Policy Choice and the Wisdom of Crowds

Nicholas G. Otis*
UC Berkeley

Abstract

Using data from seven large-scale randomized experiments, I test whether crowds of academic experts can forecast the relative effectiveness of policy interventions. Eight-hundred and sixty-three academic experts provided 9,295 forecasts of the causal effects from these experiments, which span a diverse set of interventions (e.g., information provision, psychotherapy, soft-skills training), outcomes (e.g., consumption, COVID-19 vaccination, employment), and locations (Jordan, Kenya, Sweden, the United States). For each policy comparisons (a pair of policies and an outcome), I calculate the percent of crowd forecasts that correctly rank policies by their experimentally estimated treatment effects. While only 65% of individual experts identify which of two competing policies will have a larger causal effect, the average forecast from bootstrapped crowds of 30 experts identifies the better policy 86% of the time, or 92% when restricting analysis to pairs of policies who effects differ at the $p < 0.10$ level. Only 10 experts are needed to produce an 18-percentage point (27%) improvement in policy choice.

*notis@berkeley.edu. I gratefully acknowledge financial support from GiveWell, the Global Priorities Institute, the National Institute of Aging (T32-AG000246), the Russell Sage Foundation, and the Weiss Fund for Development Economics. I thank David Bernard, Diag Davenport, Stefano DellaVigna, Eugen Dimant, Will Dow, Anna Dreber, Amanda Geiser, David Mckenzie, Don Moore, Kate Orkin, and participants at AFE 2022 for helpful comments. I would also like to thank the authors of the predicted studies for sharing their data, and the forecasters for their time. Zimai Lan provided excellent research assistance, and Chaning Jang, Irene Ngina, and Pauline Wanjeri provided outstanding management of field operations. This project was approved by the U.C. Berkeley Committee for the Protection of Human Subjects.

1 Introduction

Policy decisions should be made using the best available information. I test whether academic experts can forecast the efficacy of policy interventions and benchmark these predictions against gold-standard causal evidence from seven well-powered randomized policy experiments. If accurate, such predictions could be used to prioritize which interventions to test in randomized experiments or scale when experimentation is infeasible.

I leverage a unique data set in which 863 academic experts provide 9,295 predictions of policy interventions tested in seven well-powered randomized controlled trials (RCTs). These trials are diverse, and include, for example, an experiment testing behavioral and financial incentives to increase vaccination in Sweden (Campos-Mercade et al., 2021) and a trial in rural Kenya evaluating the relative effects of cash transfers and psychotherapy (Haushofer et al., 2021). I use the experimentally estimated effects of these interventions to rank pairs of competing policies within each study, producing a total of 161 policy comparisons (a pair of policies and an outcome). A separate group of academic experts provided predictions of the causal effects of the interventions tested in each of the seven experiments. I generate thousands of bootstrapped *crowds* of experts and assess how often the mean (or median) predicted experimental treatment effect from a crowd of size c correctly ranks pairs of policies by their estimated causal effects.

The predicted experiments offer robust causal evidence against which policy choices from crowd predictions can be evaluated. This separates my work from a large literature focused on predicting trends in economic and social indicators like employment and COVID-19 infection rates, but not the causal effects interventions have on these outcomes. Additionally, several features of the predicted experiments make them ideal for studying the ability of crowds of experts to identify better-performing policies. First, these studies are heterogeneous, (a) spanning multiple countries (Jordan, Kenya, Sweden, the United States), (b) include interventions ranging from light-touch behavioral nudges to intensive interventions like wage subsidies, cash transfers, and psychotherapy, and (c) include a range of several welfare-relevant outcomes (e.g., intimate partner violence, educational expenditure, and COVID-19 vaccination). Second, a separate sample of experts provide forecasts for each of the seven experiments, meaning that these results are not limited to a specific group of experts. Third, the analysis is restricted to *behavioral* (as opposed to subjective) outcomes. As an example of why this is important, note that two of the three non-financial interventions in Campos-Mercade et al. (2021) significantly increased self-reported vaccination *intentions*, but not actual vaccination.

