Forecasting in the Field

Nicholas G. Otis*
U.C. Berkeley

Abstract

Who knows what about the impacts of large policy interventions? This paper uses 20,000 forecasts of 50 causal effects from three large experiments in Kenya made by academics, people similar to intervention recipients, and nonexperts to examine belief accuracy. Average predicted effects track experimental results well: average absolute error from the mean academic forecast is only 0.11 standard deviations, and the average correlation between predicted and observed effects is 0.71. Recipient types are less accurate than academics on average but are at least as accurate for interventions and outcomes that are likely to be more familiar to them. The mean forecast of each group outperforms more than 75% of the comprising individuals, and averaging just five forecasts substantially reduces error, indicating strong "wisdom-of-crowds" effects. Three measures of academic expertise (rank, citations, and conducting research in East Africa) and two measures of confidence do not correlate with accuracy. Among recipient-types, high-accuracy "superforecasters" can be identified using observables. Small groups of these superforecasters are as accurate as academic respondents. I conclude by providing applications to intervention choice, information maximization, and hypothesis testing.

*notis@berkeley.edu. I gratefully acknowledge financial support from the Weiss Fund for Development Economics, the National Institute of Aging (T32-AG000246), the Russell Sage Foundation, GiveWell, and the Global Priorities Institute. I would like to thank Claire Boone, William Dow, Emily Eisner, Dan Honig, Dean Karlan, Supreet Kaur, Don Moore, Josh Rosenberg, Catherine Thomas, Eva Vivalt, and especially Stefano DellaVigna, as well as seminar participants at the U.C. Berkeley, IDinsight, Oxford, and the CEGA summer development series for their helpful comments and suggestions. I would also like to thank the teams of the participating studies: Stefan Dercon, Dennis Egger, Rob Garlick, Tilman Graff, Johannes Haushofer, Anett John, Mahreen Mahmud, Robert Mudida, Edward Miguel, Paul Niehaus, Kate Orkin, Richard Sedlmayr, Jeremy Shapiro, and Michael Walker. Christina Chao, Chazel Hakim, Benjamin Kern, Zimai Lan, Xiaoguang Liang, Angela Li, Sérgio Nascimento, and Erika Page provided excellent research assistance for this project. I would also like to thank the study participants for their time and effort, and the Busara Center for Behavioral Economics for data collection, especially Chaning Jang, Irene Ngina, and Pauline Wanjeri for their excellent project management. This project was approved by the U.C. Berkeley Committee for the Protection of Human Subjects.

Forecasts of experimental results can provide important insights into the selection of policies, the design of experiments, and the production of knowledge (DellaVigna et al., 2019). To policymakers, forecast accuracy can signal whose recommendations should be given more weight. To researchers, accurate forecasts can inform which policies should be evaluated. In the case of a null result, expert forecasts can highlight why a finding is interesting, potentially mitigating publication bias. Finally, forecasts of experimental effects can quantify how much new information a study produces, since many results seem obvious ex post.

I collected forecasts of the results of three large, randomized field experiments in Kenya ($n = 5,500$ to $10,500$) before their results were public: Haushofer et al. (2020) examine the effects of a mental health intervention benchmarked against a cash transfer. Orkin et al. (2020) evaluate the effects of an aspirations and goal-setting intervention, also benchmarked against cash. Egger et al. (2020) evaluate the general equilibrium effects of a large cash transfer program. Together these studies include 15 experimental conditions, and 18 outcomes (e.g., consumption, mental health, intimate partner violence).[1] Over 20,000 forecasts were collected from 1,309 respondents: 134 academics (*academic types*), 612 Kenyan respondents similar to the intervention recipients (*recipient types*), and 563 respondents from an accessible online nonexpert pool (*nonexpert types*).[2]

I document four main results. **First**, the average forecast tracks the experimental results remarkably well across a range of accuracy measures. For academics, 66% of the mean forecasts are not significantly different from the observed effect at the 0.05 level, even though the three studies are well powered, and the mean forecast is precisely estimated. The absolute error on the mean forecast is only 0.11 standard deviations (SD), while the average correlation (measured at the outcome level and averaged across outcomes) with observed effects is 0.71.

While academics provide the most accurate predictions averaging across all conditions and outcomes, there is substantial heterogeneity. The mean recipient-type forecast is more accurate than the mean academic forecast for 36% of predictions. I identify two key dimensions of heterogeneity by type: recipient types are approximately as or more accurate than academics for behavioral (as opposed to subjective) outcomes, and for cash transfer and spillover conditions (as opposed to conditions involving the mental health or aspirations and goal-setting intervention). Both are dimensions where recipient types are likely more familiar with the predicted effect, allowing them to leverage their local knowledge. For example, across the twelve predicted household general equilibrium effects from cash transfers on household consumption and SD, the average recipient-type forecast deviates from the

[1] Predictions were collected for 55 outcomes: 50 at the individual/household level and five at the market level, which examine general equilibrium effects of cash transfers on prices. Forecast error is orders of magnitude higher for these five market price outcomes. For example, the average absolute error on the mean academic forecast is 12.1 times larger for the five market price outcomes compared to the 50 individual/household level outcomes. I provide a discussion of these outcomes in Appendix A but focus on the household outcomes in the paper body. Note that these outcomes would naturally be excluded from correlational and rank-based accuracy measures, since only one effect is forecast for each price outcome.

[2] 15,680 forecasts remain after applying pre-registered exclusion criteria.

observed effect by only 0.07 SD, compared to 0.12 SD among the academic sample and 0.19 SD for the nonexperts.

Next, I compare individual predictions to the mean predictions from groups of forecasters. While error from the average forecast is low, there is substantial belief heterogeneity for each effect within each group of forecasters. The **second** result is that pooling independent forecasts substantially improves prediction accuracy–a phenomenon known as the "wisdom-of-crowds" (Surowiecki, 2005). The mean academic forecast is more accurate than 75% of individual academics in terms of absolute error, while the mean recipient-type forecast is more accurate than 86% of the comprising individuals' forecasts. I provide evidence from bootstrapped simulations that "crowds" of just ten academics have 31% lower error than the comprising individuals. At the individual level, academic types are still more accurate than recipient- and non-expert types. While the mean forecast of recipient- and nonexpert-types produce similar levels of error, at the individual level, recipient-types are substantially more accurate than nonexperts, with an average absolute error of 0.28 SD compared to 0.40 SD.

Third, I examine the correlates of individual forecast accuracy. Among the academic sample, I explore measures of *vertical expertise*: academic rank (e.g., being an assistant vs associate professor) and citations. More senior academics are if anything less accurate, and number of citations is uncorrelated with accuracy. As a measure of *contextual expertise*, I examine whether the respondent has conducted research in East Africa, which is not significantly associated with accuracy. Next, I present evidence from Likert-scale and quantitative measures of confidence and calibration. Using the Likert-scale measure, more confident individuals are less accurate, and there is strong evidence of overprecision: academic experts think 41% of their predictions will fall within 10% of the true effect, when in fact only 3% do.

Turning to the recipient-type sample, I find that use of a physical aid where participants make predictions using a printed number line leads to small but robust improvements in accuracy across a range of accuracy measures, though recipient types are still less accurate than the academic types. Among recipient types, the strongest predictor of accuracy is which survey enumerator administered the survey. This could result from variation in enumerator quality, or from enumerators "projecting" their own predictions on their respondents. I recontacted the five enumerators who administered the survey, collecting their incentivized predictions of the same set of experimental effects, finding that their forecasts correlate more strongly with the predictions of the respondents they interviewed than with other enumerators' respondents. Additionally, I find that individual enumerators are approximately as accurate as academic respondents.[3]

Fourth, I examine whether high accuracy "superforecasters" can be identified using observables and measures of confidence (Tetlock and Gardner, 2015). Superforecaster identification relies on the strength of accuracy correlates described above. Using a simple pre-registered k-fold OLS procedure to avoid over-fitting, I can identify high-accuracy fore-

[3]The only enumerator who received a perfect score on a screening quiz used during the hiring process is more accurate than the mean prediction of the entire academic sample.

casters among the recipient types, though not among the academic and nonexpert types with weaker accuracy correlates. The average absolute error among the top 20% of recipient types is 32% lower than the full recipient-type sample, and the "crowd" prediction of five superforecasters is approximately as accurate as the average academic respondent.

Next, I conduct simple back-of-the-envelope calculations to connect these descriptive results about forecast accuracy with my motivating applications of forecasting to intervention choice, hypothesis testing, and generating new information. **First**, I calculate the effect of crowd size on correctly ranking a subset of experimental conditions involving comparable interventions. A crowd of ten academics is 18 percentage points more likely to "choose" the more effective treatment than a single academic. **Second**, I examine whether conducting hypothesis tests against the average academic forecast meaningfully changes the interpretation of the results compared to the status-quo null hypothesis of no effect. Of the 30 experimental effects that were not significant compared to the standard null hypothesis of no effect, six (20%) become significant when compared to the mean academic prior. **Finally**, I consider the question of how identify experiments with high informational value (defined as the absolute difference between the average prediction and the observed effect). I find that forecast variance (a measure of disagreement about what the effects will be) correlates strongly (0.78) with the amount of information produced. Put another way, mean forecasts are more accurate for effects where there was less disagreement among individuals.

This paper contributes to an emerging literature on forecasts of social science results, which has three main strands. Several papers have explored the accuracy of forecasts of experimental replications, or the stability of experimental results (Camerer et al., 2016, 2018; DellaVigna and Pope, 2018b). There is also a rich literature on forecasts of geopolitical events (Tetlock and Gardner, 2015; Tetlock, 2017). More broadly, this research follows a long tradition in economics of exploring beliefs about future states. Manski (2004) provides a seminal review of this literature, and Delavande et al. (2011); Delavande (2014) review this literature in a developing country context. The literature most relevant to this study examines beliefs about the causal effects of interventions in developing countries (DellaVigna et al., 2020; Thomas et al., 2020; Abebe et al., 2019; Casey et al., 2019; Groh et al., 2016; Bloom et al., 2018; Andrade et al., 2014).

A related literature has explored decision-making among policymakers in development (Vivalt and Coville, 2020; Hjort et al., 2019) and other fields (Ambuehl et al., 2019). For example, Hjort et al. (2019) show that policymakers in Brazil demand information on intervention effectiveness, update on this information, and incorporate it into policy decisions. In environments where there is limited evidence of the effects of potential policies, this study examines forecasts as a new source of information on what works.

Outside of development, DellaVigna and Pope (2018a) conducted one of the first large-scale studies exploring forecasts of experimental results in economics. They collected forecasts of results of 15 experimental treatments in a 10,000-subject experiment aimed at motivating effort on the online platform Amazon Mechanical Turk (MTurk) from 208 academic experts, as well as university students, and MTurkers. The correlation between academic

forecasts and experimental treatment effects was 0.77. In a companion paper, DellaVigna and Pope (2018c) show that average forecasts were more accurate than estimated treatment effects from a systematic meta-analysis. My results replicate many of the key findings of DellaVigna and Pope (2018c) using evidence from three large field experiments in Kenya, as opposed to a large online real-effort experiment in the U.S.

Finally, I contribute to a literature on generalizability and replicability in economics (Bates and Glennerster, 2017; Vivalt, 2019; Christensen and Miguel, 2018; Meager, 2019). Recent empirical work has highlighted concerns about the scalability, generalizability, and replicability of intervention effects; as evidence-based policymaking is occurring on a large scale in the Global South, it is important that research and policy decisions are made using the best available information. The remaining sections are as follows. Section 1 describes the study design. Section 2 presents results, and Section 3 concludes.

1 Study Design

1.1 Forecast Studies and Outcomes

Study Selection. Three criteria were used to select the predicted studies. First, each study needed to be well-powered to reduce the role of sampling error in the predicted studies influencing forecast accuracy. The sample sizes for the selected studies range from n=5,500 to 10,500. Second, the studies needed to be pre-registered with a pre-analysis plan. This allowed outcomes to be selected before any of the experimental results were known. Third, each study had to have multiple treatments (to test correlational measures of accuracy), and multiple outcomes (e.g., health, education, assets) to test the breadth of forecaster ability. Overviews of the three studies are provided below, and more details can be found in the pre-analysis plans for the studies.

Study 1 Overview. Egger et al. (2020) randomly vary the village-level saturation of unconditional cash transfers worth approximately $1,000 (nominal) within clusters (called sublocations) of two to 15 villages in Siaya County, Kenya, allowing them to identify general equilibrium effects. In high-intensity sublocations, poor (eligible) households in two-thirds of the randomly assigned villages received cash, while in low-intensity sublocations, eligible households in only one third of the villages received cash. For this study, I collected forecasts of two household outcomes: (1) annual household consumption, and (2) total household assets. Based on the experimental design, there are eight different types of households: eligible and ineligible households, in high or low-intensity sublocations, in treated or untreated villages. I provided forecasters with average outcomes for eligible and ineligible households living in low-intensity sublocations in villages not receiving cash, and collect forecasts for the remaining six groups.

Study 2 Overview. Haushofer et al. (2020) randomly assign households in Nakuru County, Kenya to receive either an unconditional cash transfer of approximately $500 (nominal), a mental health intervention developed by the World Health Organization called Problem

Management Plus (PMP), or both the cash and the PMP intervention. Randomization took place in two stages: first, villages were randomized into one of four conditions (pure control, cash, PMP, or PMP and cash), after which those individuals not in pure control villages were randomly assigned to their villages' respective treatment, or a control (spillover) condition. The PMP intervention was delivered over a period of five weeks by community health workers, and emphasized problem solving, managing stress and problems, behavioral activation, and strengthening social support. Forecasts were collected for four outcomes: (1) monthly household consumption expenditure, (2) mental health measured through the General Health Questionnaire (GHQ), (3) subjective well-being (SWB), and (4) the proportion of women reporting physical intimate partner violence (IPV) from their male partners. Forecasters were provided with the average outcome of individuals in villages where nobody received an intervention as a reference.[4]

Study 3 Overview. Orkin et al. (2020) randomize villages in Homa Bay and Siaya Counties in Kenya to one of four treatments. Eligibility for all treatments was determined using criteria correlated with per-capita consumption. One group of villages received a role-model and goal-setting intervention consisting of viewing two ten-minute videos in which role models similar to the audience overcame obstacles and set and achieved goals related to long-term aspirations. After viewing the videos, participants took part in an hour-long facilitated drawing and discussion exercise, and received a calendar depicting the role models and stickers (to represent goals), which they were encouraged to put on the calendar. A second group received a placebo intervention–they watched a video, completed facilitated exercises, and received a calendar and stickers, but these were missing the role modeling and aspirational components. All eligible households in a third group of villages were assigned to receive a cash transfer of about $1,100 nominal, and the placebo intervention. A fourth group received both the cash intervention and the role-model intervention. Forecasts were collected for six outcomes: (1) household assets, (2) educational expenditure, (3) monthly household consumption expenditure, and aspirations for (4) child education, (5) total non-land assets, and (6) monthly income. Forecasters were provided with the average outcomes for eligible placebo-control households as a reference.

1.2 Forecaster Samples and Data Collection

I collected forecasts of experimental treatment effects from three groups: academic, recipient, and nonexpert types. All participants completed the survey over Qualtrics. The structure of the survey was: (1) background and IRB information, (2) description of the randomized controlled trial, (3) comprehension questions, (4) forecast elicitation, and (5) confidence and calibration questions. Comprehension questions were designed to test whether respondents had a basic understanding of the predicted experiment. To make forecasts incentive compatible, participants are informed that some individuals will be selected at random to receive a bonus based on the accuracy of their predictions. For all participants, I pre-registered that

[4]Due to a survey error, forecasters were provided with PPP as opposed to nominal reference levels of consumption for the control group for the monthly household consumption expenditure outcome. Calibration exercises can show that an unrealistic MPCs would be required for this to have a meaningful impact on overall accuracy.

participants failing any comprehension questions would be excluded from the main analysis.

Academic Types. Potential academic respondents were sent an invitation to participate in the forecasting study over email. These invited respondents were drawn from three pools. The first pool was comprised of 150 authors of research on cash transfers (since 2013), holding PhDs in economics or a related field (e.g., agricultural and resource economics or public policy). Authors were identified through Google Scholar searches for terms such as "cash transfer" and "unconditional cash". These authors were assigned to forecast the results of one of the three studies. For each of the studies, I also contacted 50 PhD-holding authors of research (also since 2013, and identified using Google Scholar) from the substantive area of each of the three studies (mental health, general equilibrium effects, or goal-setting and aspirations), and which predominantly focused on developing countries. The final email sample included 150 PhD students in economics or related fields, with a stated focus on development economics. Fifty students were assigned be invited to forecast the results of each study.

Each potential respondent received a tailored email inviting them to participate in the Qualtrics forecasting survey, which mentioned their research (PhD students received a more generic email). In this email, respondents could opt in to receive feedback on the accuracy of their predictions. If no response was received after two weeks, a reminder email was sent. [5] A total of 523 invitations were sent, and I received a total of 138 responses, for a response rate of 138/(523-64)=0.30 after excluding 64 respondents who were replaced, for example due to invalid email addresses or a Qualtrics survey error. Following prespecification, four respondents reported having heard about specific study results from the then unpublished studies (for example, through discussions with the project teams) and eight respondents who failed at least one comprehension question were excluded from the primary analysis, though results are robust to their inclusion. The remaining 126 respondents made a total of 2,092 forecasts.

