-compliance. To help respondents comply with instructions, enumerators walked respondents through an innocuous example and also presented them with a handout that listed the different statements and their answer options in addition to reading them out loud.

C.2 Randomized response technique

In an RRT (Warner, 1965), respondents are presented with the sensitive statement of interest and a randomization device such as a six-sided die. Depending on the outcome of the randomization device (such as rolling a six), respondents are instructed to adjust their answer to the question.

¹Therefore, the double list experiment comprises two lists of non-sensitive items with four items each. To minimize the variance of estimates, items chosen between both lists should be correlated with each other (e.g., Chuang et al., 2021; Blair and Imai, 2012).

In order to make individual behavior unobservable to the researcher and preserve anonymity, the enumerator does not observe the outcome of the randomization device (for example, looks away while the respondent rolls the die). So long as the probability of the respondent being instructed to tell the truth is known and different from 50% (such as the probability to roll a six) and respondents comply with the instructions, this allows us to back out the true prevalence rate. We can write the probability to report having engaged in the sensitive behavior of interest as

$$r = ps + (1 - p)(1 - s) \quad (\text{C.1})$$

where r is the share of respondents who report that 0 or 2 statements are true, p is the probability of a true answer, and s is the “true” share of respondents whose business did not file tax (pay for a trade license). Thus, we can estimate s as

$$s = \frac{r + p - 1}{2p - 1} \quad (\text{C.2})$$

Different variants of the RRT exist depending on the choice of the randomization device. We implement a crosswise design which uses a question whose aggregate probability distribution is known to the researcher but whose individual true value is not (Blair et al., 2015). We follow the canonical example described by Heck et al. (2018) in which the respondent is presented with two statements. The first statement asks whether the respondent was born between May and July (or any other combination of months) while the second statement asks about the business’s tax non-compliance.² Respondents are only asked to answer whether either 0 or 2 (no or both) statements are true *or* one statement is true. To facilitate compliance with the experiment’s instructions, enumerators presented respondents with a handout and two cards of different colors (pink and green). They instructed respondents to show the pink card if none or two statements were true and the green card if exactly one statement was true. This approach proved practical and more easily understandable during piloting. In our case, the question on the month of birth functions as a randomization device unknown to the researcher but for which the aggregate prevalence is known (in most countries, birthdays are nearly uniformly distributed across the year). Equation C.1 then applies in the same way as with a physical randomization device with p as the probability that the respondent is born between May and July.

This set-up has some distinct advantages in our context. First, it does not require a physical randomization device whose outcome needs to be hidden from the enumerator, thus simplifying the interview process. Second, it is easier to comprehend than other variants of the RRT and answer instructions resemble those used in our list experiment. This reduced the complexity of the experiment on the respondents’ side. Third, a crosswise design does not require respondents to lie depending on the outcome of the randomization device, as is the case in other implementations of the randomized response technique which has been shown to lead

²Specifically, the statements were “*I was born in May, June, or July*” and “*My business did NOT file any corporate or presumptive tax to URA [did NOT pay for a new or renewed trade license to KCCA] in 2022*”.

to non-compliance with instructions. We implement the RRT and estimate the TPR, FPR, and CCR using the regression estimator developed in [Blair et al. \(2015\)](#).

Similar to our approach for the double list experiment, we conduct the RRT only once with respect to one of the two tax levels (municipal or national business tax). Respondents randomized into completing the LE with respect to the firm’s compliance with municipal taxes completed the RRT regarding national taxes and vice versa. All respondents, therefore, completed both the double LE and RRT in 2023, but no respondent completed any single experiment more than once; or both with respect to the same tax level.

C.3 Bayesian Truth Serum (BTS)

Respondents in the BTS treatment group are told that their answers will be scored according to a mathematical formula that gives a higher score for more truthful and more accurate answers and that they will receive a financial reward depending on their score. Mathematically, the score is calculated as

$$\begin{aligned} C &= iscore + score \\ &= \sum_{k=1}^2 x_k^r \log \left(\frac{\bar{x}_k}{\bar{y}_k} \right) + \alpha \sum_{k=1}^2 \bar{x}_k \log \left(\frac{y_k^r}{\bar{x}_k} \right) \end{aligned} \quad (C.3)$$

where subscript k denotes the two answer options ($k = 1$ for having complied and $k = 2$ for not having complied with tax regulations) and x_k^r is equal to 1 for answer k given by respondent r to the tax compliance question and 0 for the other answer option. Further, $\bar{x}_k$ is the share of businesses that said that they did (not) pay their taxes for respondents who also said that their business did (not) pay their taxes, $\bar{y}_k$ is the (geometric) average of predicted frequencies for each answer option; and y_k^r is respondent r ’s prediction of how many people did ($k = 1$) or did not ($k = 2$) pay their taxes. The constant α is a weight parameter that decides how much weight to give to the second term, the accuracy of respondents’ prediction of the distribution of answers in aggregate.