I present three main results. First, individual experts exhibit considerable disagreement on the relative effectiveness of policies, with only 65% of forecasters correctly predicting which policy in a given pair will have the larger effect. Second, there are large improve-

ments in policy rankings from eliciting predictions from multiple forecasters, aggregating their predictions, and using their combined forecast to select policies. Crowds of 30 forecasters identify the more effective policy in each pair 86% of the time, and 92% of the time if analysis is restricted to policy comparisons that differ significantly at the $p < 0.10$ level. These wisdom-of-crowds effects occur because individual forecasts are noisy, and aggregating predictions helps to reduce these errors (Galton, 1907; Surowiecki, 2005). Finally, the finding that crowds perform well at identifying higher-impact policies is quite general: meaningful improvements in policy choice are observed across all seven randomized experiments, though there is heterogeneity in both the level of initial disagreement among experts and the magnitude of improvements from using aggregate crowd forecasts. Importantly, these results are not mechanical. The ability of crowd-forecasts to correctly rank policies is a function of the empirical distribution of experts' beliefs, and aggregation would not produce such improvements if these beliefs were inaccurate.

This paper contributes to an emerging literature examining predictions of social science results (Dreber et al., 2015; DellaVigna et al., 2019, 2020; SSPP, 2022). As more researchers adopt this practice, it is increasingly important to provide systematic evidence on the extent to which these predictions can inform policy. My results suggest that predictions from crowds of experts provide a reasonably effective mechanism for generating ex ante evidence on the relative effectiveness of policy interventions.

The remainder of the paper proceeds as follows. Section 2 provides an overview of the sample of studies, forecasters, and empirical methods (Appendix A provides further details on the empirical framework). Section 3 presents the main study results, and Section 4 concludes.

2 Study Design

Overview of predicted studies. The predicted experiments cover a broad set of contexts and interventions: Campos-Mercade et al. (2021) evaluate the effects of financial incentives and three behavioral nudges (highlighting social impact, developing an argument around vaccination, and an informational quiz) on COVID-19 vaccination measured using administrative data in Sweden. Chopra et al. (2022) experimentally test how fact-checking impacts demand for news about an economic recovery plan among U.S. democrats, varying whether the news is from an ideologically aligned source (Fox or MSNBC). DellaVigna and Pope (2018a,b) run a large online experiment evaluating the effects of financial and nonfinancial incentives on a real-effort task designed to imitate repetitive clerical jobs. Dimant et al. (2022) collect predictions of results of an experiment evaluating seven nudges designed to increase use of masks to prevent COVID-19 (Gelfand et al., 2022). Groh et al. (2016) run

an experiment comparing the effects of soft-skills training and wage subsidies on short- and long-term employment among female college graduates in Jordan. Haushofer et al. (2021) evaluate the effects of a large-scale psychotherapy intervention on intimate partner violence and consumption in rural Kenya, which was benchmarked against a cash-transfer intervention. Orkin et al. (2020) evaluate the effects of an unconditional cash transfer and an aspirations and goal-setting intervention on asset ownership, consumption, and education expenditure in rural Kenya. Experimental details are provided in Appendix C, and Table 1 provides an overview of the experimental features in each study.

The study authors collected predictions for five of the seven experiments (Chopra et al., 2022; Campos-Mercade et al., 2021; Dimant et al., 2022; Groh et al., 2016; DellaVigna and Pope, 2018a,b). These constitute, to my knowledge, every published study eliciting predictions from academic experts of the causal effects of interventions (see Appendix B for exclusion criteria). I supplement these studies with predictions I collected for two additional experiments (Otis, 2021; Haushofer et al., 2021; Orkin et al., 2020).

Sample of forecasters. Predictions were elicited from a separate sample of academic experts for each study, and respondents provided forecasts independently, without communication. This decentralized approach to collecting predictions means that the performance of crowds of forecasters is not limited to a specific sample of respondents. However, it also leads to heterogeneity in the information I have on each group of forecasters. Across the seven studies, the modal forecaster is a faculty member or PhD student in economics. In total, my dataset contains 9,295 forecasts from 863 forecasters. Table 1 provides a breakdown of the number of forecasters and forecasts by study, and Appendix B provides details on the sample of forecasters by study.