Table A1 lists summary statistics comparing the invited academics to those who responded. Respondents resemble the invited population in terms of academic rank. For example, 25% of invited respondents were assistant or associate professors, compared to 26% of the responding sample. The median responding PhD-holding academic had been cited 1,119 times, compared to the median invited academic's 1,468 citations. Respondents who had conducted research in East Africa comprised only 40% of invited respondents, but made up 51% of respondents.

Recipient Types. The second group of forecasters are Kenyans from socioeconomic backgrounds similar to those of intervention recipients. Two groups of about 300 respondents from Nairobi and Kirinyaga County were recruited by the Kenyan research organization the Busara Center for Behavioral Economics (Busara). The Nairobi sample was more accessible, while the Kirinyaga sample was geographically and socioeconomically more similar to the

[5] Some individuals received an additional reminder if they posted an away message, for example saying that they will return in a week. If I received an automatic response that the individual would be away from email for more than approximately two weeks, a replacement was contacted, but the initial individual was still allowed to respond.

rural areas where the predicted studies took place. To increase comprehension, all respondents were required to be under age 40 and have completed primary school. The Nairobi sample came from a large survey pool maintained by Busara. Participants from this pool were recruited by phone, and interviews took place in a rented office space. The second group of surveys took place in Kirinyaga county, a location selected to avoid possible contamination with the studies being forecast, and based on feasibility, as Busara had worked there before. In Kirinyaga, community mobilizers worked with Busara enumerators to locate survey respondents, and interviews were completed in a rented community hall.

Surveys were conducted by five trained enumerators using tablet computers as well as visual aids to help convey the experimental design. I also randomized whether respondents provided verbal forecasts, or used physical aids, placing houses along a number line depicting conditional means of the different experimental groups. Respondents received 500 Kenyan Shillings (KES; about $5) for participating in the survey, inclusive of a transport fee to the survey location, plus a 50 KES bonus for arriving on time.

In addition to the screening questions used to test a basic understanding of the experimental design, enumerators rated respondent comprehension. Those rated as having "understood very little of the survey" or "understood some of the survey, but struggled with many parts" are excluded from the primary analysis, following pre-specification (the remaining options categorized respondents as having understood the survey "well" or "perfectly"). Results are robust to inclusion of all participants. Of 612 respondents, 441 passed both the survey comprehension questions and the enumerator-rated comprehension test, providing 7,380 forecasts.

Nonexpert Types. Nonexperts are from the online platform Amazon Mechanical Turk (MTurk). MTurk workers complete short online paid tasks, and are increasingly used in social science research (Paolacci and Chandler, 2014). For each study, I posted a survey on Amazon Mechanical Turk titled "Short survey on social policies" which ran until I had received about 200 responses for each of the three studies. Respondents were paid $1.25 for participating in the survey. Through the platform, I restricted myself to workers in the U.S. who have an approval rating above 95%, and who had completed more than 50 tasks. Of the 563 respondents, 384 passed the comprehension questions, providing 6,208 forecasts.

1.3 Data Preparation

Pre-Registration and Pre-Analysis Plan. This study is registered on the AEA's Social Science Registry under AEARCTR-0003600. For each of the three predicted studies, I uploaded a document pre-specifying which outcomes would be predicted before experimental results were known. Additionally, I uploaded a pre-analysis plan in which I specify statistical analyses, exclusion criteria, and robustness checks. Non-pre-registered analyses are labelled as such in the table or figure's *notes.*

Forecast Preparation. Forecasters are provided with the conditional mean of a reference condition, and then predict the conditional means of the other conditions. For each types'

predictions, I winsorize the top 5% of forecasts by magnitude (while maintaining the original sign) at the outcome level to further screen careless forecasts, for example from people who may have misinterpreted the questions, added extra zeros, or were otherwise inattentive (Vivalt, 2017; McKenzie, 2018; Rees-Jones and Taubinsky, 2019).

Accuracy Measures. Results focus on two complementary pre-registered accuracy measures. *Negative absolute forecast error* measures the (negative) absolute difference between the forecast and observed experimental effect:

$$\delta^{NAE} = -|y^{SD} - \tilde{y}^{SD}| \ , \tag{1}$$

where y^{SD} is the observed experimental effect in SD, and $\tilde{y}^{SD}$ is the predicted experimental effect (measured using the same SD). Predicted experimental effects are calculated by subtracting reference means presented in the forecasting surveys from the forecaster's predicted conditional mean for each experimental group. Experimental treatment effects were provided by the project teams. Note that $\tilde{y}^{SD}$ can refer to an individual forecast ($\tilde{y}_i^{SD}$), or can refer to the average forecast of a group ($\bar{\tilde{y}}_n^{SD} = \frac{1}{n}\sum_{j=1}^{N} \tilde{y}_i^{SD}$), where n is the group size. This will be used when examining the effect of aggregation on accuracy ("wisdom-of-crowds" effects). Negative absolute forecast error provides a useful metric for quantifying the magnitude of prediction errors. A second accuracy measure examines the *correlation* between forecast and observed effects at the outcome level:

$$\delta_k^{cor} = \frac{\text{cov}(\boldsymbol{y}_k, \tilde{\boldsymbol{y}}_k)}{\sigma_{\boldsymbol{y}_k}\sigma_{\tilde{\boldsymbol{y}}_k}} \ , \tag{2}$$

where $\text{cov}(\boldsymbol{y}_k, \tilde{\boldsymbol{y}}_k)$ is the covariance between $\boldsymbol{y}_{k,}$ a vector of observed effects for outcome k, and $\tilde{\boldsymbol{y}}_{k,}$ the forecast effect equivalent.[6] σ denotes the respective standard deviations. This measure complements negative absolute forecast error, by capturing the extent to which predicted and observed effects for different treatments "move together" for a given outcome.[7] We are again able to calculate accuracy either at the individual level by taking the correlation between the observed experimental effects for an outcome and the individuals' vector of forecast effects, or for a group of size n, by substituting the vector of *average* forecast effects. Results for additional pre-registered accuracy measures are presented in Table A3. Results are generally robust to alternative pre-registered robustness checks (e.g., winsorizing the top 1% of forecasts or dropping comprehension-based exclusion criteria).

[6]Note that the *SD* superscript has been dropped since location- and scale-invariance properties of the correlation coefficient imply that standardization will not impact accuracy.

[7]The fact that average forecasts perform well on both of these accuracy measures indicates that people are both distinguishing between different conditions for a given outcome, and that their predictions of specific effects are close to the observed experimental result.

2 Results

2.1 Average and Individual Accuracy

Average Predictions. How well do average forecasts predict the experimental treatment effects? In Figure 1, the 50 experimental effects are depicted by vertical black dashes, surrounded by gray 95% confidence intervals. Blue and red diamonds denote the average forecast for each treatment effect and the accompanying 95% confidence interval. Forecast effects that are not significantly different from the experimental effect at the $p < 0.05$ significance level (as measured through a z-test) are colored blue, while red forecasts are significantly different. The average forecast is precisely estimated and the predicted studies are well powered, making the z-test a reasonable summary measure of prediction accuracy. Among academics, 36 of 50 forecasts are not significantly distinguishable from the observed experimental effect, and the average absolute difference from the true effect is only 0.11 SD. The average forecast correlates strongly with the observed experimental effect. Taking the correlation within each outcome and averaging across all outcomes yields a mean correlation of 0.71.

Heterogeneity. The average predictions of recipient and non-expert-types are less accurate than the academic sample, though this is not the case for all outcomes. For example, recipient types are more accurate than academic types for 18 of the 50 outcomes (for more details, see Table A2). Recipient types generally perform best on outcomes and conditions that they are likely to be familiar with. It is reasonable to assume that recipient types will be more familiar with behavioral outcomes such as consumption and assets, compared to the subjective measures (e.g., Likert scale measures of happiness, or measures of aspirations). Since cash transfer programs are widespread in Kenya, it also seems reasonable to assume that recipient types will be more familiar with these types of interventions than non-cash interventions like mental-health counseling, which are relatively uncommon. Stratifying results along these dimensions, we can see that the recipient types' average error for the mean forecast among the non-cash and subjective outcomes is 0.37 SD, compared to 0.07 for the cash and behavioral outcomes, and for which recipient types outperform academics whose mean error is 0.10 SD (see Figure 2 for details).

Individual Predictions. How does the accuracy of *individual* forecasts compare to the mean prediction by each type? In Table 1, I show that averaging independent predictions leads to substantial accuracy improvements. Among academic types, the average absolute error of individual academic forecasts is 0.17 SD, compared to 0.11 SD for the mean "crowd" prediction. Examining individual forecast accuracy further differentiates recipient types from nonexperts: both groups performed similarly in terms of average prediction accuracy, but the average absolute forecast error for recipient types is 0.28 SD compared to 0.40 SD for nonexperts. Pre-registered regression results comparing group differences across a range of accuracy measures and robustness checks are depicted in Table A3.

Benchmarking. To benchmark individual accuracy, I compare experimental forecasts to random predictions drawn from the uniform distribution within 1.50 SD around a mean of 0,

and within 0.75 SD of a mean of 0.10 SD. Individual forecasts from academic and recipient types are more accurate than these benchmark forecasts in terms of negative absolute error. All three types are more accurate than these random draws in terms of correlation-based measures of accuracy, since the benchmark results in no correlation between observed and predicted effects.

2.2 Wisdom-of-Crowds

Table 1 also highlights that the mean "crowd" forecast is far more accurate than the predictions of the individuals comprising the crowd. The crowd prediction among academics is more accurate (in terms of absolute error) than 75% of the comprising individuals' predictions, and the average negative absolute error is 35% lower for the crowd compared to the individual predictions (similar results hold for correlational accuracy measures). However, there are some effects (e.g., academic predictions of the combined effect of therapy and cash on consumption) for which more than half of the individuals are more accurate than the group prediction (see Table A2).

How many forecasts are required to produce the wisdom-of-crowds? Figure 3 depicts cumulative distribution functions (c.d.f.'s) comparing the negative absolute error of individual forecasts (in blue) to the average predictions of 5,000 bootstrapped "crowds" of size 5 (red) and 10 (yellow). These simulations show that averaging just five predictions results in average absolute error (denoted by the circular points on the c.d.f.'s) close to the full sample accuracy (the black dashed line). Figure 3 also highlights that within each group there are individuals that are more accurate than the crowd prediction. This motivates two questions. First, what are the correlates of forecast accuracy (subsection 2.3)? Second, can these features be used to identify high-accuracy type forecasters (subsection 2.4)?

2.3 Determinants of Forecast Accuracy

Academic Types. Among the academic sample, I examine the correlation between expertise and forecast accuracy. My results are consistent with DellaVigna and Pope (2018a), who find that traditional measures of expertise do not correlate strongly with forecast accuracy. Table 2 presents two measures of *vertical* expertise (academic rank and citations) and one measure of *contextual* (or *horizontal*) expertise (whether the respondent has conducted research in East Africa).

Panel A of Table 2 displays the association between academic rank and negative absolute forecast error. Full and associate professors are less accurate than assistant professors–though not significantly so. Academic rank is a fairly coarse measure of vertical expertise. As alternative measures of an academics' influence, I consider their citations, as measured on Google Scholar (or Research Gate, if the respondent could not be found on Google Scholar). Excluding PhD students, the median academic respondent in my sample had been cited over 1,100 times. Figure 4 depicts the association between log citations (winsorized at the top 5%) and negative absolute forecast error. Panel B of Table 2 confirms the visual trend in Figure 4: that there is no clear association between forecast accuracy and citations. Finally,

in Panel C we use data on whether each (PhD-holding) respondent has conducted research in East Africa, again finding a negligible association with forecast accuracy. Together, this evidence suggests that academic expertise is not a strong predictor of having accurate beliefs about the effects of interventions, at least for the three measures employed in this study.

Do forecasters have private information about the accuracy of their predictions? I explore this question by eliciting two measures of confidence in one's forecasts. Panel A of Figure 5 depicts the correlation between a Likert scale measure of confidence and forecast accuracy, which asks *How confident are you in your predictions for this study? If you are confident it means that you believe your predictions are very accurate*, with response options *not at all*, *not very*, *somewhat*, and *very confident*. Individuals reporting higher confidence in their predictions are if anything less accurate. For the second measure, I ask participants to predict the proportion of forecasts that fall within 10% of the (then unknown) true experimental effect. Calibrated forecasters would fall on the 45-degree line depicted in Panel B. Participants predicted that 43% of their forecasts will fall within 10% of the true effect, while only 3% of forecasts actually fell in these bounds. Panel C shows that the measure of confidence used in Panel B is basically uncorrelated with absolute forecast error.

Recipient and Nonexpert Types. Among the academic sample, measures of expertise and confidence are not strong predictors of forecast accuracy. Are there strong determinants of forecast accuracy among the recipient and nonexpert types? Table 3 depicts correlates of forecast accuracy among these samples. Consistent with a literature on measuring subjective expectations in a developing country context (Delavande, 2014), I find that use of a (randomly assigned) physical aid in eliciting forecasts significantly improves accuracy, though the effect is small. Respondents in this condition used small cut-outs of houses to order treatments along a number line, as opposed to providing verbal forecasts to enumerators, which reduced forecast error by 0.04 SD. I also randomly assigned half of the recipient and nonexpert types to receive an additional sentence emphasizing the importance of accuracy by stating the minimum and maximum amount they could receive based on the accuracy of their predictions. For example, the MTurk respondents received the message: *...someone who is very accurate could get as much as $100, while someone who is very inaccurate could earn $0*. Similar to DellaVigna and Pope (2018a), I find that accuracy-incentive salience has a small and insignificant effect on forecast accuracy.

While the association between education, income, and accuracy is weak, individuals who grew up in more rural areas (outside Nairobi) from the Nairobi pool had more accurate beliefs than individuals who were born in Nairobi and drawn from the Nairobi pool. This provides suggestive evidence that *some* exposure to contexts similar to where the interventions took place (more rural areas, outside Nairobi) may result in more accurate forecasts, though this could also be explained by selection.

Enumerator Effects. We also see strong enumerator effects on forecast accuracy: the largest difference in accuracy between enumerators' participants is 0.10 SD. One interpretation of this result is that the enumerators "project" their own beliefs onto respondents in the forecast elicitation process. To test this idea, I recontacted the five enumerators and elicited

their predictions of the experimental results. Unlike participants who predicted the results of only one study, each enumerator provided forecasts for all three studies, resulting in a total of 275 forecasts. The first result of this exercise is that enumerators' own beliefs correlate on average more strongly with those of the respondents who they administered surveys to (cor=0.89) compared to those of other enumerators (cor=0.69). This result is consistent with enumerators projecting their beliefs onto their respondents, though it is also consistent with enumerators updating their beliefs from their respondents' predictions, or with better enumerators more effectively administering the survey and also having more accurate beliefs. This strong correlation between enumerator and respondent accuracy raises the question of how accurate the enumerators themselves are.

At the individual level, the average negative absolute error of the enumerators (0.18 SD) is approximately the same as for academics (0.17 SD; see Figure A1 the respective c.d.f.s). I leverage a screening test used when hiring the enumerators, which measured nuanced comprehension of the survey instruments. The one enumerator that received a perfect score on this test provided forecasts that were more accurate than the full sample of academics. This evidence suggests that enumerators may be a valuable source of information on the effectiveness of interventions, perhaps because they combine contextual and hands-on academic knowledge, and that screening tools may be an effective way to identify those with more accurate beliefs.

2.4 Identifying "Superforecasters"

Next we examine whether high-accuracy individuals can be identified based on observables. This builds on a substantial body of research in psychology examining forecasting tournaments focused on predicting geopolitical events, where accuracy correlates strongly within individuals (Tetlock and Gardner, 2015), as well as a smaller body of work examining identification of superforecasters with respect to predictions of experimental results (DellaVigna and Pope, 2018a).

Method. My strategy to identify superforecasters follows the k-fold procedure outlined in DellaVigna and Pope (2018a). For each experiment and type, I first partition the sample into 10 folds. I omit the first fold, and regress a measure of average individual absolute error on observables and confidence measures, which I use to generate fitted values for the omitted part of the sample. I then iterate this procedure until I have a fitted value for the respondents in each fold. After this, I calculate the top 20% of respondents by fitted values ("superforecasters"). I repeat this procedure 1,000 times. To calculate the average prediction accuracy from a crowd of superforecasters, for each pool of superforecasters, I take 1,000 bootstrapped samples of sizes $n = 1$ and 5. For each crowd, I calculate the negative absolute error from the mean forecast for each treatment, and then average this error across the crowds' full set of predictions.

Academic and Nonexpert Types. Note that if there is a zero correlation between accuracy and observables, we would observe a *negative* correlation between fitted values (OLS-predicted accuracy) and observed accuracy because of regression to the mean. For example,

if a high accuracy respondent is in an omitted fold, then their fitted value will be lower because the constant from an OLS regression (the predicted accuracy from the remaining sample) omits their high accuracy. This mean regression is decreasing in sample size and the strength of accuracy correlates. This is what we observe among the academic and nonexpert types, as shown in Figure 6.