The first part of the formula is the so-called “information score” (“iscore”). The information score is maximized for answers that are “surprisingly common”; that is, answers that turn out to be more common on aggregate than what had been collectively predicted. Following Bayesian reasoning, respondents (not) complying with taxes can be expected to predict higher shares of (non-) compliance than their counterparts. This is because, with Bayesian reasoning, the respondent’s own tax compliance behavior functions as an “informative sample of $N=1$ ” for their prediction of the population prevalence of tax compliance ([Prelec, 2004](#)). As a result, any respondent can expect that their true behavior should stand the highest chance of being the surprisingly common (and thus payout-maximizing) answer. The second part of the formula relates respondents’ predictions to their actual prevalence and rewards respondents for accurately predicting the stated prevalence of tax (non-) compliance in the survey.

Appendix D: Validity tests for survey techniques measuring sensitive behavior

We conduct a number of common validity tests to investigate whether these would have reliably alerted us to the breakdown of survey techniques for sensitive behavior.

First, [Chuang et al. \(2021\)](#) propose to exploit the double list experiment setup to test whether the estimated prevalence of sensitive behavior is statistically indistinguishable in both lists. If the assumptions of the list experiment are met, both lists (differing only in the base items they contain) should provide the same estimate barring sampling error. While true, the list experiment estimator is commonly underpowered to reject even substantial differences in estimated rates between both lists. In our case, even though our sample size is reasonable and differences between both lists at times sizable (e.g., 12.7 percentage points in the 2023 pooled sample), we cannot reject the null of no difference between them (Table A.23). More importantly, even a large number of false positives can be consistent with no differences between both lists whatsoever if misreporting is similar between both lists. For example, this is evident in our estimates of the non-compliance rate with URA business taxes in 2024 which differ by only 1.2 percentage points between both lists, despite an accuracy that does not even beat a random guess.¹

Second, we conduct the tests of ceiling and floor effects (Table A.24) and for design effects in the list experiment proposed in [Blair and Imai \(2012\)](#) (Table A.25). In neither case would the tests suggest reasons for concern.

Third, we include the 13-item short form of the Marlowe-Crowne social desirability module in the 2024 survey. This scale is commonly used to “control for” respondents’ propensity to report in a socially desirable manner in regressions (see, e.g., [Dhar et al., 2022](#)).² We then test whether—in line with the premise of scores being predictive of socially desirable answers—higher social desirability scores are associated with lower self-reported tax non-compliance. Further, we test whether respondents with higher scores have lower true positive rates (fewer ‘admissions of guilt’) and whether there is any association with false positive rates (non-strategic misreporting). We find no statistically significant correlation between higher social desirability scores and self-reported tax non-compliance. Coefficients are small and, if anything, suggest that respondents with higher social desirability scores were slightly *more* likely to report having not complied with taxes, driven by higher rates of non-strategic misreporting among them in the list experiment (Table A.26).

¹[Chuang et al. \(2021\)](#) also propose an internal validity test for the randomized response technique premised on setting $p = 0.5$ in equation C.1 and observing whether the RRT recovers an estimate not significantly different from 0.5. While we do not implement this test, it suffers from the same limitations. Statistical power is often low and passing the test can still be consistent with substantial measurement error. For example, any non-strategic misreporting that takes the form of picking an answer at random precisely biases the estimate toward 0.5.

²The module asks respondents 13 questions on behaviors and attitudes that are considered ‘too good to be true’. It then counts the number of questions that were answered in the socially desirable manner, assigning a score between 0 and 13 to each respondent. Higher scores are meant to proxy for a higher propensity to report in a socially desirable manner.

Appendix E: Additional information on enumerator effects

An exploration of measurement errors in survey data is incomplete if it disregards that measurement-as-designed may differ from measurement-as-implemented. Critically, if implementation fidelity differs between measurement options, this may result in measures yielding different estimates in practice even when they would result in the same estimate when implemented as designed. We exploit variation in the assignment of our 27 enumerators to interviews during the 2024 survey as a window into the survey implementation process.