Overview of empirical methods. I evaluate policy choice from crowds of forecasters using a simple and transparent empirical method: A crowd of size c chooses Policy A over Policy B if the aggregate (mean or median) prediction for the effect of Policy A is larger than for Policy B. The crowd has chosen correctly if the estimated causal effect from Policy A is larger than Policy B. In each policy experiment, I generate 5,000 crowds of $c = 1, \ldots, 30$ forecasters and calculate each crowd's aggregate predicted causal effect for each policy. I then test whether, for each *policy combination* (an outcome and a pair of policies), the crowd prediction correctly identifies which policy has a larger causal effect. Finally, I calculate the percent of the 5,000 crowds (at each crowd size) identifying the better performing policy.

For example, in Campos-Mercade et al. (2021) I calculate each crowd's aggregate predicted causal effect for the four non-control policy interventions on vaccination. 29 of the 52 individual forecasters (crowds of size 1) correctly predicted that financial incentives would

lead to a larger increase in vaccination than an intervention delivering information about vaccine safety and efficacy. At c=5, I generate 5,000 crowds of five forecasters bootstrapped-sampled from the full group of 52 forecasters. In 87% of these five-person crowds, the mean crowd forecast correctly identifies financial incentives as the better performing policy, and at c=10, 95% of crowds make the correct policy choice.

My main set of results pool the percent of correct policies for crowds of a given size across all seven experiments, weighting each experiment equally and weighting each policy comparison equally within experiments. Details on my empirical framework are provided in Appendix A.

3 Results

3.1 Disagreement Among Individual Experts

The first result is that there is substantial disagreement among individual academic experts regarding which policy will be more effective (see Appendix B for study and sample exclusion criteria and Appendix C for sample details). Individual forecasters select the better (higher estimated causal effect) policy among pairs of interventions 65% of the time; experts are better than chance (50%) at selecting the higher impact policy, but there is much room for improvement. We also observe variation in accuracy by outcome. For example, 98% of forecasters correctly predict that unconditional cash transfers increase consumption more than psychotherapy (in Kenya), while only 30% correctly predict that a message highlighting how wearing a mask prevents people from "being contaminated by a disgusting virus" will be more effective at inducing recipients to take a pledge to wear a mask than a message emphasizing how many Americans COVID-19 has killed.

3.2 The Wisdom of Crowds

How well do crowds perform compared to individual experts? Panel A of Figure 1 presents the percent of correct policy choices from crowds of size $c = 1, \ldots, 30$. Compared to individual experts, crowds of ten are on average 27% (18 pps) more likely to select the better performing policy, and crowds of thirty experts are on average 32% (21 pps) more likely to identify the better policy, choosing the higher impact intervention 86% of the time. Panel A of Figure A1 extends these results to crowds of up to size 100, showing that there are still small improvements in policy choice among larger crowds, with 100-forecaster crowds selecting the better policy 87% of the time.

These results are not merely mechanical. Taking bootstrapped samples of experimental participants (as opposed to forecasters) would produce a mechanical improvement in the

correct ranking of policies as larger samples generate closer approximations of the full experimental results. However, rankings based on the predictions of groups of forecasters need not converge to the correct ranking of policies by their causal effects, as even large crowds can be wrong (i.e., can have aggregate forecasts that predict lower impact policies will be more effective). For some policy comparisons larger crowds do perform worse, and the net improvement in policy choice as crowd size increases reflects a novel empirical result as opposed to a statistical artifact.