Recipient Types. For the recipient-type sample, which have reasonably strong accuracy correlates and a much larger sample than the academic respondents, high-accuracy individuals can be successfully identified. For this group, the c.d.f. of the "superforecasters" is shifted to the right of the general sample forecasts, meaning that the superforecaster "crowd" is more accurate. We can observe that a crowd of five recipient-type superforecasters has a mean prediction approximately equal to the average individual academic. At the individual level, the full sample of recipient-type respondents has an average absolute error approximately 25% higher than the superforecasters.

2.5 Applications

How do these results connect to the three motivations described in the introduction: (1) selecting interventions, (2) hypothesis testing against forecasts (priors), and (3) generating new information from experimentation?

First, consider the question of *what is the right number of people to inform intervention selection*? Consider two discrete policy choices: the selection of either the cash transfer or the mental health intervention from Haushofer et al. (2020), or the cash transfer or goal-setting intervention from Orkin et al. (2020). For 1,000 bootstrapped samples at each size $n = 1$ to 20, I calculate the proportion of crowds whose mean forecast correctly ranks treatments on each of the ten outcomes from these two studies. For example, in Orkin et al. (2020) the cash transfer intervention was more effective at increasing expenditure on child education than the aspirations intervention (the respective treatment effects are 0.08 and 0.01 SD). Panel A of Figure 7 depicts the proportion of crowds who rank each treatment effect correctly by crowd size. We observe large effects of crowd size on ranking: the mean prediction of ten forecasters is 18 percentage points more likely to rank treatments correctly than a single forecaster.

Panel B reweights the results from Panel A based on the difference in treatment effects between the two interventions. Returning to the example above, the difference in treatment effects is $0.08 - 0.01 = 0.07$ SD. A crowd that "selects" the better treatment here is said to produce a benefit of 0.07 SD. Averaging across all outcomes, the errors in ranking from a crowd of ten versus a single forecaster corresponds to a reduction in treatment effects of about 20%. While these calculations substantially oversimplify the decision environment (e.g., standardized outcomes are weighted equally), they suggest the power of aggregating independent beliefs as a decision aid.

Second, do forecasts meaningfully change how we interpret results? As an example, consider the effect of the PMP counseling intervention on depression, which is -0.05 SD

(se=0.06), and thus not significant. When compared to the expert predictions (a 0.12 SD effect), the result becomes highly significant. Turning to the full set of 50 effects, 30/50 were not significant at the $p < 0.05$ level. If we use the average expert prior as an alternative null hypothesis, 20% of these effects become significant. This simple calculation highlights that using expert priors as a complement to the traditional null hypothesis of no effect can lead to meaningful changes in how we interpret results.

Finally, consider an alternative motivation for experimentation: trying to find which interventions will provide the most new information. Here, we want to maximize the absolute distance between the observed effect and the average academic prior, which provides an approximation of the new information provided by a study. Is there a way to determine which interventions will provide the most new information ex ante? Figure 8 shows that forecast variance (within group) predicts accuracy, implying that we learn the most from studies where experts disagree ex ante. The correlation between forecast variance and accuracy is 0.78. If we interpret error from the mean forecast as "accuracy" as opposed to "information" (as we have done in the rest of the paper), this result suggests that mean forecasts are less accurate for effects where there was more ex ante disagreement. Note that this strong correlation is not mechanical, since higher variance in predictions does not imply that the *mean* prediction will be less accurate.

3 Conclusion

This study presents evidence from over 20,000 forecasts of 50 pre-registered experimental treatment effects from three experiments in Kenya, made by academics, people more similar to intervention recipients, and nonexperts. Forecasts of these large field experiments confirm several key results from an emerging literature on forecasting experimental results (DellaVigna and Pope, 2018a).

First, the average academic forecast predicts the observed experimental effects quite well. While academics are the most accurate overall, the recipient-type sample is at least as accurate for more familiar (behavioral) outcomes and (cash-transfer) interventions. **Second**, there are strong "wisdom-of-crowds" effects: individual forecasts are much less accurate than a group's mean prediction. Averaging just five forecasts leads to substantial improvements in accuracy. **Third**, I examine the correlates of individual forecast accuracy, showing that three measures of expertise among the academic sample do not correlate with forecast accuracy. If anything, more senior or more confident academics tend to have less accurate beliefs about experimental effects. Among recipient types, survey enumerator is the strongest predictor of forecast accuracy. Enumerator forecasts correlate with their respondents' forecasts, and enumerators are themselves quite accurate. **Finally**, I can identify high-accuracy "superforecasters" (at least among the recipient-type sample), and crowds of five of these superforecasters are about as accurate as academic respondents.

These back-of-the-envelope calculations in the Applications section highlight some of the ways that forecasts might be used to improve how policies are selected, how results are

interpreted, and how we learn from experimentation. One concern is whether these results apply to other environments. This concern is partly assuaged by the fact that this study looks at multiple outcome domains (e.g., health, consumption, aspirations), effect types (e.g., main effects from single treatments, additive effects of combined treatments, and spillover conditions), and collects predictions from large and heterogeneous groups of forecasters. Further research will be required to see whether the findings presented here are replicated in other contexts or among different groups of forecasters.

References

Abebe, G., Fafchamps, M., Koelle, M., and Quinn, S. (2019). *Learning Management Through Matching: A Field Experiment Using Mechanism Design*. Working paper.

Ambuehl, S., Bernheim, B. D., and Ockenfels, A. (2019). Projective paternalism. Technical report, National Bureau of Economic Research.

Andrade, G. H. D., Bruhn, M., and McKenzie, D. (2014). A helping hand or the long arm of the law? experimental evidence on what governments can do to formalize firms. *The World Bank Economic Review*, 30(1):24–54.

Bates, M. A. and Glennerster, R. (2017). The generalizability puzzle. *Stanford Social Innovation Review*, 15(3).

Bloom, N., Mahajan, A., McKenzie, D., and Roberts, J. (2018). *Do management interventions last? evidence from India*. The World Bank.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637.

Casey, K., Glennerster, R., Miguel, E., and Voors, M. (2019). Skill versus voice in local development. Technical report, National Bureau of Economic Research.

Christensen, G. and Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80.

Delavande, A. (2014). Probabilistic expectations in developing countries. *Annu. Rev. Econ.*, 6(1):1–20.

Delavande, A., Giné, X., and McKenzie, D. (2011). Measuring subjective expectations in developing countries: A critical review and new evidence. *Journal of development economics*, 94(2):151–163.

DellaVigna, S., Otis, N., and Vivalt, E. (2020). Forecasting the results of experiments: Piloting an elicitation strategy. In *AEA Papers and Proceedings*, volume 110, pages 75–79.

DellaVigna, S. and Pope, D. (2018a). Predicting experimental results: who knows what? *Journal of Political Economy*, 126(6):2410–2456.

DellaVigna, S. and Pope, D. (2018b). *Stability of Experimental Results: Forecasts and Evidence*. Working paper.

DellaVigna, S. and Pope, D. (2018c). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

DellaVigna, S., Pope, D., and Vivalt, E. (2019). Predict science to improve science. *Science*, 366(6464):428–429.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. (2020). General equilibrium effects of cashtransfers: experimental evidencefrom kenya.

Groh, M., Krishnan, N., McKenzie, D., and Vishwanath, T. (2016). The impact of soft skills training on female youth employment: evidence from a randomized experiment in jordan. *IZA Journal of Labor & Development*, 5(1):9.

Haushofer, J., Mudida, R., and Shapiro, J. (2020). *The Comparative Impact of Cash Transfers and Psychotherapy on Psychological and Economic Well-being*. Working paper.

Hjort, J., Moreira, D., Rao, G., and Santini, J. F. (2019). How research affects policy: Experimental evidence from 2,150 brazilian municipalities. Technical report, National Bureau of Economic Research.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, 72(5):1329–1376.

McKenzie, D. (2018). Can business owners form accurate counterfactuals? eliciting treatment and control beliefs about their outcomes in the alternative treatment status. *Journal of Business & Economic Statistics*, 36(4):714–722.

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.

Orkin, K., Garlick, R., Mahmud, M., Sedlmayr, R., Haushofer, J., and Dercon, S. (2020). *Aspirations, Assets, and Anti-Poverty Policies*. Working paper.

Paolacci, G. and Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188.

Rees-Jones, A. and Taubinsky, D. (2019). Measuring schmeduling. *The Review of Economic Studies*.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Tetlock, P. E. (2017). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

Tetlock, P. E. and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction.* Random house.

Thomas, C. C., Otis, N. G., Abraham, J. R., Markus, H. R., and Walton, G. M. (2020). Toward a science of delivering aid with dignity: Experimental evidence and local forecasts from kenya. *Proceedings of the National Academy of Sciences*, 117(27):15546–15553.

Vivalt, E. (2017). *How Much Can Impact Evaluations Inform Policy Decisions?* Working Paper. Stanford, CA: Stanford University.

Vivalt, E. (2019). *How much can we generalize from impact evaluations?* Working paper.

Vivalt, E. and Coville, A. (2020). How do policymakers update their beliefs?

Table 1: Forecast Benchmark

Type	Negative absolute error in standard deviations			Pearson correlation		
	Average accuracy of individual forecasts (sd) (1)	Accuracy of mean forecast (wisdom-of-crowds) (2)	% of forecasters more accurate than crowd (3)	Average accuracy of individual forecasts (sd) (4)	Accuracy of mean forecast (wisdom-of-crowds) (5)	% of forecasters more accurate than crowd (6)
Academic	-0.17 (0.20)	-0.11	24.44	0.59 (0.45)	0.71	21.81
Recipient	-0.28 (0.30)	-0.19	15.53	0.43 (0.42)	0.55	28.23
Nonexpert	-0.40 (0.49)	-0.20	18.72	0.31 (0.55)	0.62	15.91
Benchmark for comparison						
Random guess in -1.5 to 1.5	-0.76			0.00		
Random guess in -0.65 to 0.85	-0.38			0.00		

Notes: This table reports the negative absolute error (cols. 1 to 3) and correlation between predicted and observed experimental effects (cols. 4 to 6) for each respondent type (rows 1 to 3), and compared a benchmark (rows 4 to 5). Benchmarks are based on 1,000,000 draws from the uniform distribution from -1.5 to 1.5, or from -0.65 to 0.85 (the mean prediction of the academic group was approximately 0.10). Cols. 1 and 4 present the mean individual negative forecast error and correlation with their respective standard deviations. Cols. 2 and 5 present the negative absolute forecast error and correlation for the mean (crowd) forecast. Cols. 3 and 6 display the percent of forecasters whose average prediction error (across all of their predictions) is lower than the average group forecast (forecasters who made constant predictions receive a correlation of 0). Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level.

Table 2: The Effect of Academic Expertise on Accuracy

	Negative absolute error (SD)
Panel A	
Ref: Assistant prof	
PhD student	-0.003 (0.020)
Researcher or postdoc	-0.008 (0.024)
Associate professor	-0.015 (0.025)
Full professor	-0.025 (0.023)
n_i=123, n_f=2056	
Panel B	
log(cites)	0.001 (0.004)
n_i=81, n_f=1360	
Panel C	
Ref: No research in East Africa	
Research in East Africa	0.009 (0.017)
n_i=81, n_f=1360	

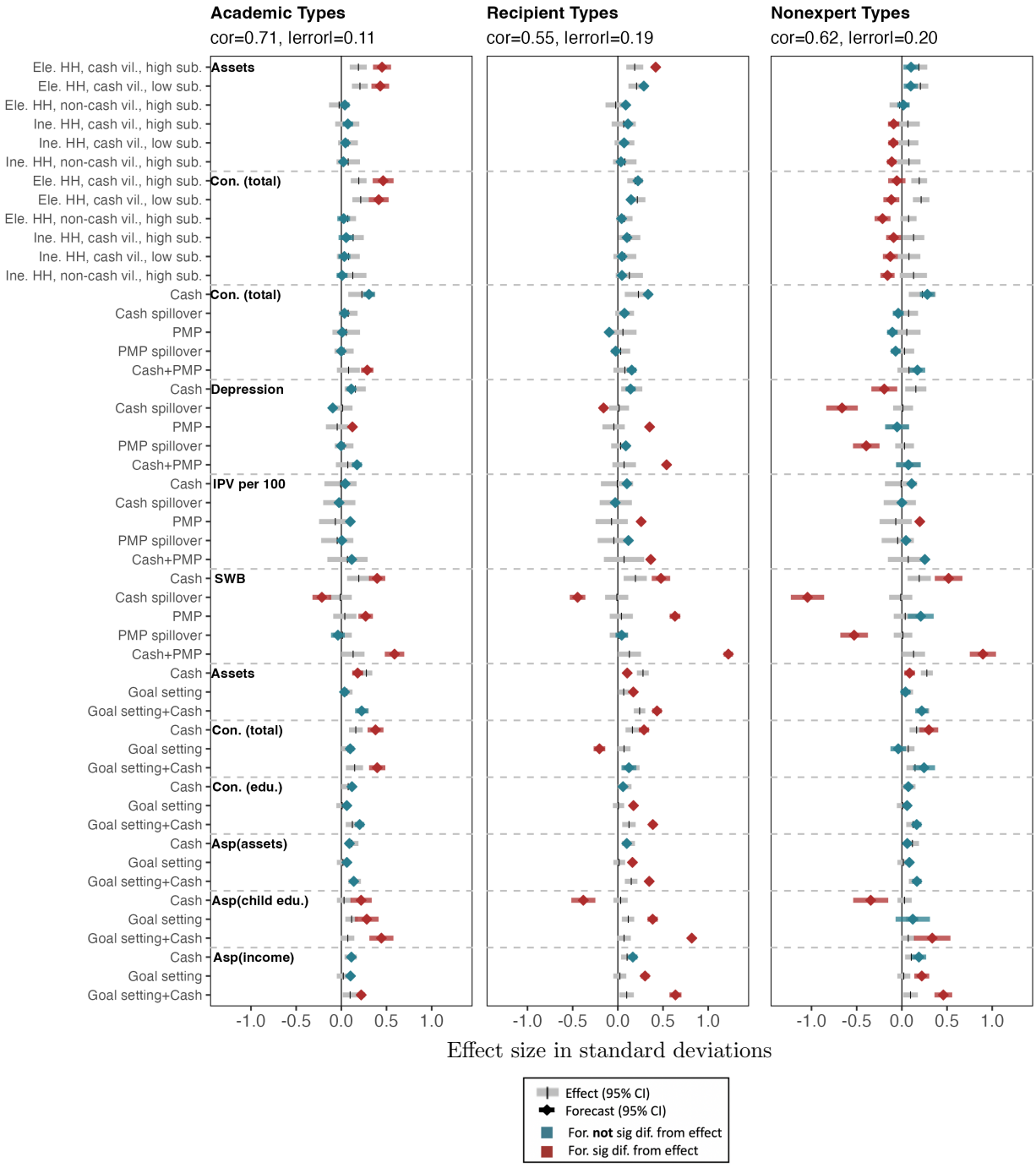
Notes * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level. Column 1 presents negative absolute forecast error, with standard errors clustered at the individual level displayed in parentheses. n_i refers to the number of individual forecasters, and n_f refers to the total number of forecasts. All models include condition×outcome fixed effects. Each panel represents a separate OLS regression. Panels B and C exclude PhD students. Observations are at the individual forecast×condition×outcome level. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level.