We find evidence for substantial variation in enumerator behavior across two dimensions: diligence in recording information (proxied by whether optional notes were taken and the length of notes) and effort in eliciting answers (proxied by net interview length¹) (Figures B.4a and B.4b). On average, enumerators took notes at the end of 26% of interviews. When enumerators took a note, the average length was 114 characters, roughly corresponding to the length of one or two sentences.² Five enumerators did not take any notes whereas an equal number took notes at the end of 60% or more of their interviews. When taking notes, half of the 22 enumerators with any notes wrote fewer than 70 characters on average. One particularly diligent enumerator took notes in 94% of their interviews with an average length of 357 characters per note taken. We also find net interview length to vary between enumerators. On average, enumerators spent a net time of 25 minutes interviewing respondents. Eight enumerators went at a fast pace and took under twenty minutes on average. Twelve enumerators took between 25 and 30 minutes whereas for two enumerators, the average net interview time lasted for over 40 minutes.

Having observed that enumerators appear to differ in their survey administration, we next ask whether this matters for data quality and measurement uncertainty. Using the consistency and recall checks, we show that response quality varies between enumerators depending on the discretion enumerators have for deviating from standard protocol.

For both the consistency and recall check, we find appreciable between-enumerator variation in the share of respondents failing the check (Figures B.4c and B.4d). However, this variation is much greater for the recall check than for the consistency check. For the consistency check, average failure rates by enumerator vary between 7% and 54% and most cluster tightly around the overall average of 30%. Conversely, failure rates for the recall check are dispersed across the entire domain and range from almost no respondent failing the check (1%) to nearly everyone failing (98%). As a result, the standard deviation of enumerator-level failure rates for the recall check is 2.4 times that of the consistency check.

We formally confirm these patterns by decomposing overall variance into the share explained by between- versus within-enumerator variance using a multilevel mixed-effects model (Table A.16). Enumerators account for a large share of the total variation precisely where they had more discretion in terms of survey implementation. Most importantly, enumerators account for 50% of variation in data quality for the recall check but less than 10% of variation for the

¹That is, excluding any breaks in between modules that respondents needed to take, for example, to tend to customers

²A typical example from our survey corresponding to this length was: “KCCA should clean the city and work on roads. URA should be transparent and show accountability for taxes paid.”

consistency check. These patterns are strongly indicative of differences in enumerator behavior that matter most where enumerators have room for discretionary choices in the way they administer the interview.

Accounting for the potential non-random allocation of enumerators to interviews does not change these findings meaningfully: Given the assignment mechanism of enumerators to interviews (Section 3.5.2), the most plausible alternative explanation for enumerator effects is geographic sorting of respondents by characteristics correlating with response quality. Notably, these characteristics would also need to correlate with enumerator discretion across methods in order for us to misattribute the patterns in Table A.16 to enumerator implementation heterogeneity. We address this concern in two ways (Table A.17). First, we control for a set of respondent (gender, education, business ownership, tax knowledge, Marlowe-Crowne social desirability score) and firm characteristics (log number of staff and a retail dummy) and find our results to hold in both direction and magnitude (Panel A). Second, we additionally include parish-level random effects alongside enumerator random effects (Panel B). Parishes are quite granular geographical units for our survey with a median of nine interviews per parish. Virtually all of the explained variation in data quality pertains to enumerators.

What factors influence enumerator behavior? We investigate three hypotheses. First, we compare the interviews of ten enumerators who had already participated in the 2023 survey to those of the remaining enumerators who only joined the survey in 2024. Based on conversations with the survey firm, enumerators who participated in both surveys differ from other enumerators in two ways: (i) they had prior experience with a similar survey tool and fieldwork context; (ii) they had performed well during the 2023 survey in the eyes of the survey firm and were therefore prioritized from the survey firm’s enumerator pool to participate again in 2024. In addition, we investigate the role of enumerators learning over the course of the survey (number of interviews previously conducted) and of enumerator fatigue (number of interviews previously conducted on the same day).