Significant policy comparisons. Panel B of Figure 1 excludes policy comparisons (a pair of policies and an outcome) that were not significant at the $p < 0.10$ level. For example, wage subsidies led to a 37-percentage point increase in employment among Jordanian women relative to soft-skills training in the *short run*, and crowds of size $c = 30$ correctly identified wage subsidies as the more effective policy 95% of the time (compared to just 57% for individual experts). However, both interventions produce *long-run* effects of 3 pps or less, and the difference between conditions is not significant at the $p < 0.10$ level. Panel B would therefore exclude the long-run outcome (see Table A1 for details on the number of experimental features and forecasts meeting this exclusion criteria). In other words, this robustness check focuses on pairs of policies where we can be somewhat more confident that one policy is more effective than the other. However, it does not eliminate the risk that differences between policies are due to sampling variation, which would, if anything, lower the accuracy of expert forecasts. For this set of policies, the wisdom-of-crowds improvement over individual experts from crowds of size $c = 10$ ($c = 30$) is 22 pps (26 pps), with crowds of 30 experts selecting the better-performing policy 92% of the time, and crowds of size 100 selecting the better policy 94% of the time (see Panel B of Figure A1 for results with crowds up to size 100).

Weighting strategies. The results presented so far give equal weight to each of the seven studies, and also weight each policy comparison (a pair of policies and an outcome) equally within studies. This reduces the likelihood that results are driven by a specific experimental setting, set of treatments, or group of forecasters. In a series of robustness checks, I present results weighting by the number of (1) forecasters, (2) forecasts, and (3) policy comparisons. Details on these weights are provided in Section A. Figure 2 presents results for my primary weighting scheme (Panel A) and the three alternative weighting schemes (Panels B-D). The improvement in policy choice from 1 forecaster to crowds of 30 ranges from 14 percentage points (23%) for forecast weights to 21 percentage points (32%) for the primary study weights. Restricting analysis to significantly different policies (at the $p < 0.10$ level), the smallest accuracy improvement from crowds of size 1 to size 30 is 19 percentage points (29%).

Forecast aggregation. Figure 2 also presents results using the median as opposed to the mean crowd forecast to aggregate predictions. The median forecast is less sensitive to outliers than the mean, and differences in performance between aggregation methods depend largely on whether these outliers shift crowd predictions towards or away from the correct ranking. Results are stable across aggregation methods, though the mean forecast generally yields more accurate policy rankings. For example, the improvement in policy choice from crowds of size 1 to 30 (when using the default study weights) is 17 pp for the median forecast, compared to 21 pp for the mean.

Heterogeneity by study. Figure 3 disaggregates results at the study level, averaging the percent of successful policy choices within each study and weighting each policy comparison equally (Figure A2 provides a robustness check restricting to significantly different policy comparisons, and Figure A3 extends results to crowds of up to size 100). For example, in SOFTSKILLS there are two outcomes, short- and long-term employment, and for each I compare two policies (soft-skills training and wage subsidies). At $c = 10$, the mean percent of correct choices are 81% and 50% respectively, yielding an average percent of correct policy choices of (81+50)/2=65.5%. Strong wisdom-of-crowds effects are observed in each study, and while there is heterogeneity in magnitude, the effects are consistently large. For example, in VACCINATION we observe an improvement of 45% (25 pps) for individual experts compared to ten-person crowds, while in EFFORT the improvement is 21% (13 pps).

4 Discussion

It is often not clear ex ante which policies will be most effective. This paper evaluates the wisdom of crowds in the context of policy choice by testing the extent to which groups of academic experts make more accurate predictions than individuals, using novel data in which academic experts forecast the causal effects of interventions whose impacts were estimated in large, randomized experiments. This allows me to identify whether crowds of experts can correctly predict which policy will produce larger causal effects.

While even small improvements in policy choice from crowd predictions could lead to meaningful welfare improvements among policy recipients, my results suggest that improvements in policy choice from crowd predictions are large and robust across a diverse range of empirical settings. I show that crowds of 30 forecasters make substantially more accurate policy choices, selecting the better performing policy 86% of the time, or 92% of the time when restricting analysis to policy comparisons with significant differences.