Table 3: Determinants of Accuracy: Recipient- and Nonexpert-Types

	Negative absolute forecast error (SD)		
	(1)	(2)	(3)
Panel A: Recipient Types			
Ref: No Physical Aid			
Physical aid	0.038*** (0.009)	0.037*** (0.009)	0.038*** (0.008)
Ref: No Salient Incentives			
Salient Incentives	0.002 (0.009)	0.001 (0.009)	-0.001 (0.008)
Ref: Nairobi Sample, From Nairobi			
Kirinyaga sample, From Nairobi		0.027 (0.029)	0.032 (0.025)
Kirinyaga sample, Not From Nairobi		0.005 (0.014)	0.003 (0.013)
Nairobi sample, Not From Nairobi		0.032** (0.014)	0.024* (0.014)
Ref: Secondary School or Less			
More Than Secondary		-0.004 (0.010)	-0.006 (0.009)
Ref: Above Median Income			
Below Median Income		0.015 (0.009)	0.018** (0.009)
Ref: Enumerator 1			
Enumerator 2			0.041*** (0.016)
Enumerator 3			0.017 (0.016)
Enumerator 4			0.101*** (0.015)
Enumerator 5			0.068*** (0.015)
n_i=441, n_f=7380			
Panel B: Nonexpert Types			
Ref: No Salient Incentives			
Salient Incentives	0.010 (0.025)	0.010 (0.025)	
Ref: Less Than college			
Completed College (or above)		0.034 (0.027)	
Ref: Below $30,000			
Above 30,000		-0.018 (0.026)	
n_i=384, n_f=6208			

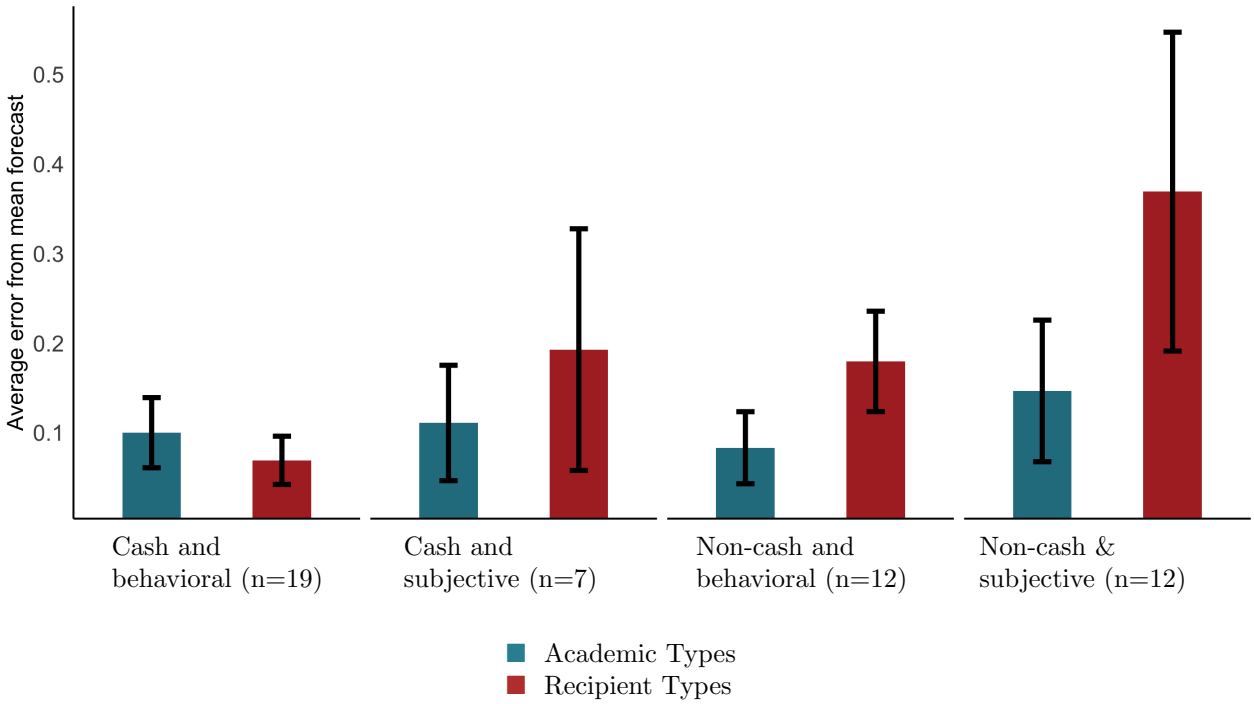
Notes: * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level. Panels A and B display results from regressing negative absolute forecast error on sociodemographic variables. Standard errors clustered at the individual level are displayed in parentheses. n_i refers to the number of individual forecasters, and n_f refers to the total number of forecasts. All models include condition×outcome fixed effects. Observations are at the individual forecast×condition×outcome level. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level. The enumerator analysis in Panel A was not pre-registered.

Figure 1: Average Forecast and Experimental Effects



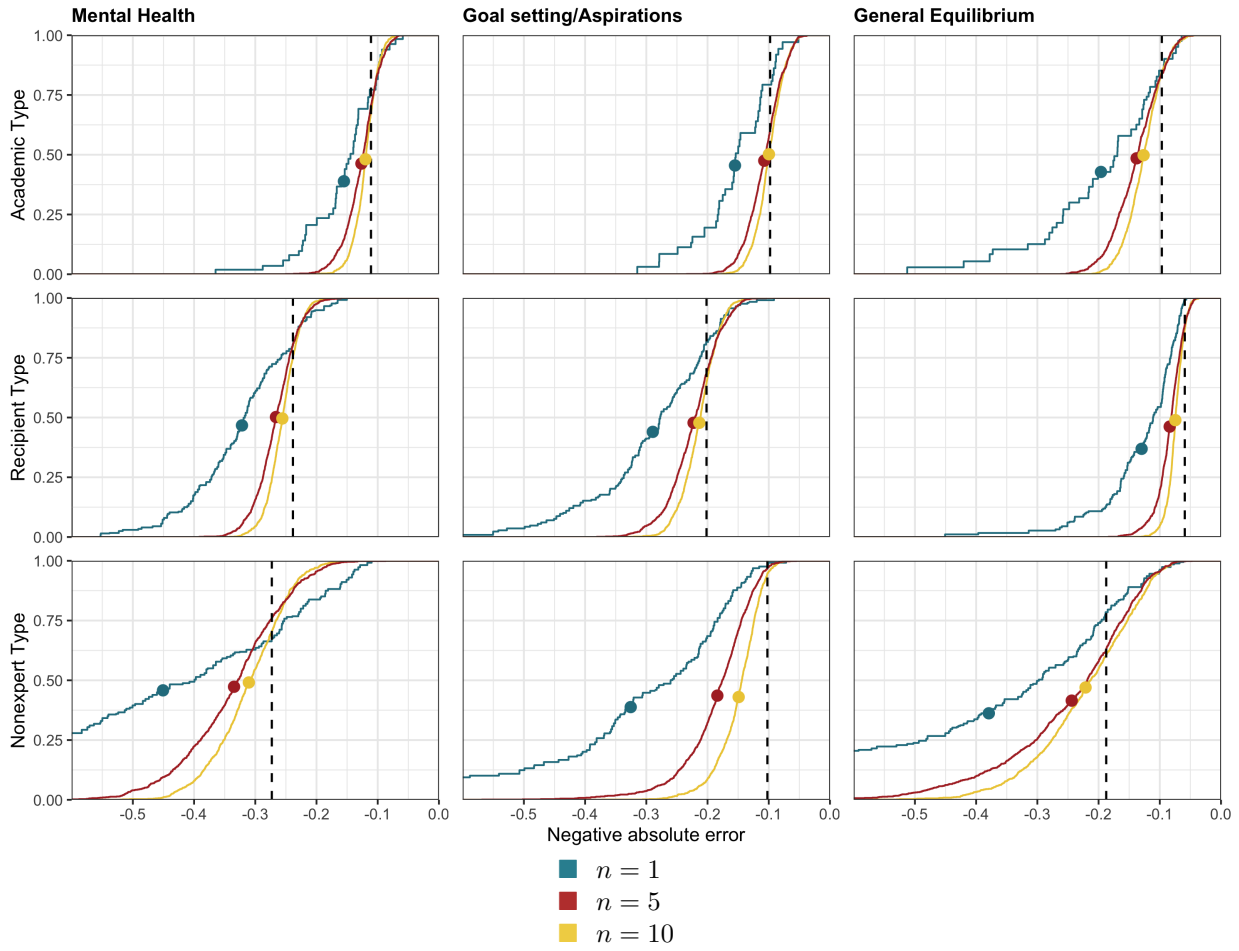
Notes: Grey bars represent 95% confidence intervals around the experimental effect. Blue or red bars represent 95% confidence intervals around the average forecast experimental effect. Red bars denote significant ($p<0.05$) differences between the forecast and experimental effect, as measured through a z-test, and blue bars denote differences that are not significant. Average correlation ("cor") is measured by first taking the correlation between the average forecast and experimental effects at the outcome level, and then averaging across outcomes. Average absolute error ("|error|") is calculated by taking the absolute difference between the average forecast and experimental effect at the effect level, and then averaging across all effects. Individual-level forecasts used to calculate the mean prediction are winsorized at the 5% level by magnitude at the type×outcome level.

Figure 2: Average Forecast Accuracy by Outcome and Treatment type



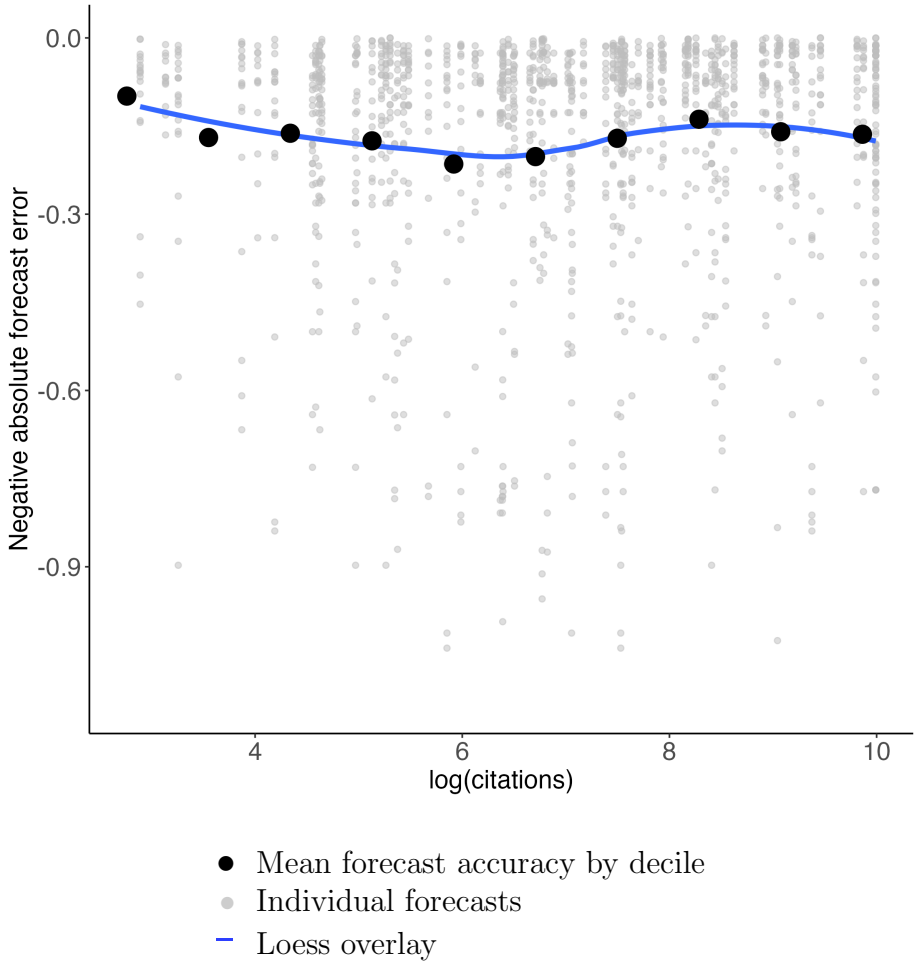
Notes: Bars depict average absolute forecast error from the mean forecasts by outcome type (subjective outcomes or behavioral outcomes) and effect type (cash transfer and spillover conditions, or conditions including a noncash intervention and associated spillover conditions). Subjective outcomes are: subjective well-being, depression and aspirations about assets, childhood education, and income. Behavioral outcomes are: assets, consumption, and intimate partner violence. Non-cash interventions include the mental health intervention, the spillover of the mental health intervention, and the combined cash transfer and mental health intervention from Haushofer et al. (2020), and the aspirations and goal-setting, and the combined cash, aspirations and goal setting intervention from Orkin et al. (2020). The cash transfer interventions include the pure cash transfer and spillover conditions from the previous two studies, and the full set of cash transfer and spillover conditions from Egger et al. (2020). Error bars depict a 95% confidence interval around the mean. Observations are at the effect level. Individual-level forecasts used to calculate the mean prediction are winsorized at the 5% level by magnitude at the type×outcome level. This analysis was not pre-registered.

Figure 3: Cumulative Distribution Functions of Negative Absolute Error by Experiment and Type



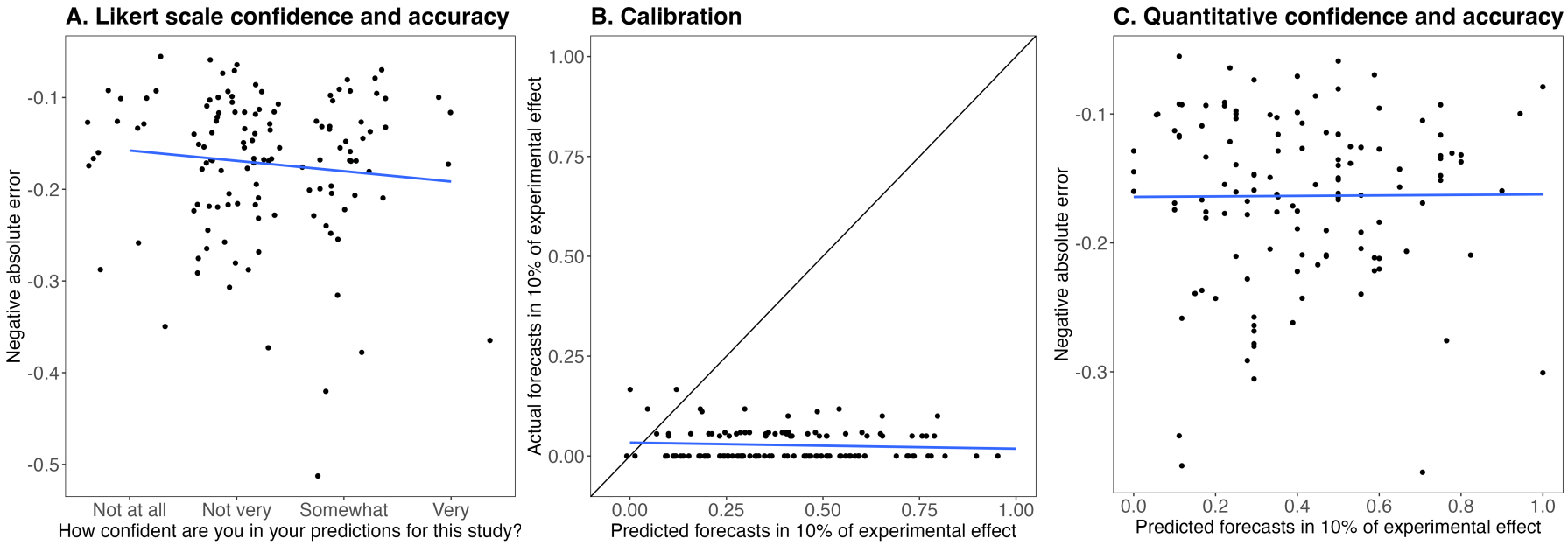
Notes: Negative absolute error for a crowd of size $n = \{1, 5, 10\}$ is calculated by taking a bootstrapped sample of size n, and then calculating the average negative absolute error of the groups' mean prediction for each treatment effect (observations are at the crowd-size×experiment level). This procedure is repeated 5,000 times to generate c.d.f.'s of error for each crowd size. Dotted lines denote the average negative absolute error for the full sample. Points on each c.d.f denote the average negative absolute error across all groups for a given crowd size. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level.

Figure 4: Citations and Forecast Accuracy



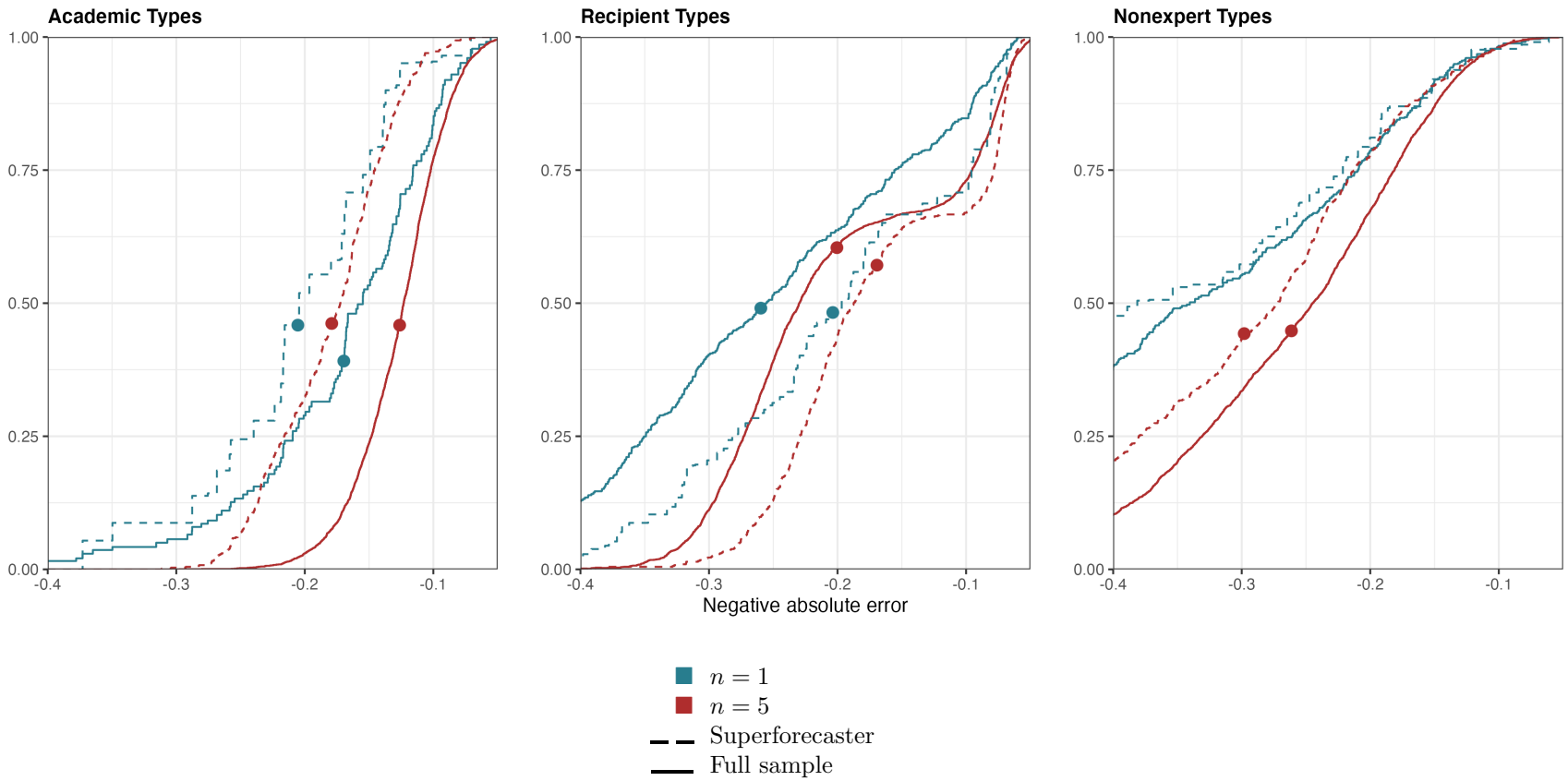
Notes: Logged citations (winsorized at the 5% level) from the PhD-holding academic respondents are presented on the x-axis. The y-axis displays negative absolute forecast error, where forecast and experimental effects are measured in standard deviations. Observations are at the individual forecast×condition×outcome level. Individual-level forecasts are winsorized at the 5% level by type×outcome.