We find a robust association between repeat enumerators and survey implementation (Table A.18). Enumerators who had already participated in the 2023 survey were 24 percentage points (40%) less likely to have had a respondent fail the recall check, 6 percentage points (27%) more likely to take notes, took notes that were on average 74 characters (91%) longer³, and spent 5 additional minutes (+22%) interviewing respondents. We also find indicative evidence for learning effects over the course of the survey. Using a model with enumerator fixed effects (Table A.19), we estimate that once enumerators reached the average number of total interviews an enumerator conducted (71 interviews), their respondents were 14 percentage points less likely to fail the consistency check, 7 percentage points less likely to fail the recall check; that enumerators were 14 percentage points less likely to take notes and that they conducted interviews that were about 8.5 minutes shorter than at the start of data collection. We do not find a significant association between the number of interviews an enumerator had already conducted on a day and data quality or survey implementation. When we test whether some enumerators were able

³This result is driven by one experienced enumerator who took very long notes. Excluding this enumerator, there is no longer a significant difference between both groups.

to collect more accurate data on tax non-compliance using either direct questioning or a list experiment, we do not find any statistically significant differences in false or true positive rates between enumerators who had already participated in the 2023 survey and those that were new to the survey in 2024 (Table [A.27](#)).

Appendix F: Practical guidelines for applied researchers

This appendix provides guidance on incorporating measurement uncertainty into applied work. The guidelines roughly follow the order in which the relevant decisions actually arise in a research project:

- [F.1](#) Deciding whether the problem applies to you
- [F.2](#) Assembling a menu of measures
- [F.3](#) Fielding the measures in your survey
- [F.4](#) Pre-registration
- [F.5](#) Estimation
- [F.6](#) Probing robustness
- [F.7](#) Reporting results
- [F.8](#) Guidance beyond descriptive survey settings
- [F.9](#) Practical questions that recurred when we presented this work (FAQ)

F.1 When is measurement uncertainty a relevant concern?

What is measurement uncertainty? Measurement uncertainty occurs when researchers single out one particular approach to measuring a target variable or outcome and leave aside other plausible options they could have chosen instead. For example, this could mean choosing among alternative ways to elicit risk preferences or of asking a sensitive question; or it could mean choosing one data source over another that covers the same target quantities (such as different remote sensing products). Measurement uncertainty is distinct from sampling uncertainty which is what conventional standard errors capture. Standard errors indicate how much an estimate would change when drawing a fresh sample from the same population using the *same* measure. Conversely, measurement uncertainty describes how much the estimate would move if the researcher had made a *different, equally defensible* measurement choice on the same sample. Relative to the total survey error framework (Biemer, 2010), which concerns the sources of error conditional on a chosen design, measurement uncertainty adds a layer above. That is, measurement uncertainty is the uncertainty inherent in choosing a particular design in the first place and conditioning all subsequent inference on that one design.

Screening questions. Measurement uncertainty is relevant for your study if you answer yes to all three questions:

- ☐ **Is there more than one methodologically defensible way to measure the target quantity?**

This is the case for most quantities that are not directly observable (preferences, beliefs, attitudes, sensitive or stigmatized behavior) and wherever no single approach has been

established as a validated ‘gold standard’. Variations may comprise different elicitation methods or other design choices such as framing or phrasing.

□ **Is the relative accuracy of alternative approaches hard to verify in the research setting?**

If an individual-level benchmark is available (even for a subset) and it unambiguously identifies a measurement approach as the first-best, there is no measurement uncertainty. However, this will often not be the case and benchmarking requires caution, for example, when relying on aggregate heuristics such as the notion that higher estimates of sensitive behavior must reflect more valid measurement. In sum, many applied measurement problems will allow for alternative design choices without a clear benchmark, or best-practice guidance from strong validation studies, to settle them.

□ **Could the headline estimate plausibly have differed with a different measurement approach?**

If several approaches would converge on the same expected value in arbitrarily large samples, there is no measurement uncertainty to account for. Note that this is an empirical question and not one that can be settled by assertion.

We would also like to emphasize the following:

1. Measurement uncertainty is there whether or not you measure it.

Testing for it does not create it and does not penalize you for diligence. Committing to a single measure simply ignores the uncertainty and reveals but one point in the range of results the study could have plausibly produced.

2. Measurement uncertainty does not shrink with sample size.

Sampling uncertainty vanishes as N grows. However, the same is not true for measure-specific biases. Measurement uncertainty declines only with more consistent measurement between plausible alternatives, either because estimates turn out to be robust across them, or because validated best-practice tools have shrunk the set of plausible alternatives.