Figure 5: Forecast Accuracy, Confidence, and Calibration



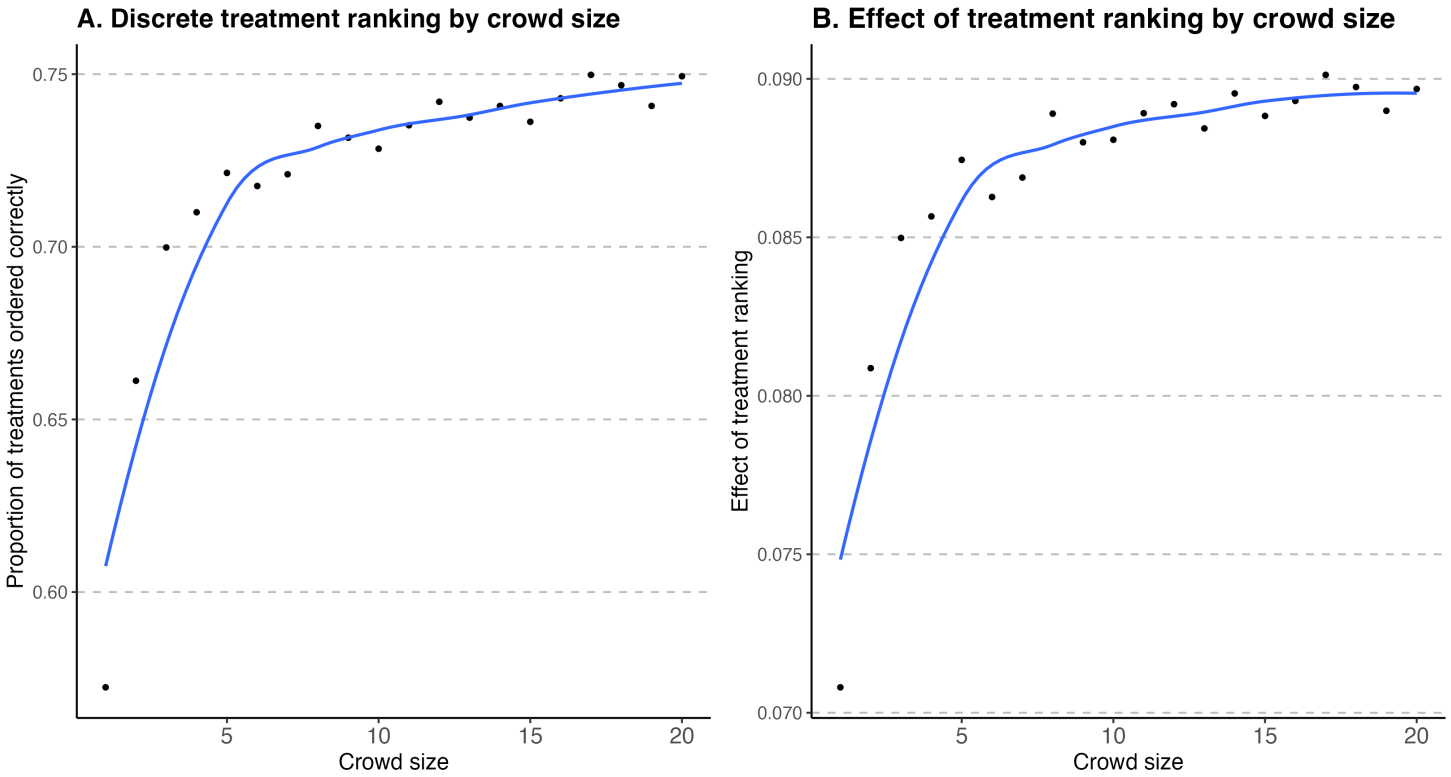
Notes: Panel A presents negative absolute forecast error by self-reported confidence. The x-axis depicts responses to the question: "How confident are you in your predictions for this study? If you are confident it means that you believe your predictions are very accurate." Data are presented with a horizontal jitter. The y-axis displays respondents' average negative absolute error. Panel B presents the predicted vs observed proportion of forecasts within 10% of the experimental effect. The 45-degree line represents perfect calibration. In Panel C, the x-axis is the same as in Panel B, and the y-axis is negative absolute error. The blue lines depict the fit of a linear model. Observations are at the individual level. For panels A and C, negative absolute error is calculated for each of the individual's forecasts, and then averaged within the individual. Panels B and C include the 5 market price outcomes, since these would have been included in the set of predictions considered by respondents when making their estimate. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level.

Figure 6: Cumulative Distribution Functions of Superforecaster Negative Absolute Error



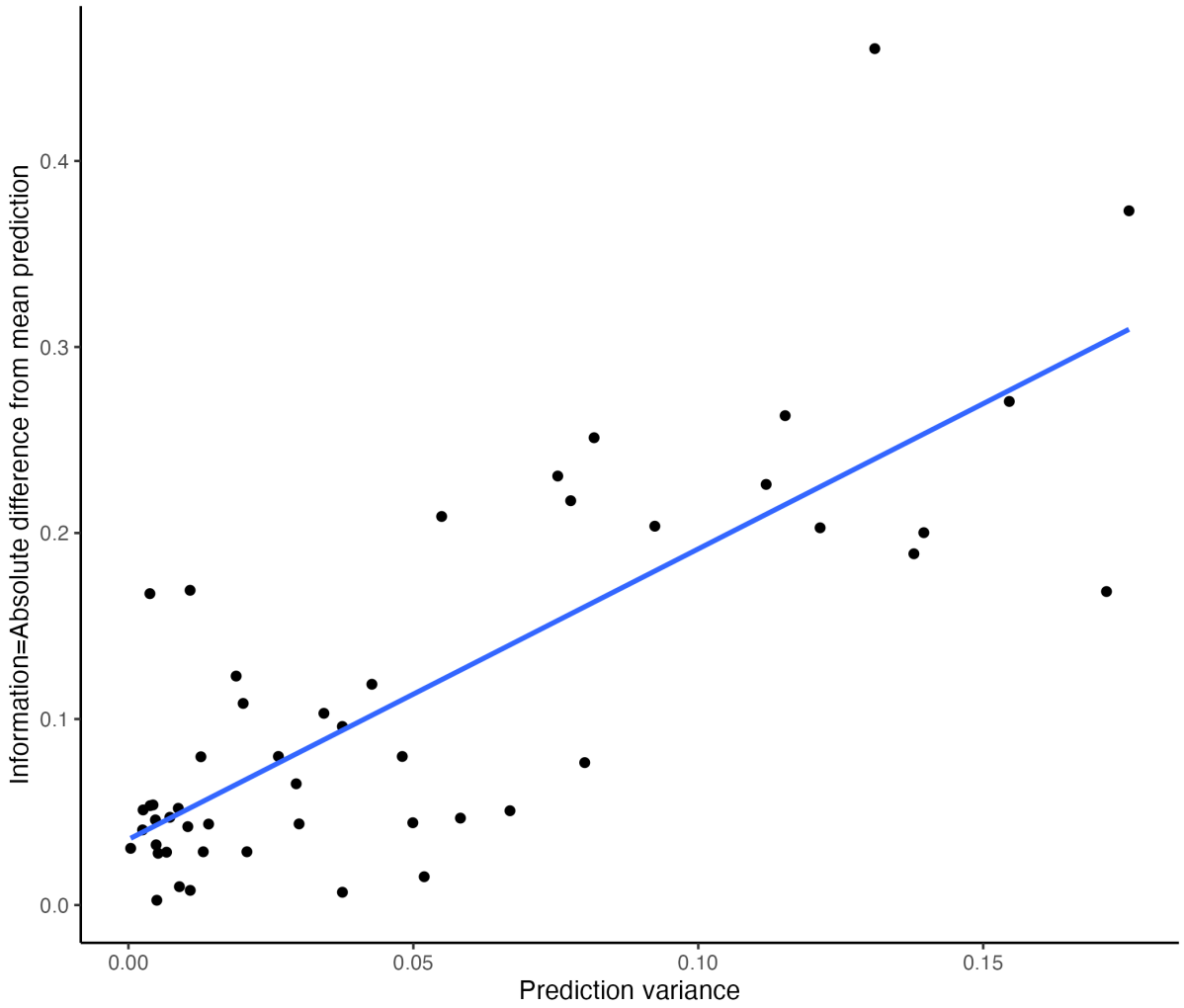
Notes: Negative absolute error for a crowd of size $n = \{1, 5\}$ is calculated by taking a bootstrapped sample of size n, and then calculating the average negative error of the groups' mean prediction for each treatment effect (observations are at the crowd-size $\times$experiment level). This procedure is repeated 1,000 times to generate c.d.f.'s of error for each crowd size. Dotted lines denote crowds drawn from the top 20% of forecasters based on a k-fold procedure. Points on each c.d.f. denote the average negative absolute error across all groups for a given crowd size. Individual-level forecasts are winsorized by magnitude at the 5% level by type$\times$outcome.

Figure 7: Simulated Wisdom-of-Crowds Policy Choice



Notes: In Panel A, the y-axis presents the proportion of crowds from 5,000 bootstrapped simulations that correctly rank two treatment effects over a set of 10 outcomes. For Haushofer et al. (2020), the interventions are (1) the cash transfer and (2) the mental health intervention, and the outcomes are (1) monthly household consumption expenditure, (2) mental health measured through the General Health Questionnaire (GHQ), (3) subjective well-being, and (4) the proportion of women reporting physical intimate partner violence (IPV) from their male partners. For Orkin et al. (2020), the interventions are (1) the cash transfer and (2) the aspirations and goal setting intervention, and the outcomes are (1) household assets, (2) educational expenditure (3) monthly household consumption expenditure, and aspirations for (4) child education, (5) total non-land assets, and (6) monthly income. In Panel B, the y-axis is reweighted by the difference in (standardized) treatment effects between the two treatments. In both panels, the x-axis displays the crowd size. Individual-level forecasts are winsorized at the 5% level by magnitude at the type ×outcome level. This analysis was not pre-registered.

Figure 8: Information and Forecast Variance



Notes: The x-axis presents the variance in academic forecasts for an experimental effect using individual-level forecast data. The y-axis presents the absolute error on the mean forecast for each effect. Observations are at the effect level. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level. This analysis was not pre-registered.

A Appendix: Market price forecasts

Forecasts were collected for five market price outcomes from Egger et al. (2020). In this study, market price data was collected on a monthly basis from 61 local markets. Forecasts were elicited for market prices in high-saturation sublocations (where poor households in 2/3 of villages received cash transfers), with low-saturation markets (where poor households in 1/2 of villages received cash transfers) serving as a reference. Forecasts of high-intensity sublocation prices were elicited for five different goods: (1) two kg of maize; (2) one kg of rice; (3) one kg of beef; (4) one local calf; (5) one 32-gauge 10-foot iron sheet.

Figure A2 depicts the average absolute forecast error from the mean prediction for market price and non-market price outcomes by type. Black points depict the average absolute forecast error for each effect. Averaging across outcomes, the absolute error of the mean forecast for the five market-price outcomes for the academic, recipient, and non-expert types are 12.1, 8.01, and 22.0 times higher respectively than the 50 non-market outcomes.

All groups overestimate the effect of these price outcomes, meaning average error is the same as average absolute error. The fact that all groups predict large price increases from a higher saturation of cash transfers suggests that all types of respondents understand the basic economic intuition underlying price effects, even though evidence from Egger et al. (2020) suggests that these effects are empirically small. The prediction of large price effects, combined with the (relatively) small standard deviation on these market-level outcomes results in the high forecast error.

Appendix Tables

Table A1: Response Summary

	Invited Academics (1)	Responding Academics (2)	Recipient Types (3)	Nonexpert Types (4)
Panel A: Academic Types				
Academic rank				
Assistant professor	0.11	0.11		
Associate professor	0.14	0.15		
Professor	0.22	0.20		
PhD student	0.33	0.34		
Researcher or postdoc	0.19	0.20		
Research in East Africa				
No	0.60	0.49		
Yes	0.40	0.51		
Median citations	1468	1119		
Pred. prop. in 10% of effect		0.41		
Confidence				
Not at all / Not very		0.64		
Somewhat / Very		0.36		
Panel B: Recipient Types				
Physical aid				
No			0.48	
Yes			0.52	
Salient incentives				
No			0.51	
Yes			0.49	
Sample / location				
Nairobi Sample, from Nairobi			0.04	
Kirinyaga sample, from Nairobi			0.47	
Kirinyaga sample, not from Nairobi			0.20	
Nairobi sample, not from Nairobi			0.30	
Education				
Started/completed secondary			0.35	
Started/Completed university			0.65	
Income				
Below median inc.			0.56	
Above median inc.			0.44	
Enumerator				
Enumerator 1			0.14	
Enumerator 2			0.19	
Enumerator 3			0.24	
Enumerator 4			0.18	
Enumerator 5			0.25	
Panel C: Nonexpert Types				
Salient incentives				
No				0.52
Yes				0.48
Education				
Completed college and above				0.65
Some college and below				0.35
Income				
Above 30,000				0.63
Below 30,000				0.37
n_i	523	126	441	384
n_f		2092	7380	6208

Notes: Col. 1 provides information on the full sample of invited academics. Col. 2 provides information on the responding academics. The rows "median citations" and "research in East Africa" exclude PhD students. Cols. 3 and 4 provide information on the recipient and nonexpert types, respectively. Cols. 2-4 include only those individuals who pass pre-registered screening criteria. n_i is the number of individuals and n_f is number of forecasts.

Table A2: Summary of Accuracy by Condition

				Mean forecast effect (SD)			Neg. abs. error (mean for.)			Neg. abs. error (ind. for.)			% more accurate than crowd		
	Behavior (1)	Effect (SD) (2)	(se) (3)	Aca. (4)	Rec. (5)	MTu. (6)	Aca. (7)	Rec. (8)	MTu. (9)	Aca. (10)	Rec. (11)	MTu. (12)	Aca. (13)	Rec. (14)	MTu. (15)
General Equilibrium															
Assets															
Ele. HH, non-cash vil., high sub.	Yes	-0.03	0.06	0.04	0.09	0.02	0.06	0.11	0.04	0.08	0.13	0.30	0.75	0.55	0.28
Ele. HH, cash vil., low sub.	Yes	0.20	0.04	0.43	0.29	0.10	0.23	0.08	0.11	0.31	0.19	0.42	0.48	0.37	0.12
Ele. HH, cash vil., high sub.	Yes	0.19	0.05	0.45	0.42	0.10	0.26	0.23	0.09	0.34	0.26	0.45	0.50	0.55	0.09
Ine. HH, non-cash vil., high sub.	Yes	0.08	0.07	0.02	0.04	-0.11	0.05	0.04	0.19	0.07	0.07	0.28	0.20	0.32	0.53
Ine. HH, cash vil., low sub.	Yes	0.07	0.06	0.04	0.07	-0.10	0.03	0.00	0.17	0.08	0.08	0.28	0.16	0.00	0.52
Ine. HH, cash vil., high sub.	Yes	0.06	0.07	0.07	0.11	-0.09	0.01	0.05	0.16	0.10	0.11	0.32	0.02	0.35	0.44
Consumption															
Ele. HH, non-cash vil., high sub.	Yes	0.07	0.05	0.03	0.04	-0.21	0.05	0.03	0.29	0.13	0.08	0.46	0.23	0.16	0.54
Ele. HH, cash vil., low sub.	Yes	0.21	0.05	0.41	0.15	-0.12	0.20	0.07	0.33	0.34	0.17	0.52	0.52	0.16	0.41
Ele. HH, cash vil., high sub.	Yes	0.19	0.04	0.46	0.22	-0.06	0.27	0.03	0.25	0.39	0.16	0.54	0.43	0.09	0.30
Ine. HH, non-cash vil., high sub.	Yes	0.13	0.08	0.01	0.05	-0.16	0.12	0.08	0.29	0.16	0.11	0.42	0.27	0.26	0.56
Ine. HH, cash vil., low sub.	Yes	0.08	0.06	0.03	0.05	-0.13	0.04	0.03	0.21	0.12	0.07	0.41	0.39	0.16	0.46
Ine. HH, cash vil., high sub.	Yes	0.13	0.06	0.05	0.10	-0.09	0.08	0.03	0.22	0.17	0.11	0.43	0.39	0.13	0.43
Mean		0.12	0.06	0.17	0.13	-0.07	0.12	0.07	0.19	0.19	0.13	0.40	0.36	0.26	0.39
Mental Health															
Consumption															
Cash	Yes	0.23	0.08	0.31	0.34	0.28	0.08	0.11	0.06	0.17	0.22	0.33	0.46	0.32	0.17
Cash spillover	Yes	0.07	0.05	0.04	0.07	-0.04	0.04	0.00	0.11	0.06	0.09	0.22	0.27	0.01	0.59
PMP	Yes	0.05	0.08	0.01	-0.10	-0.11	0.04	0.15	0.16	0.08	0.22	0.23	0.30	0.47	0.60
PMP spillover	Yes	0.03	0.05	0.00	-0.02	-0.07	0.03	0.05	0.10	0.03	0.11	0.16	0.84	0.48	0.71
Cash+PMP	Yes	0.08	0.06	0.29	0.15	0.17	0.21	0.08	0.09	0.23	0.25	0.36	0.66	0.16	0.27
Depression															
Cash	No	0.15	0.06	0.11	0.14	-0.20	0.04	0.01	0.35	0.13	0.24	0.53	0.25	0.05	0.53
Cash spillover	No	0.01	0.06	-0.10	-0.16	-0.66	0.11	0.17	0.68	0.12	0.20	0.77	0.64	0.67	0.68
PMP	No	-0.05	0.06	0.12	0.35	-0.05	0.17	0.40	0.01	0.17	0.40	0.50	0.61	0.42	0.00
PMP spillover	No	0.03	0.05	0.00	0.09	-0.39	0.03	0.06	0.42	0.04	0.14	0.53	0.80	0.50	0.72
Cash+PMP	No	0.07	0.07	0.17	0.54	0.07	0.10	0.47	0.00	0.18	0.49	0.52	0.41	0.24	0.00
Intimate Partner Violence															
Cash	Yes	-0.01	0.09	0.04	0.10	0.11	0.05	0.11	0.12	0.08	0.18	0.16	0.27	0.33	0.47
Cash spillover	Yes	-0.02	0.09	-0.03	-0.03	0.00	0.00	0.01	0.02	0.04	0.08	0.13	0.14	0.06	0.35
PMP	Yes	-0.07	0.09	0.10	0.26	0.20	0.17	0.33	0.27	0.17	0.33	0.29	0.70	0.52	0.44
PMP spillover	Yes	-0.04	0.09	0.01	0.12	0.04	0.05	0.16	0.09	0.06	0.17	0.14	0.57	0.63	0.47
Cash+PMP	Yes	0.07	0.11	0.12	0.36	0.26	0.05	0.30	0.19	0.07	0.30	0.21	0.46	0.45	0.41
Subjective well-being															
Cash	No	0.19	0.06	0.40	0.48	0.52	0.20	0.28	0.32	0.30	0.57	0.66	0.54	0.26	0.40
Cash spillover	No	-0.01	0.06	-0.22	-0.45	-1.04	0.20	0.43	1.03	0.30	0.52	1.12	0.41	0.48	0.49
PMP	No	0.04	0.06	0.27	0.63	0.21	0.23	0.59	0.17	0.28	0.60	0.59	0.57	0.44	0.31
PMP spillover	No	0.01	0.05	-0.04	0.04	-0.53	0.05	0.03	0.54	0.12	0.29	0.68	0.61	0.25	0.59
Cash+PMP	No	0.13	0.06	0.59	1.22	0.90	0.46	1.10	0.77	0.51	1.11	1.00	0.43	0.32	0.26
Mean		0.05	0.07	0.11	0.21	-0.02	0.12	0.24	0.28	0.16	0.33	0.46	0.50	0.35	0.42
Goal setting/Aspirations															
Assets															
Cash	Yes	0.28	0.03	0.18	0.10	0.09	0.10	0.17	0.19	0.18	0.24	0.30	0.21	0.36	0.26
Goal setting	Yes	0.06	0.03	0.04	0.17	0.04	0.03	0.11	0.02	0.06	0.14	0.17	0.18	0.72	0.18
Goal setting+Cash	Yes	0.24	0.03	0.22	0.43	0.22	0.01	0.19	0.02	0.18	0.29	0.30	0.03	0.57	0.01
Consumption															
Cash	Yes	0.16	0.04	0.38	0.29	0.30	0.22	0.13	0.14	0.26	0.26	0.39	0.53	0.36	0.32
Goal setting	Yes	0.07	0.04	0.10	-0.20	-0.04	0.03	0.27	0.11	0.11	0.41	0.33	0.18	0.38	0.25
Goal setting+Cash	Yes	0.15	0.05	0.40	0.12	0.25	0.25	0.02	0.10	0.28	0.42	0.48	0.50	0.01	0.23
Consumption (educational)															
Cash	Yes	0.08	0.04	0.12	0.06	0.07	0.04	0.02	0.00	0.07	0.15	0.15	0.40	0.15	0.00
Goal setting	Yes	0.01	0.03	0.06	0.17	0.06	0.05	0.17	0.05	0.06	0.19	0.14	0.53	0.66	0.33
Goal setting+Cash	Yes	0.12	0.04	0.20	0.39	0.16	0.08	0.26	0.04	0.13	0.30	0.21	0.50	0.55	0.20
Aspirations: Assets															
Cash	No	0.12	0.04	0.09	0.10	0.06	0.03	0.02	0.06	0.07	0.13	0.15	0.21	0.19	0.24
Goal setting	No	0.02	0.03	0.06	0.16	0.08	0.04	0.15	0.07	0.05	0.17	0.15	0.63	0.53	0.42
Goal setting+Cash	No	0.15	0.04	0.14	0.35	0.17	0.01	0.20	0.02	0.08	0.25	0.20	0.03	0.49	0.05
Aspirations: Child Education															
Cash	No	0.03	0.04	0.22	-0.38	-0.34	0.19	0.41	0.38	0.31	0.69	0.73	0.42	0.51	0.53
Goal setting	No	0.11	0.04	0.28	0.38	0.12	0.17	0.27	0.01	0.34	0.37	0.69	0.32	0.62	0.01
Goal setting+Cash	No	0.07	0.04	0.44	0.82	0.34	0.37	0.75	0.27	0.48	0.75	0.85	0.45	0.05	0.12
Aspirations: Income															
Cash	No	0.10	0.04	0.11	0.17	0.19	0.01	0.06	0.09	0.08	0.21	0.25	0.13	0.32	0.39
Goal setting	No	0.02	0.04	0.10	0.30	0.22	0.08	0.28	0.20	0.09	0.28	0.30	0.71	0.74	0.55
Goal setting+Cash	No	0.10	0.04	0.22	0.64	0.46	0.12	0.54	0.36	0.15	0.54	0.44	0.47	0.67	0.54
Mean		0.10	0.04	0.19	0.23	0.14	0.10	0.22	0.12	0.17	0.32	0.35	0.36	0.44	0.26
Overall Mean		0.08	0.06	0.15	0.20	0.03	0.11	0.19	0.20	0.17	0.28	0.40	0.41	0.36	0.36
Behavioral Mean		0.09	0.06	0.15	0.14	0.04	0.09	0.11	0.14	0.15	0.19	0.31	0.39	0.33	0.35
Subjective Mean		0.07	0.05	0.16	0.29	0.01	0.14	0.33	0.30	0.20	0.42	0.56	0.45	0.41	0.36

Notes: Col. 1 captures whether the outcome variable is behavioral (yes) or subjective (no). Cols. 2 and 3 depict the observed experimental effect (in standard deviations) and standard error. Cols. 4 to 6 display the mean forecast effect among academic, recipient, and nonexpert types. Cols. 7 to 9 present the negative absolute error of the crowd forecast. Cols. 10 to 12 display the average negative absolute error of individual forecasts. Cols. 13 to 15 depict the percent of individuals who are more accurate than the crowd (mean prediction) for each outcome. Individual-level forecasts are winsorized at the 5% level by magnitude at the type ×outcome level.

Table A3: Group Differences in Accuracy and Robustness

	Neg. abs. error (SD)		Error (SD)		Neg. quadratic error (SD)		SE-weighted neg. abs. error		Pearson correlation		Spearman correlation		Neg. abs. rank deviation		Identify worst	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
A. Passing comprehension, 5% winsorization																
Academic type (ref)	-0.17***		0.07***		-0.07***		-3.40***		0.59***		0.53***		-1.14***		0.61***	
	(0.01)		(0.01)		(0.01)		(0.16)		(0.02)		(0.02)		(0.04)		(0.03)	
Recipient type	-0.13***	-0.13***	0.05***	0.05***	-0.13***	-0.12***	-2.79***	-2.64***	-0.17***	-0.17***	-0.15***	-0.14***	-0.15***	-0.18***	-0.20***	-0.20***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.32)	(0.26)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.03)	(0.04)	(0.04)
Phys. tool	0.04***	0.04***	-0.02	-0.01	0.04***	0.04***	0.96***	0.90***	0.01	0.00	0.00	0.00	0.03	0.04	0.04	0.04
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.31)	(0.23)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)
Salient incentives	0.01	0.00	0.00	0.01	0.01	0.00	-0.01	0.02	0.02	0.02	0.00	0.01	0.01	0.00	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.30)	(0.22)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)
Nonexpert type	-0.23***	-0.23***	-0.13***	-0.12***	-0.33***	-0.33***	-4.41***	-4.46***	-0.27***	-0.27***	-0.23***	-0.24***	-0.17***	-0.16***	-0.11***	-0.11***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)	(0.37)	(0.37)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.03)	(0.04)	(0.04)
Salient incentives	-0.00	-0.01	-0.02	-0.01	0.01	-0.01	-0.30	-0.29	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.06	-0.06
	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.05)	(0.52)	(0.52)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)
n_i	15680	15680	15680	15680	15680	15680	15680	15680	3613	3613	3634	3634	15680	15680	2733	2733
n_f	951	951	951	951	951	951	951	951	948	948	948	948	951	951	609	609
B. Full sample, 5% winsorization																
Academic type (ref)	-0.23***		0.09***		-0.20***		-4.45***		0.59***		0.53***		-1.16***		0.61***	
	(0.01)		(0.01)		(0.02)		(0.24)		(0.02)		(0.02)		(0.04)		(0.03)	
Recipient type	-0.10***	-0.12***	0.04***	0.04***	-0.06**	-0.12***	-2.27***	-2.42***	-0.17***	-0.16***	-0.16***	-0.15***	-0.12**	-0.15***	-0.19***	-0.18***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.35)	(0.32)	(0.03)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.04)	(0.04)
Phys. tool	0.10***	0.01**	-0.06***	-0.05***	0.19***	-0.07***	1.58***	0.28**	-0.01	-0.01	-0.01	-0.01	0.02	0.05**	0.02	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.16)	(0.14)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)
Salient incentives	-0.00	-0.00	0.01	0.01	-0.00	0.00	-0.17	-0.04	-0.00	-0.01	-0.01	-0.01	0.02	0.01	-0.00	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.37)	(0.29)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.02)	(0.03)	(0.03)
Nonexpert type	-0.27***	-0.31***	-0.23***	-0.23***	-0.54***	-0.65***	-5.63***	-6.13***	-0.31***	-0.31***	-0.27***	-0.27***	-0.14***	-0.14***	-0.19***	-0.19***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.08)	(0.08)	(0.55)	(0.55)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.03)	(0.04)	(0.04)
Salient incentives	-0.01	-0.02	-0.03	-0.03	-0.03	-0.04	-0.56	-0.45	-0.01	-0.01	-0.02	-0.02	-0.02	-0.04	-0.00	-0.00
	(0.03)	(0.03)	(0.04)	(0.04)	(0.12)	(0.12)	(0.78)	(0.75)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)
n_i	21814	21814	21814	21814	21814	21814	21814	21814	5019	5019	5067	5067	21814	21814	3927	3927
n_f	1306	1306	1306	1306	1306	1306	1306	1306	1299	1299	1299	1299	1306	1306	876	876
C. Passing comprehension, 1% winsorization																
Academic type (ref)	-0.19***		0.06***		-0.19*		-3.93***		0.59***		0.53***		-1.14***		0.61***	
	(0.01)		(0.02)		(0.10)		(0.37)		(0.02)		(0.02)		(0.04)		(0.03)	
Recipient type	-0.14***	-0.13***	0.07***	0.07***	-0.07	-0.06	-2.84***	-2.66***	-0.17***	-0.16***	-0.16***	-0.15***	-0.15***	-0.18***	-0.19***	-0.19***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.11)	(0.11)	(0.52)	(0.47)	(0.03)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.04)	(0.04)
Phys. tool	0.05***	0.05***	-0.03*	-0.03*	0.08***	0.08***	1.27***	1.21***	-0.01	-0.01	-0.01	-0.01	0.03	0.04	0.03	0.03
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.39)	(0.30)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)
Salient incentives	0.01	0.01	0.00	0.01	0.02	0.02	0.06	0.14	-0.00	-0.01	-0.01	-0.01	0.01	0.00	0.01	0.01
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.38)	(0.30)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)
Nonexpert type	-0.26***	-0.26***	-0.09***	-0.09***	-0.48***	-0.49***	-4.95***	-5.00***	-0.31***	-0.31***	-0.27***	-0.27***	-0.17***	-0.16***	-0.11***	-0.11***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.15)	(0.15)	(0.67)	(0.65)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.03)	(0.04)	(0.04)
Salient incentives	-0.01	-0.02	-0.02	-0.02	-0.05	-0.06	-0.52	-0.45	-0.01	-0.01	-0.02	-0.02	-0.02	-0.03	-0.07*	-0.07
	(0.04)	(0.04)	(0.04)	(0.04)	(0.17)	(0.17)	(0.87)	(0.86)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)	(0.04)	(0.04)
n_i	15680	15680	15680	15680	15680	15680	15680	15680	5019	5019	5067	5067	15680	15680	2733	2733
n_f	951	951	951	951	951	951	951	951	1299	1299	1299	1299	951	951	609	609
Fixed effects		X		X		X		X		X		X		X		X

Notes * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level. Standard errors, in parentheses are clustered at the individual level. n_i refers to the number of individual forecasters, and n_f refers to the total number of forecasts. Cols 1-2 depict negative absolute forecast error. Cols 3-4 present forecast error. Cols 5-6 depict negative quadratic forecast error. Cols 7-8 reweight negative absolute forecast error by the standard error of the estimated effect. Cols 9-10 and 11-12 depict Pearson and Spearman correlation coefficients. Cols 13-14 depict the negative absolute difference between predicted and observed treatment ranking. Cols 15-16 depict an indicator variable equal to 1 if the respondent correctly forecasts which treatment would be least effective. Cols 2,4,6, and 14 include treatment ×outcome fixed effects. Cols 10, 12 and 18 include outcome fixed effects. In Panel A, the sample is restricted respondents passing comprehension questions, and forecasts are winsorized at the 5% level by magnitude at the type ×outcome level. In Panel B, the sample is not restricted, and forecasts are winsorized at the 5% level by magnitude at the outcome ×type level. In Panel C, the sample is restricted respondents passing comprehension questions, and forecasts are winsorized at the 1% level by magnitude at the type×outcome level. Cols 1-8 and 13-15 are at the treatment effect level. Cols 9-12 and 15-16 are at the outcome level. Cols 15-16 are restricted to comparisons between discrete non-spillover conditions (the cash transfer or the mental health intervention or combined intervention from Haushofer et al. (2020) and the cash transfer or aspirations/goal-setting or combined intervention from Orkin et al. (2020)).