3. Accounting for measurement uncertainty can also make inference *tighter*.

Our framework allows us to exploit a much greater wealth of information than single-measure inference. Several measures may remove noise that is inherent in a single measure. The prior encodes external information—such as qualitative observation, plausible ranges from previous literature, or expert prediction—which is otherwise hard to incorporate into inference in a principled manner. Where plausible measures broadly agree, the total information gain can dominate the widening from $\tau^2 > 0$ and produce a narrower posterior than any single measure’s confidence interval. A wide posterior is therefore a finding (about the measurement problem) and not a cost imposed by the model per se.

4. Measurement uncertainty is broadly relevant.

The guidance we provide is written with descriptive and prevalence estimates from survey data in mind. However, our framework also applies to experimental work whenever treatment may interact differentially with outcome measures, and to non-survey settings in which researchers choose among plausibly equivalent measurement or analytical approaches (Section F.8).

Minimum viable implementation

1. Add at least one plausible alternative to your primary measure that differs substantially in design.
2. Pre-register the menu of measures and the rule for aggregating them.
3. Extract one estimate and standard error per measure.
4. Build a location prior from prior literature, qualitative evidence, or elicited predictions. This should be constructed independently of the data.
5. Fit the hierarchical model with a weakly informative prior on τ^2 (e.g., via Stata's `bayesmh` or R's `baggr`).
6. Report every individual measure, the prior, the posterior median and 95% credible interval, the ratio of the posterior standard deviation to the average single-measure standard error, and the sensitivity of the posterior to the menu of measures.

F.2 Step 1: Assembling the menu of measures

The goal is to map the uncertainty space efficiently via a small number of measures that together give a good picture of how much the estimate depends on measurement choices. Three criteria determine which measures belong on the menu.

Plausibility. Include measures that are methodologically defensible and among which genuine uncertainty exists about which is most accurate. Do not narrow the menu prematurely on the basis of untested preferences. If measurement uncertainty genuinely matters little, the posterior will reflect this. High-quality validation studies are the appropriate basis for excluding an approach.

Coverage. Each measure should represent a substantial departure in design from the others, so that together they span the key measurement choices you faced. What counts as a key choice is context-specific. In our setting it included the elicitation format (direct question, indirect questioning technique, belief elicitation, third-party assessment) and the framing of the sensitive statement (as measuring compliance or *non-compliance*). A useful test: write down the choices you deliberated when designing the instrument, and make sure the menu varies each of them.

Independence. Two nearly identical measures contribute little new information about τ^2 . Worse, if measures share a common methodological component and this correlation is left un-

modeled, the model treats them as independent pieces of evidence, overstates the effective number of measures, and understates posterior uncertainty. Where measures share methodological machinery you may group them into method-families and model the correlation explicitly (Section F.6).

How many measures? There is no universal number and the right one depends on how large the space of plausible alternatives is. Our results support three pragmatic statements.

- **Going from one measure to two is where much of the correction happens.** In our application, even the most agreeing pair of measures—the worst case across all pairs—closes roughly three-quarters of the gap between the average single-measure standard error and the posterior standard deviation obtained from the full set of six (Table A.20). This may not hold in every setting but highlights the usefulness of making measurement uncertainty estimable *at all* by collecting at least one additional measure.
- **Small menus are noisy and depend on which measures you happened to pick.** With $k = 2$, the posterior standard deviation in our application ranged from 2.1 to 5.1 times the average single-measure standard error depending on the pair chosen. Some pairs are too conservative rather than too confident. The spread narrows steadily as measures are added (Figure B.7).
- **More measures can therefore sharpen, not just widen, inference.** Additional independent measures reduce this noise and tend to tighten the posterior.

F.3 Step 2: Fielding the measures

Within or between respondents? Both designs are viable and we use both. Fielding all measures to all respondents makes full use of the sample for every measure but raises two issues. First, contamination: respondents who recognize that they are being asked the same thing twice may anchor on, or seek consistency with, their earlier answer. Where this is likely (most obviously when the same question is re-asked with only a small variation), we recommend randomizing the variant between respondents instead. For example, we do so for the phrasing experiment (Section 3.5.1). Second, measures fielded to the same respondents may have correlated sampling errors, which the baseline model assumes away. Verifying and adjusting for this via a joint bootstrap is inexpensive (Section F.6).

Where measures are fielded within respondents, we recommend ordering them deliberately: place those measures first that make the topic least salient and best protect the anonymity of individual answers, place direct questions last, and randomize the order among measures that are equivalent in salience and privacy protection. This limits both priming and any incentive for respondents to reconcile later answers with earlier ones.