Table A4: Summary of Accuracy by Condition
(Table A2 winsorized at the 1% level)

				Mean forecast effect (SD)			Neg. abs. error (mean for.)			Neg. abs. error (ind. for.)			% more accurate than crowd		
	Behavior (1)	Effect (SD) (2)	(se) (3)	Aca. (4)	Rec. (5)	MTu. (6)	Aca. (7)	Rec. (8)	MTu. (9)	Aca. (10)	Rec. (11)	MTu. (12)	Aca. (13)	Rec. (14)	MTu. (15)
General Equilibrium															
Assets															
Ele. HH, non-cash vil., high sub.	Yes	-0.03	0.06	0.04	0.09	0.08	0.07	0.11	0.11	0.08	0.13	0.30	0.75	0.55	0.28
Ele. HH, cash vil., low sub.	Yes	0.20	0.04	0.46	0.30	0.21	0.26	0.10	0.00	0.31	0.19	0.42	0.48	0.37	0.12
Ele. HH, cash vil., high sub.	Yes	0.19	0.05	0.50	0.48	0.19	0.31	0.29	0.01	0.34	0.26	0.45	0.50	0.55	0.09
Ine. HH, non-cash vil., high sub.	Yes	0.08	0.07	0.02	0.04	-0.09	0.05	0.04	0.17	0.07	0.07	0.28	0.20	0.32	0.53
Ine. HH, cash vil., low sub.	Yes	0.07	0.06	0.04	0.07	-0.05	0.03	0.00	0.13	0.08	0.08	0.28	0.16	0.00	0.52
Ine. HH, cash vil., high sub.	Yes	0.06	0.07	0.07	0.12	-0.04	0.01	0.05	0.11	0.10	0.11	0.32	0.02	0.35	0.44
Consumption															
Ele. HH, non-cash vil., high sub.	Yes	0.07	0.05	0.02	0.04	-0.16	0.05	0.04	0.23	0.13	0.08	0.46	0.23	0.16	0.54
Ele. HH, cash vil., low sub.	Yes	0.21	0.05	0.42	0.16	-0.06	0.21	0.06	0.27	0.34	0.17	0.52	0.52	0.16	0.41
Ele. HH, cash vil., high sub.	Yes	0.19	0.04	0.47	0.27	0.01	0.28	0.08	0.18	0.39	0.16	0.54	0.43	0.09	0.30
Ine. HH, non-cash vil., high sub.	Yes	0.13	0.08	0.00	0.05	-0.12	0.12	0.08	0.25	0.16	0.11	0.42	0.27	0.26	0.56
Ine. HH, cash vil., low sub.	Yes	0.08	0.06	0.03	0.05	-0.07	0.05	0.03	0.15	0.12	0.07	0.41	0.39	0.16	0.46
Ine. HH, cash vil., high sub.	Yes	0.13	0.06	0.05	0.12	-0.01	0.08	0.01	0.14	0.17	0.11	0.43	0.39	0.13	0.43
Mean		0.12	0.06	0.18	0.15	-0.01	0.13	0.07	0.14	0.19	0.13	0.40	0.36	0.26	0.39
Mental Health															
Consumption															
Cash	Yes	0.23	0.08	0.33	0.37	0.43	0.10	0.14	0.21	0.17	0.22	0.33	0.46	0.32	0.17
Cash spillover	Yes	0.07	0.05	0.04	0.07	0.00	0.04	0.00	0.07	0.06	0.09	0.22	0.27	0.01	0.59
PMP	Yes	0.05	0.08	0.02	-0.10	-0.11	0.04	0.15	0.16	0.08	0.22	0.23	0.30	0.47	0.60
PMP spillover	Yes	0.03	0.05	0.00	-0.02	-0.05	0.03	0.05	0.08	0.03	0.11	0.16	0.84	0.48	0.71
Cash+PMP	Yes	0.08	0.06	0.30	0.17	0.30	0.22	0.10	0.23	0.23	0.25	0.36	0.66	0.16	0.27
Depression															
Cash	No	0.15	0.06	0.11	0.12	-0.20	0.04	0.03	0.35	0.13	0.24	0.53	0.25	0.05	0.53
Cash spillover	No	0.01	0.06	-0.10	-0.20	-0.69	0.12	0.22	0.71	0.12	0.20	0.77	0.64	0.67	0.68
PMP	No	-0.05	0.06	0.12	0.35	-0.05	0.17	0.40	0.01	0.17	0.40	0.50	0.61	0.42	0.00
PMP spillover	No	0.03	0.05	0.00	0.09	-0.41	0.03	0.06	0.44	0.04	0.14	0.53	0.80	0.50	0.72
Cash+PMP	No	0.07	0.07	0.19	0.54	0.06	0.12	0.47	0.01	0.18	0.49	0.52	0.41	0.24	0.00
Intimate Partner Violence															
Cash	Yes	-0.01	0.09	0.05	0.10	0.11	0.06	0.11	0.12	0.08	0.18	0.16	0.27	0.33	0.47
Cash spillover	Yes	-0.02	0.09	-0.03	-0.03	0.00	0.00	0.01	0.02	0.04	0.08	0.13	0.14	0.06	0.35
PMP	Yes	-0.07	0.09	0.10	0.26	0.20	0.17	0.33	0.27	0.17	0.33	0.29	0.70	0.52	0.44
PMP spillover	Yes	-0.04	0.09	0.01	0.12	0.04	0.05	0.16	0.09	0.06	0.17	0.14	0.57	0.63	0.47
Cash+PMP	Yes	0.07	0.11	0.12	0.37	0.26	0.06	0.30	0.19	0.07	0.30	0.21	0.46	0.45	0.41
Subjective well-being															
Cash	No	0.19	0.06	0.41	0.47	0.52	0.21	0.28	0.32	0.30	0.57	0.66	0.54	0.26	0.40
Cash spillover	No	-0.01	0.06	-0.23	-0.50	-1.04	0.22	0.49	1.03	0.30	0.52	1.12	0.41	0.48	0.49
PMP	No	0.04	0.06	0.27	0.63	0.21	0.23	0.59	0.17	0.28	0.60	0.59	0.57	0.44	0.31
PMP spillover	No	0.01	0.05	-0.04	0.04	-0.53	0.06	0.03	0.54	0.12	0.29	0.68	0.61	0.25	0.59
Cash+PMP	No	0.13	0.06	0.61	1.22	0.90	0.48	1.09	0.77	0.51	1.11	1.00	0.43	0.32	0.26
Mean		0.05	0.07	0.11	0.20	0.00	0.12	0.25	0.29	0.16	0.33	0.46	0.50	0.35	0.42
Goal setting/Aspirations															
Assets															
Cash	Yes	0.28	0.03	0.18	0.10	0.09	0.09	0.17	0.19	0.18	0.24	0.30	0.21	0.36	0.26
Goal setting	Yes	0.06	0.03	0.04	0.17	0.06	0.03	0.11	0.01	0.06	0.14	0.17	0.18	0.72	0.18
Goal setting+Cash	Yes	0.24	0.03	0.23	0.47	0.26	0.01	0.23	0.02	0.18	0.29	0.30	0.03	0.57	0.01
Consumption															
Cash	Yes	0.16	0.04	0.44	0.32	0.36	0.28	0.16	0.20	0.26	0.26	0.39	0.53	0.36	0.32
Goal setting	Yes	0.07	0.04	0.10	-0.19	-0.03	0.03	0.26	0.10	0.11	0.41	0.33	0.18	0.38	0.25
Goal setting+Cash	Yes	0.15	0.05	0.48	0.17	0.32	0.33	0.03	0.18	0.28	0.42	0.48	0.50	0.01	0.23
Consumption (educational)															
Cash	Yes	0.08	0.04	0.12	0.06	0.09	0.04	0.02	0.01	0.07	0.15	0.15	0.40	0.15	0.00
Goal setting	Yes	0.01	0.03	0.06	0.19	0.06	0.05	0.18	0.06	0.06	0.19	0.14	0.53	0.66	0.33
Goal setting+Cash	Yes	0.12	0.04	0.21	0.43	0.20	0.09	0.30	0.08	0.13	0.30	0.21	0.50	0.55	0.20
Aspirations: Assets															
Cash	No	0.12	0.04	0.11	0.10	0.14	0.01	0.01	0.02	0.07	0.13	0.15	0.21	0.19	0.24
Goal setting	No	0.02	0.03	0.06	0.17	0.16	0.04	0.15	0.14	0.05	0.17	0.15	0.63	0.53	0.42
Goal setting+Cash	No	0.15	0.04	0.16	0.38	0.26	0.02	0.24	0.11	0.08	0.25	0.20	0.03	0.49	0.05
Aspirations: Child Education															
Cash	No	0.03	0.04	0.02	-0.45	-0.50	0.01	0.48	0.52	0.31	0.69	0.73	0.42	0.51	0.53
Goal setting	No	0.11	0.04	0.04	0.38	0.03	0.07	0.27	0.09	0.34	0.37	0.69	0.32	0.62	0.01
Goal setting+Cash	No	0.07	0.04	0.20	0.82	0.17	0.13	0.75	0.10	0.48	0.75	0.85	0.45	0.05	0.12
Aspirations: Income															
Cash	No	0.10	0.04	0.11	0.20	0.34	0.01	0.10	0.24	0.08	0.21	0.25	0.13	0.32	0.39
Goal setting	No	0.02	0.04	0.10	0.32	0.34	0.08	0.29	0.32	0.09	0.28	0.30	0.71	0.74	0.55
Goal setting+Cash	No	0.10	0.04	0.24	0.73	0.62	0.14	0.63	0.53	0.15	0.54	0.44	0.47	0.67	0.54
Mean		0.10	0.04	0.16	0.24	0.17	0.08	0.24	0.16	0.17	0.32	0.35	0.36	0.44	0.26
Overall Mean		0.08	0.06	0.15	0.20	0.06	0.11	0.21	0.21	0.17	0.28	0.40	0.41	0.36	0.36
Behavioral Mean		0.09	0.06	0.16	0.16	0.08	0.10	0.12	0.13	0.15	0.19	0.31	0.39	0.33	0.35
Subjective Mean		0.07	0.05	0.12	0.28	0.02	0.12	0.35	0.34	0.20	0.42	0.56	0.45	0.41	0.36

Notes: Col. 1 captures whether the outcome variable is behavioral (yes) or subjective (no). Cols. 2 and 3 depict the observed experimental effect (in standard deviations) and standard error. Cols. 4 to 6 display the mean forecast effect among academic, recipient, and nonexpert types. Cols. 7 to 9 present the negative absolute error of the crowd forecast. Cols. 10 to 12 display the average negative absolute error of individual forecasts. Cols. 13 to 15 depict the percent of individuals who are more accurate than the crowd (mean prediction) for each outcome. Individual-level forecasts are winsorized at the 5% level by magnitude at the type ×outcome level.

Table A5: Summary of Accuracy by Condition
(Table A2 with full sample)

				Mean forecast effect (SD)			Neg. abs. error (mean for.)			Neg. abs. error (ind. for.)			% more accurate than crowd		
	Behavior (1)	Effect (SD) (2)	(se) (3)	Aca. (4)	Rec. (5)	MTu. (6)	Aca. (7)	Rec. (8)	MTu. (9)	Aca. (10)	Rec. (11)	MTu. (12)	Aca. (13)	Rec. (14)	MTu. (15)
General Equilibrium															
Assets															
Ele. HH, non-cash vil., high sub.	Yes	-0.03	0.06	0.04	0.08	0.00	0.06	0.10	0.03	0.08	0.14	0.31	0.76	0.56	0.27
Ele. HH, cash vil., low sub.	Yes	0.20	0.04	0.43	0.30	0.06	0.22	0.10	0.15	0.31	0.20	0.43	0.49	0.36	0.17
Ele. HH, cash vil., high sub.	Yes	0.19	0.05	0.45	0.42	0.07	0.26	0.24	0.11	0.34	0.28	0.45	0.51	0.51	0.09
Ine. HH, non-cash vil., high sub.	Yes	0.08	0.07	0.02	0.03	-0.12	0.06	0.05	0.19	0.07	0.08	0.30	0.20	0.30	0.48
Ine. HH, cash vil., low sub.	Yes	0.07	0.06	0.04	0.06	-0.11	0.03	0.01	0.19	0.08	0.09	0.31	0.16	0.05	0.48
Ine. HH, cash vil., high sub.	Yes	0.06	0.07	0.07	0.09	-0.11	0.01	0.03	0.17	0.10	0.11	0.34	0.02	0.23	0.42
Consumption															
Ele. HH, non-cash vil., high sub.	Yes	0.07	0.05	0.03	0.02	-0.21	0.05	0.05	0.29	0.12	0.11	0.50	0.24	0.30	0.51
Ele. HH, cash vil., low sub.	Yes	0.21	0.05	0.41	0.15	-0.16	0.20	0.06	0.37	0.33	0.21	0.58	0.36	0.15	0.39
Ele. HH, cash vil., high sub.	Yes	0.19	0.04	0.46	0.26	-0.10	0.27	0.06	0.30	0.38	0.21	0.60	0.44	0.19	0.30
Ine. HH, non-cash vil., high sub.	Yes	0.13	0.08	0.01	0.04	-0.15	0.12	0.08	0.28	0.16	0.13	0.46	0.27	0.22	0.50
Ine. HH, cash vil., low sub.	Yes	0.08	0.06	0.03	0.05	-0.15	0.04	0.03	0.22	0.12	0.11	0.46	0.40	0.10	0.40
Ine. HH, cash vil., high sub.	Yes	0.13	0.06	0.05	0.11	-0.09	0.08	0.02	0.22	0.17	0.14	0.48	0.40	0.11	0.41
Mean		0.12	0.06	0.17	0.13	-0.09	0.12	0.07	0.21	0.19	0.15	0.44	0.35	0.26	0.37
Mental Health															
Consumption															
Cash	Yes	0.23	0.08	0.28	0.32	0.16	0.06	0.09	0.06	0.17	0.22	0.44	0.20	0.26	0.13
Cash spillover	Yes	0.07	0.05	0.03	0.05	-0.13	0.05	0.02	0.20	0.07	0.09	0.30	0.40	0.12	0.63
PMP	Yes	0.05	0.08	0.00	-0.10	-0.16	0.05	0.16	0.21	0.08	0.22	0.32	0.28	0.54	0.54
PMP spillover	Yes	0.03	0.05	-0.01	-0.03	-0.13	0.04	0.06	0.16	0.04	0.12	0.26	0.76	0.48	0.63
Cash+PMP	Yes	0.08	0.06	0.27	0.12	0.09	0.19	0.04	0.01	0.22	0.26	0.45	0.60	0.10	0.03
Depression															
Cash	No	0.15	0.06	0.10	0.07	-0.34	0.05	0.08	0.49	0.14	0.30	0.66	0.22	0.32	0.59
Cash spillover	No	0.01	0.06	-0.08	-0.20	-0.80	0.09	0.21	0.82	0.12	0.26	0.92	0.62	0.64	0.60
PMP	No	-0.05	0.06	0.11	0.34	-0.25	0.16	0.39	0.20	0.18	0.41	0.60	0.56	0.42	0.25
PMP spillover	No	0.03	0.05	0.00	0.04	-0.57	0.03	0.01	0.60	0.04	0.16	0.71	0.14	0.03	0.58
Cash+PMP	No	0.07	0.07	0.17	0.51	-0.13	0.10	0.44	0.20	0.18	0.49	0.60	0.40	0.26	0.20
Intimate Partner Violence															
Cash	Yes	-0.01	0.09	0.05	0.06	0.11	0.06	0.07	0.11	0.09	0.20	0.21	0.46	0.10	0.41
Cash spillover	Yes	-0.02	0.09	-0.03	-0.04	0.01	0.00	0.01	0.03	0.05	0.10	0.18	0.12	0.06	0.27
PMP	Yes	-0.07	0.09	0.10	0.26	0.17	0.17	0.33	0.24	0.17	0.34	0.30	0.64	0.49	0.39
PMP spillover	Yes	-0.04	0.09	0.00	0.10	0.04	0.05	0.15	0.09	0.07	0.18	0.18	0.54	0.59	0.34
Cash+PMP	Yes	0.07	0.11	0.13	0.36	0.22	0.06	0.29	0.15	0.08	0.30	0.23	0.42	0.42	0.39
Subjective well-being															
Cash	No	0.19	0.06	0.42	0.50	0.37	0.23	0.31	0.18	0.31	0.62	0.76	0.54	0.24	0.30
Cash spillover	No	-0.01	0.06	-0.23	-0.47	-0.99	0.21	0.45	0.98	0.31	0.56	1.13	0.42	0.47	0.40
PMP	No	0.04	0.06	0.29	0.60	-0.02	0.26	0.56	0.06	0.30	0.62	0.69	0.56	0.45	0.00
PMP spillover	No	0.01	0.05	-0.06	-0.02	-0.56	0.07	0.03	0.58	0.14	0.34	0.80	0.60	0.25	0.51
Cash+PMP	No	0.13	0.06	0.60	1.23	0.71	0.48	1.10	0.58	0.52	1.12	1.00	0.42	0.30	0.20
Mean		0.05	0.07	0.11	0.19	-0.11	0.12	0.24	0.30	0.16	0.34	0.54	0.44	0.33	0.37
Goal setting/Aspirations															
Assets															
Cash	Yes	0.28	0.03	0.18	0.15	0.00	0.10	0.13	0.28	0.19	0.28	0.38	0.20	0.19	0.51
Goal setting	Yes	0.06	0.03	0.03	0.20	-0.05	0.03	0.14	0.11	0.06	0.17	0.26	0.23	0.75	0.55
Goal setting+Cash	Yes	0.24	0.03	0.22	0.48	0.10	0.02	0.24	0.14	0.18	0.33	0.38	0.03	0.60	0.24
Consumption															
Cash	Yes	0.16	0.04	0.37	0.31	0.17	0.21	0.14	0.01	0.26	0.32	0.53	0.51	0.36	0.00
Goal setting	Yes	0.07	0.04	0.09	-0.15	-0.07	0.02	0.22	0.14	0.11	0.45	0.49	0.18	0.23	0.33
Goal setting+Cash	Yes	0.15	0.05	0.38	0.19	0.14	0.24	0.04	0.01	0.28	0.51	0.62	0.51	0.07	0.00
Consumption (educational)															
Cash	Yes	0.08	0.04	0.12	0.10	0.07	0.05	0.02	0.01	0.07	0.17	0.23	0.38	0.14	0.01
Goal setting	Yes	0.01	0.03	0.06	0.18	0.04	0.06	0.18	0.04	0.06	0.21	0.21	0.51	0.67	0.27
Goal setting+Cash	Yes	0.12	0.04	0.21	0.41	0.14	0.08	0.29	0.02	0.13	0.34	0.28	0.49	0.56	0.03
Aspirations: Assets															
Cash	No	0.12	0.04	0.09	0.13	-0.02	0.03	0.01	0.14	0.07	0.16	0.20	0.20	0.14	0.59
Goal setting	No	0.02	0.03	0.06	0.21	0.00	0.04	0.19	0.02	0.05	0.21	0.18	0.64	0.60	0.24
Goal setting+Cash	No	0.15	0.04	0.13	0.42	0.06	0.01	0.27	0.09	0.08	0.32	0.23	0.03	0.59	0.25
Aspirations: Child Education															
Cash	No	0.03	0.04	0.06	-0.37	-0.97	0.03	0.40	1.00	0.45	0.67	1.36	0.26	0.54	0.71
Goal setting	No	0.11	0.04	0.09	0.35	-0.62	0.03	0.24	0.74	0.53	0.40	1.32	0.00	0.58	0.75
Goal setting+Cash	No	0.07	0.04	0.24	0.77	-0.40	0.17	0.70	0.47	0.66	0.74	1.43	0.13	0.08	0.18
Aspirations: Income															
Cash	No	0.10	0.04	0.12	0.28	0.15	0.01	0.18	0.05	0.08	0.32	0.32	0.13	0.53	0.05
Goal setting	No	0.02	0.04	0.10	0.40	0.16	0.08	0.38	0.14	0.09	0.38	0.34	0.69	0.74	0.46
Goal setting+Cash	No	0.10	0.04	0.22	0.82	0.37	0.12	0.73	0.27	0.15	0.73	0.48	0.46	0.70	0.45
Mean		0.10	0.04	0.15	0.27	-0.04	0.07	0.25	0.20	0.19	0.37	0.51	0.31	0.45	0.31
Overall Mean		0.08	0.06	0.14	0.20	-0.08	0.10	0.20	0.24	0.18	0.31	0.51	0.37	0.35	0.35
Behavioral Mean		0.09	0.06	0.15	0.15	0.00	0.09	0.11	0.15	0.15	0.21	0.37	0.38	0.32	0.33
Subjective Mean		0.07	0.05	0.13	0.30	-0.20	0.12	0.35	0.40	0.23	0.46	0.72	0.37	0.42	0.39

Notes: Col. 1 captures whether the outcome variable is behavioral (yes) or subjective (no). Cols. 2 and 3 depict the observed experimental effect (in standard deviations) and standard error. Cols. 4 to 6 display the mean forecast effect among academic, recipient, and nonexpert types. Cols. 7 to 9 present the negative absolute error of the crowd forecast. Cols. 10 to 12 display the average negative absolute error of individual forecasts. Cols. 13 to 15 depict the percent of individuals who are more accurate than the crowd (mean prediction) for each outcome. Individual-level forecasts are winsorized at the 5% level by magnitude at the type ×outcome level.

Table A6: The Effect of Academic Expertise on Accuracy
(Table 2 with simple forecast error)

	Forecast error (SD)
Panel A	
Ref: Assistant prof	
PhD student	-0.003 (0.029)
Researcher or postdoc	-0.027 (0.034)
Associate professor	0.035 (0.033)
Full professor	0.039 (0.029)
n_i=123, n_f=2056	
Panel B	
log(cites)	0.003 (0.005)
n_i=81, n_f=1360	
Panel C	
Ref: No research in East Africa	
Research in East Africa	0.012 (0.022)
n_i=81, n_f=1360	

Notes: * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level. Column 1 presents simple forecast error, with standard errors clustered at the individual level. n_i refers to the number of individual forecasters, and n_f refers to the total number of forecasts. All models include condition×outcome fixed effects. Each panel represents a separate OLS regression. Panels B and C exclude PhD students. Observations are at the individual forecast×condition×outcome level. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level.

Table A7: The Effect of Academic Expertise on Accuracy
(Table 2 with correlational accuracy measures)

	Pearson cor.	Spearman cor.
	(1)	(2)
Panel A		
Ref: Assistant prof		
PhD student	-0.003 (0.029)	-0.003 (0.029)
Researcher or postdoc	-0.027 (0.034)	-0.027 (0.034)
Associate professor	0.035 (0.033)	0.035 (0.033)
Full professor	0.039 (0.029)	0.039 (0.029)
n_i=123, n_f=2056		
Panel B		
log(cites)	0.004 (0.010)	0.002 (0.010)
n_i=79, n_f=287		
Panel C		
Ref: No research in East Africa		
Research in East Africa	0.012 (0.022)	0.012 (0.022)
n_i=81, n_f=1360		

Notes: * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level. Cols. 1 and 2 present results from correlational accuracy outcomes (Pearson and Spearman coefficients). Standard errors are clustered at the individual level. n_i refers to the number of individual forecasters, and n_f refers to the total number of forecasts. All models include outcome fixed effects. Each panel represents a separate OLS regression. Panels B and C exclude PhD students. Observations are at the individual×outcome level. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level.

Table A8: Determinants of Accuracy: Recipient- and Nonexpert-Types
(Table 3 with simple forecast error)

	Forecast error		
	(1)	(2)	(3)
Panel A: Recipient Types			
Ref: No Physical Aid			
Physical aid	-0.011 (0.012)	-0.012 (0.012)	-0.016 (0.011)
Ref: No Salient Incentives			
Salient Incentives	-0.008 (0.012)	-0.005 (0.012)	-0.005 (0.011)
Ref: Nairobi Sample, From Nairobi			
Kirinyaga Sample, From Nairobi		-0.118* (0.031)	-0.105* (0.028)
Kirinyaga Sample, Not From Nairobi		-0.035* (0.017)	-0.015 (0.018)
Nairobi Sample, Not From Nairobi		-0.036* (0.018)	-0.010 (0.018)
Ref: Secondary School or Less			
More Than Secondary		-0.005 (0.013)	-0.004 (0.012)
Ref: Above Median Income			
Below Median Income		0.006 (0.012)	0.011 (0.011)
Ref: Enumerator 1			
Enumerator 2			0.031 (0.025)
Enumerator 3			0.074* (0.023)
Enumerator 4			0.045* (0.022)
Enumerator 5			-0.025 (0.021)
n_i=441, n_f=7380			
Panel B: Nonexpert Types			
Ref: No Salient Incentives			
Salient Incentives	0.013 (0.029)	0.013 (0.029)	
Ref: Less Than college			
Completed College (or above)		0.038 (0.032)	
Ref: Below $30,000			
Above 30,000		-0.027 (0.032)	
n_i=384, n_f=6208			

Notes: * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level. Panels A and B present results from simple forecast error. n_i refers to the number of individual forecasters, and n_f refers to the total number of forecasts. Standard errors clustered at the individual level are displayed in parentheses. All models include condition×outcome fixed effects. Observations are at the individual forecast×condition×outcome level. Individual-level forecasts are winsorized at the 5% level by magnitude type×outcome level

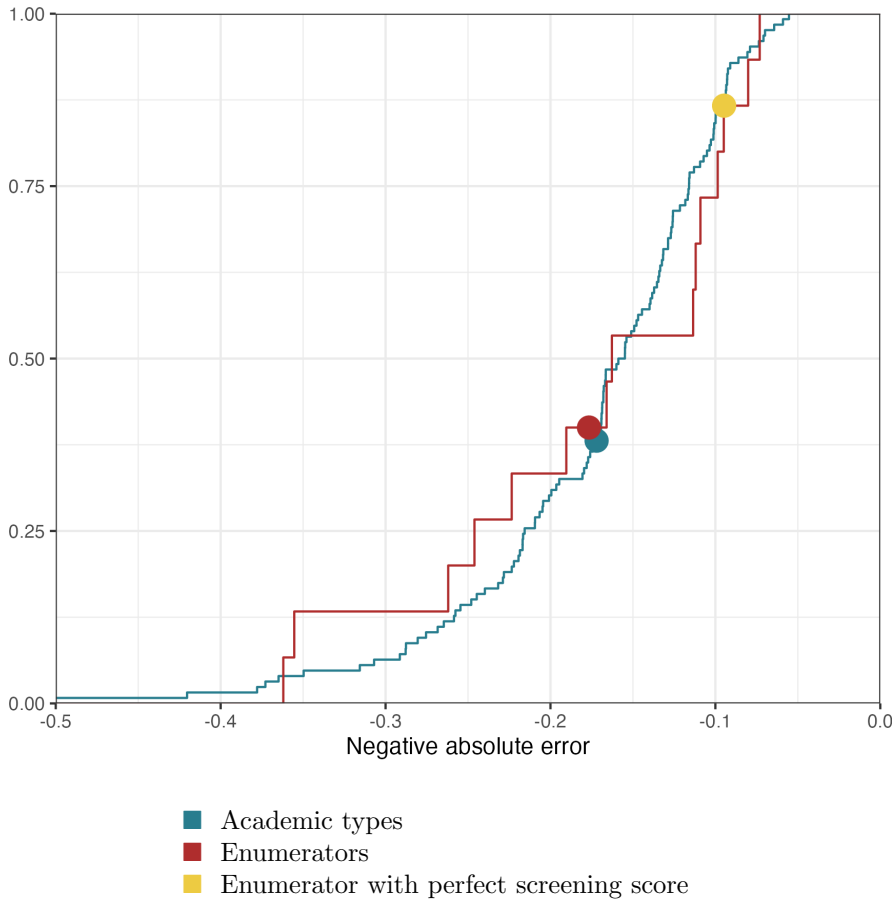
Table A9: Determinants of Accuracy: Recipient- and Nonexpert-Types
(Table 3 with correlation accuracy measures)

	Pearson correlation			Spearman correlation		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Recipient Types						
Ref: No Physical Aid						
Physical aid	0.006 (0.022)	0.005 (0.023)	0.003 (0.022)	0.004 (0.022)	0.003 (0.022)	0.003 (0.021)
Ref: No Salient Incentives						
Salient Incentives	-0.017 (0.022)	-0.017 (0.022)	-0.021 (0.021)	-0.003 (0.022)	-0.004 (0.022)	-0.009 (0.021)
Ref: Nairobi Sample, From Nairobi						
Kirinyaga Sample, From Nairobi		0.002 (0.072)	0.023 (0.063)		-0.038 (0.072)	-0.014 (0.066)
Kirinyaga Sample, Not From Nairobi		-0.024 (0.034)	-0.020 (0.033)		-0.040 (0.030)	-0.046 (0.030)
Nairobi Sample, Not From Nairobi		0.016 (0.034)	0.015 (0.034)		0.001 (0.032)	-0.015 (0.032)
Ref: Secondary School or Less						
More Than Secondary		0.026 (0.026)	0.025 (0.025)		0.014 (0.024)	0.013 (0.023)
Ref: Above Median Income						
Below Median Income		-0.016 (0.023)	-0.013 (0.022)		-0.010 (0.022)	-0.017 (0.021)
Ref: Enumerator 1						
Enumerator 2			0.066 (0.046)			0.033 (0.041)
Enumerator 3			0.087* (0.045)			0.100** (0.043)
Enumerator 4			0.172*** (0.047)			0.146*** (0.043)
Enumerator 5			0.111** (0.045)			0.141*** (0.040)
n_i=441, n_f=1776						
Panel B: Nonexpert Types						
Ref: No Salient Incentives						
Salient Incentives	0.016 (0.039)	0.022 (0.039)		0.025 (0.038)	0.031 (0.037)	
Ref: Less Than college						
Completed College (or above)		0.060 (0.044)			0.051 (0.042)	
Ref: Below $30,000						
Above 30,000		0.048 (0.043)			0.049 (0.042)	
n_i=383, n_f=1410						

Notes: * denotes significance at 10 pct., ** at 5 pct., and *** at 1 pct. level. Panels A and B display results from correlational accuracy outcomes (Pearson and Spearman coefficients). Standard errors clustered at the individual level are displayed in parentheses. n_i refers to the number of individual forecasters, and n_f refers to the total number of forecasts. All models include condition×outcome fixed effects. Observations are at the individual forecast×outcome level. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level.

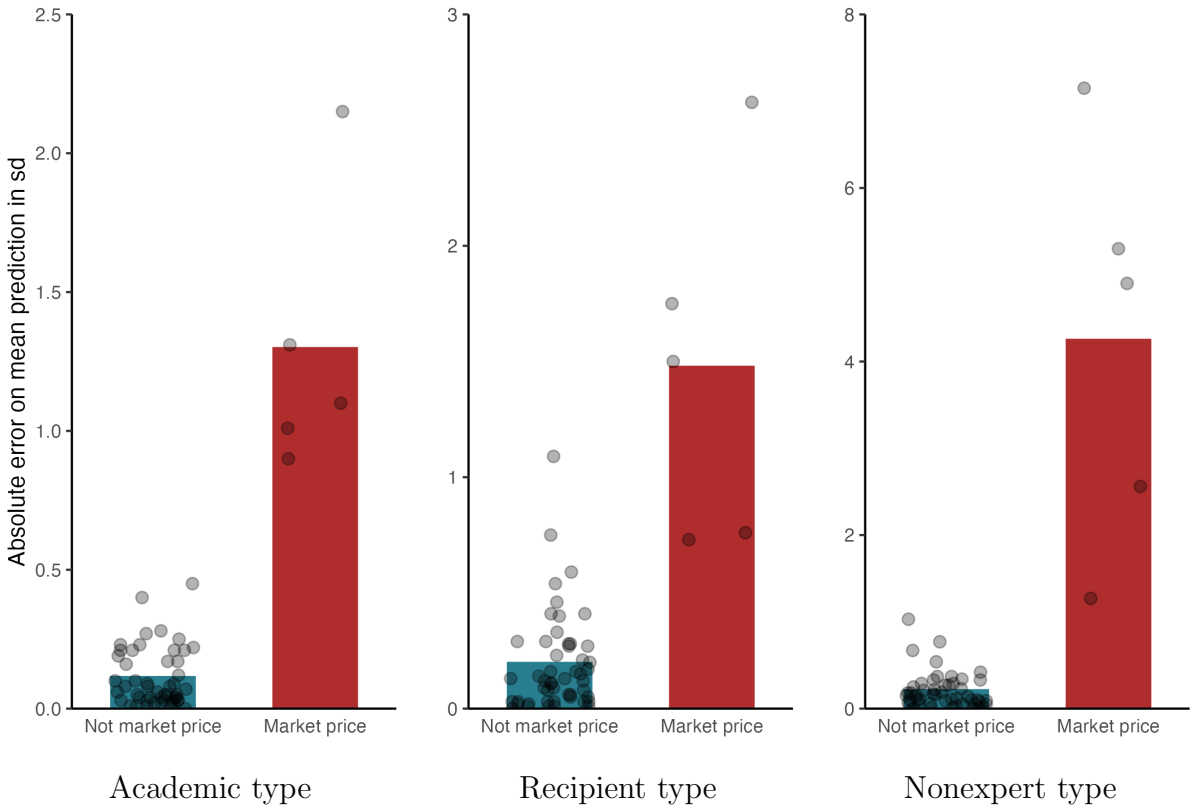
Appendix Figures

Figure A1: Cumulative Distribution Functions of Negative Absolute Error Among Academic Types and Enumerators



Notes: C.d.f.'s of negative absolute error comparing academic types (blue) and enumerators (red). The x-axis displays negative absolute forecast error comparing predicted to observed experimental results in standard deviations. Observations are at the respondent×experiment level. Points on each c.d.f denote the average negative absolute error across the entire group. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level. Enumerator forecasts are unwinsorized due to the small number of forecasts for each outcome. This analysis was not pre-registered.

Figure A2: Absolute Forecast Error for Market Price and Non-Market Price Effects



Notes: Bars depict the absolute error from the mean forecast, averages across predicted effects. Blue bars depict the average absolute error from the mean forecast for the fifty non-market outcomes, and red bars depict the average absolute error from the five market price outcomes. Points depict the average forecast error for each effect, with a horizontal jitter to display dispersion. Note that the scales differ for each figure. Predictions are at the type×effect level. Individual-level forecasts are winsorized at the 5% level by magnitude at the type×outcome level. This analysis was not pre-registered.