Reduce the errors you can reduce. Measurement uncertainty falls when plausible measures agree, so mitigating known error sources has an inferential payoff over and above what a larger

sample can deliver. Attend to well-established error modes such as question complexity, respondent burden and fatigue, or experimenter demand. Note that these considerations may trade off against each other. For example, a technique that improves anonymity may also raise cognitive demands. The net effect this has on data quality cannot be reliably judged from untested assumptions or from aggregate estimates. In our application, techniques designed to elicit truthful answers to a sensitive question raised false positives roughly as much as true positives and left overall accuracy no better than a direct question.

Treat implementation fidelity as a design variable. Measurement-as-designed can differ from measurement-as-implemented, and measures differ systematically in how much they depend on the enumerator. In our survey, enumerators accounted for under 10% of the variation in a data quality check in which they had little discretion, but around 50% in one where they were responsible for conveying the information respondents needed (Section 4.5). When a measure leans heavily on enumerator discretion, it inherits that heterogeneity. Intensive enumerator training and oversight, in-built data quality checks, and at-scale audio audits may help compress this component of between-measure variation.

F.4 Step 3: Pre-registration

Pre-registration eliminates specification search by committing to a measure ex-ante. However, committing to a single measure is exactly what keeps measurement uncertainty out of view. The two are complementary rather than in tension, provided the pre-analysis plan commits to the full aggregation procedure and not just to one measure. Pre-specify:

1. **The primary measure**, as you would anyway.
2. **The full menu of plausible alternatives**, with the design dimension each one varies.
3. **The aggregation rule**, meaning all four components of the model in Section F.5: likelihood, measurement uncertainty model, location prior, and heterogeneity prior.
4. **How the location prior will be constructed**: the source, the elicitation instrument if you are eliciting, and the aggregation formula, so your hands are tied before seeing the data.
5. **The robustness checks**: sensitivity to the menu of measures, to assumed between-measure correlations, and to the priors (Section F.6).

Other good practices in writing a pre-analysis plan apply as before.

F.5 Step 4: Estimation

The model has four components (Section 5.1.2):

$y_k \sim N(\theta_k, \sigma_k^2)$	[likelihood]
$\theta_k \sim N(\mu, \tau^2)$	[measurement uncertainty model]
$\mu \sim N(\mu_0, \sigma_0^2)$	[location prior]
$\tau^2 \sim \text{Half-Cauchy}(0, 1)$	[heterogeneity prior]

(1) Likelihood: one estimate and one standard error per measure. Each measure k contributes an estimate y_k of the same target quantity and its standard error σ_k , computed exactly as you would if that measure were your only one. If the parameter is bounded (e.g., a proportion), transform to an unbounded scale (e.g., logit) before estimation since the normal distributions above have unbounded support. Where an estimate sits near a boundary, we treat it as defining a truncated normal on the probability scale, draw simulated values by the inverse CDF method, and transform the draws (Section 5.1.3). Report results back on the original scale.

(2) Measurement uncertainty model: how the different measures relate. The θ_k are modeled as draws from a population of plausible measures with mean μ and variance τ^2 . This implies two things. First, estimates are partially pooled, with measure k receiving weight proportional to $1/(\sigma_k^2 + \tau^2)$. As τ^2 grows, differences in the precision of individual measures matter less and weights converge toward equality. This is the sense in which accounting for measurement uncertainty stops a single very precise but possibly badly biased measure from dominating. Second, $\tau^2 > 0$ adds posterior uncertainty about μ that does not disappear even if every measure were estimated without sampling error. Setting $\tau^2 = 0$ recovers a fixed-effects specification, which is what conventional single-measure practice implicitly assumes.

(3) Location prior: what you believed before. The prior on μ can draw on previous studies, qualitative observation, or elicited expert predictions. It must be constructed independently of the data you are analyzing. If you elicit predictions (we use the Social Science Prediction Platform), ask each respondent for both a point prediction and a measure of their own uncertainty. We ask for the value v such that they are 90% certain the truth lies within $\pm v$ of their prediction, implying a standard deviation of $v/1.645$. Aggregate as

$$\mu_0 = \text{median}_j(m_j), \quad \sigma_0^2 = \underbrace{\text{Var}(m_j)}_{\text{between-expert}} + \underbrace{\frac{1}{J} \sum_j s_j^2}_{\text{avg. within-expert}},$$

where m_j and s_j are respondent j 's prediction and implied standard deviation on the estimation scale. Both these terms matter for the variance prior on μ . Predictors disagree with one another, and each expert is individually uncertain. Absent a strong evidence base, a relatively diffuse prior is the appropriate default, so that the data determine the posterior.

(4) Heterogeneity prior: how much measurement uncertainty you expect. Unless you have genuine prior information about how much measurement uncertainty to expect, use a weakly

informative prior diffuse enough to accommodate both negligible and substantial values, in line with standard recommendations for variance priors (Gelman, 2006). We use Half-Cauchy(0, 1) (Section 5.1.6).

Estimation in practice. We estimate by MCMC with 20,000 posterior draws after 5,000 burn-in iterations using Stata’s `bayesmh`. Equivalent models can be fitted in R with the `baggr` package (Meager, 2019). We suggest reporting the posterior median as the point estimate and the posterior standard deviation or 95% credible interval as the measure of spread. If you report credible-interval endpoints under a range of specifications, note that these are tail quantiles and need considerably more draws to be estimated precisely than a posterior standard deviation does. We use 200,000 draws for the correlation grid in Table A.22.

F.6 Step 5: Robustness

A number of inexpensive robustness checks help investigate the sensitivity of results to modeling choices and assumptions.

- **Sensitivity to the menu of measures.**

Refit the model on every combination of k of your measures and plot the posterior standard deviation against k (Figure B.7, Table A.20). This shows readers how much the conclusion depends on the particular measures you happened to field, and whether you are in the region where adding measures still changes the answer.

- **Between-measure correlation.**

Group measures that share methodological machinery into method-families and allow measures within a family to be correlated through a common method effect, $\theta_k = \mu + \sqrt{\rho} a_{m(k)} + \sqrt{1 - \rho} e_k$. Because each measure yields a single estimate, ρ typically cannot be credibly estimated from the data. We suggest varying it over a grid and reporting how the posterior responds (see Table A.22). Positive correlation widens the posterior, so a specification assuming independence is a lower bound on total uncertainty.

- **Heterogeneity prior.**

This is the prior least disciplined by data, so it warrants explicit sensitivity analysis. Refit under alternatives, such as those that put more mass on large τ^2 or that do not permit $\tau^2 \rightarrow 0$. We use Half-Cauchy(0, 2.5) and Inverse-Gamma(0.5, 0.5) (see Table A.21).

- **Location prior.**

Report how the posterior responds to reasonable alternative constructions. Where elicited predictions imply substantial probability mass outside the parameter’s support, report results with and without them. We drop predictions implying 20% or more mass outside the unit interval as a robustness check.

- **Correlated sampling errors.**

The likelihood assumes $\text{Cov}(y_k, y_j \mid \theta) = 0$, which need not hold exactly when measures are fielded to the same respondents. A possible solution is to estimate the sampling covariance between measures via a joint bootstrap and compare it to τ^2 . In our case, the largest covariance was under half a percent of τ^2 , so the effect on results was negligible.

F.7 Step 6: What to report

Focus discussion on the whole shape of the posterior. Our framework provides an uncertainty quantification and not a de-biasing tool. Rather than focusing on a measure of central tendency (such as the posterior median), we suggest reporting the probability that the estimate falls within a decision-relevant range or exceeds (falls below) a pertinent threshold. Such probabilistic statements—supported by the Bayesian model the framework uses but not by frequentist inference—communicate results in a useful yet honest format to readers.

Quantifying measurement uncertainty. The natural comparison is the posterior standard deviation against the average single-measure standard error. However, this understates the role of measurement uncertainty, because the model also *gains* information relative to a single measure which partly offsets the widening. The cleaner isolation compares the hierarchical (random-effects) model against a fixed-effects model built from the same prior and the same likelihood but assuming measures differ only by chance.

Be explicit about what the framework does not do. The model will only be as accurate as the data that go into it allow. If measure-specific biases approximately offset, the posterior median will sit near the truth. If biases are directionally aligned, or if the fielded measures are a non-random subset of plausible measures sharing a common bias, the posterior inherits this bias. This limitation is shared by any aggregation procedure. What the framework therefore delivers is honest uncertainty quantification *given the available evidence*. We contend that this is still useful even if biases do not offset. As long as measures are not aligned in direction *and* magnitude of bias, the model will deliver an estimate that is biased but with a wide confidence range. This compares to a single measure’s estimate which can be just as biased but is so *confidently*, that is, with misleadingly narrow uncertainty.

F.8 Beyond descriptive survey settings

Measurement uncertainty presents a relevant concern beyond descriptive estimates and survey data. Our framework applies to these settings, wherever several plausible approaches exist and each produces an estimate for the same target quantity.

Experiments. Measurement uncertainty affects experimental work whenever treatment interacts differentially with the way the outcome is measured, so that the estimated treatment effect depends on the measurement choice. Given how sensitive respondent behavior is to small design

variations, we contend that the reasonable default is to assume such interactions exist and to falsify this by testing rather than by assumption.

Discretionary analytical choices. The same logic applies where the discretion is analytical rather than instrumental. For example, data analyzed independently by separate research teams (Menkveld et al., 2024), or more generally the “humans-in-the-loop” problem in which researcher choices materially move results but go unaccounted for (Stevenson and Fischman, 2026). Where competing approaches are each defensible under their own assumptions, treating the disagreement as uncertainty to be quantified is preferable to the “dueling certitudes” that result from treating each as certain (Manski, 2011).

Other data sources. Nothing in the framework is specific to survey self-reports. It applies wherever measurement products are substitutable but disagree. For instance, the choice between remotely sensed measures of weather, climate extremes, or crop yields is one such case.

F.9 Frequently asked questions

Below, we present short answers to frequent doubts that came up when we presented this work.

Is this just a robustness check under a different name? No, and the distinction lies in how robustness is quantified. A conventional robustness check reports whether a result survives an alternative choice, leaving the reader to adjust their confidence informally. Here, the disagreement between plausible measures is an input to estimation. It enters the posterior and changes the interval you report. The heuristic that researchers already apply, trusting a result more when several approaches agree, is exactly what the model makes explicit and quantitative.

Does this license an approach where pooling rescues poor measurement? The opposite. Because poor measurement choices now carry a quantified inferential cost, the framework forces measurement decisions to the fore rather than leaving them implicit. Pooling does not turn poor measurement into good measurement, and we do not claim it does.

Am I being punished for being thorough? Readers want tight inference and clear conclusions. Whether quantified or not, the uncertainty exists either way. The only question is whether the reported interval acknowledges it. Further, consider that the direction is not automatic. Because the model incorporates several measures and a prior, it can produce a *tighter* posterior than a single measure’s confidence interval when plausible measures agree. Posteriors will be wide only in those cases where single-measure inference would have misled the most. Wide posteriors tell you and readers that the measurement problem is unsolved and create a visible return to better measurement. Conventional inference does neither.

Is this affordable? Survey space is scarce, and fielding multiple measures competes with everything else in the instrument. The demands here are comparable to those of other established

measurement-error remedies: [Gillen et al. \(2019\)](#) require duplicate elicitations of the same measure, and [de Quidt et al. \(2018\)](#) suggest repeating tasks with an induced demand treatment. One or two additional measures may already deliver much of the benefit.

I can only afford one extra measure. Is it worth it? Yes. The move from one measure to two is what pulls measurement uncertainty into view in the first place. With a single measure, measurement uncertainty is not merely unaccounted for but unidentified. Be aware that a two-measure posterior is noisy and depends on which pair you fielded.

My measures disagree enormously. Has something gone wrong? Not necessarily—that is the finding. Large disagreement means the estimate is highly sensitive to a choice that could reasonably have gone another way. What you should not do is resolve the disagreement by picking the measure you find most favorable after seeing the results. This would represent the specification search pre-registration is meant to prevent.

Why not correct the measurement error instead? Where the structure of the measurement error can be credibly identified, correcting it remains the first choice. For instance, the obviously related instrumental variables (ORIV) estimator of [Gillen et al. \(2019\)](#) recovers the underlying parameter from duplicate elicitations when error is classical. That route is closed for many quantities economists study. The errors we document are systematic rather than classical, specific to each method, and capable of moving in the same direction. In these cases, averaging or instrumenting across measures need not converge on the truth and easily conveys a sense of ill-advised confidence.

I have no prior information. What should I do? Use a diffuse location prior. The prior is a channel for genuine extraneous knowledge, not a requirement to invent one. A relatively flat prior lets the data determine the posterior. If you want an evidence-based prior but no literature exists, eliciting predictions is inexpensive and yielded a usable, if diffuse, prior in our case.

Will a larger sample fix this? No. Sampling uncertainty falls with N ; measurement uncertainty does not. It falls only when plausible measures agree more closely. This is why investment in validated, best-practice measurement is a public good for inference: each validated tool shrinks the set of plausible alternatives, and therefore the measurement uncertainty every subsequent study faces.