This paper contributes to an emerging literature exploring forecasts in the social sciences (Dreber et al., 2015; DellaVigna et al., 2019, 2020). This literature demonstrates that accuracy improvements can often be achieved by employing weighting schemes that incorporate heterogenous forecaster effort or ability (Tetlock and Gardner, 2015; DellaVigna and Pope, 2018a). For simplicity and transparency, I employ a forecaster-level weighting scheme that weights forecasters equally when aggregating predictions, meaning the results are likely a lower bound on the achievable crowd performance in policy choice. While I have focused on policy choice by academic experts, future research could incorporate predictions from other groups, such as people similar to policy recipients who may possess local contextual knowledge (Thomas et al., 2020). For example, Otis (2022) use crowd forecasts from local participants in Kenya to select nudges for evaluation in a large randomized controlled trial, and find that these interventions significantly improve adoption of a coronavirus notification service.

In light of these results, how should researchers and policymakers weigh forecasts against more traditional sources of evidence? While my results suggest that forecasts from crowds of academic experts provide substantial information on the relative performance of policy interventions, they should be viewed as a complement and not a substitute for more robust identification strategies like well-powered randomized experiments. Further work, both theoretical and empirical, is needed to assess the relative value of different tools for policy choice.

References

Abaluck, J., Kwong, L. H., Styczynski, A., Haque, A., Kabir, M. A., Bates-Jefferys, E., Crawford, E., Benjamin-Chung, J., Raihan, S., Rahman, S., et al. (2021). Impact of community masking on covid-19: A cluster-randomized trial in bangladesh. *Science*, page eabi9069.

Campos-Mercade, P., Meier, A. N., Schneider, F. H., Meier, S., Pope, D., and Wengström, E. (2021). Monetary incentives increase covid-19 vaccinations. *Science*, page eabm0475.

Casey, K., Glennerster, R., Miguel, E., and Voors, M. J. (2021). Long run effects of aid: Forecasts and evidence from sierra leone. Technical report, National Bureau of Economic Research.

Chadimová, K., Cahlíková, J., and Cingl, L. (2022). Foretelling what makes people pay: Predicting the results of field experiments on tv fee enforcement. *Journal of Behavioral and Experimental Economics*, page 101902.

Chopra, F., Haaland, I., and Roth, C. (2022). Do people demand fact-checked news? evidence from us democrats. *Journal of Public Economics*, 205:104549.

DellaVigna, S. and Linos, E. (2022). Rcts to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1):81–116.

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). 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.

Dimant, E., Pieper, D., Clemente, E. G., Dreber, A., and Gelfand, M. J. (2022). Politicizing mask-wearing: predicting the success of behavioral interventions among republicans and democrats. *Forthcoming in Science Advances*.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., and Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. W. (2020). General equilibrium effects of cash transfers: experimental evidence from kenya. Technical report, National Bureau of Economic Research.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451.

Gelfand, M., Li, R., Stamkou, E., Pieper, D., Denison, E., Fernandez, J., Choi, V., Chatman, J., Jackson, J., and Dimant, E. (2022). Persuading republicans and democrats to comply with mask wearing: An intervention tournament. *Journal of Experimental Social Psychology*, 101:104299.

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):1–23.

Haushofer, J., Mudida, R., and Shapiro, J. (2021). The comparative impact of cash transfers and a psychotherapy program on psychological and economic well-being. *Available at SSRN 3759722*.

Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J. E., et al. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences*, 119(6):e2115126119.

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., et al. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, 600(7889):478–483.

Orkin, K., Garlick, R., Mahmud, M., Sedlmayr, R., Haushofer, J., and Dercon, S. (2020). Aspirations, assets, and anti-poverty policies. Technical report, Mimeo.

Otis, N. G. (2021). Forecasting in the field. *Working paper*. Retrieved from `https://nicholasotis.com/Research/Otis_ForecastingField.pdf`.

Otis, N. G. (2022). The efficacy of crowdsoured nudges: Experimental evidence. *Working paper*. Retrieved from `https://nicholasotis.com/Research/Otis_Crowdsourcing.pdf`.

SSPP (2022). Social science prediction platform.

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

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

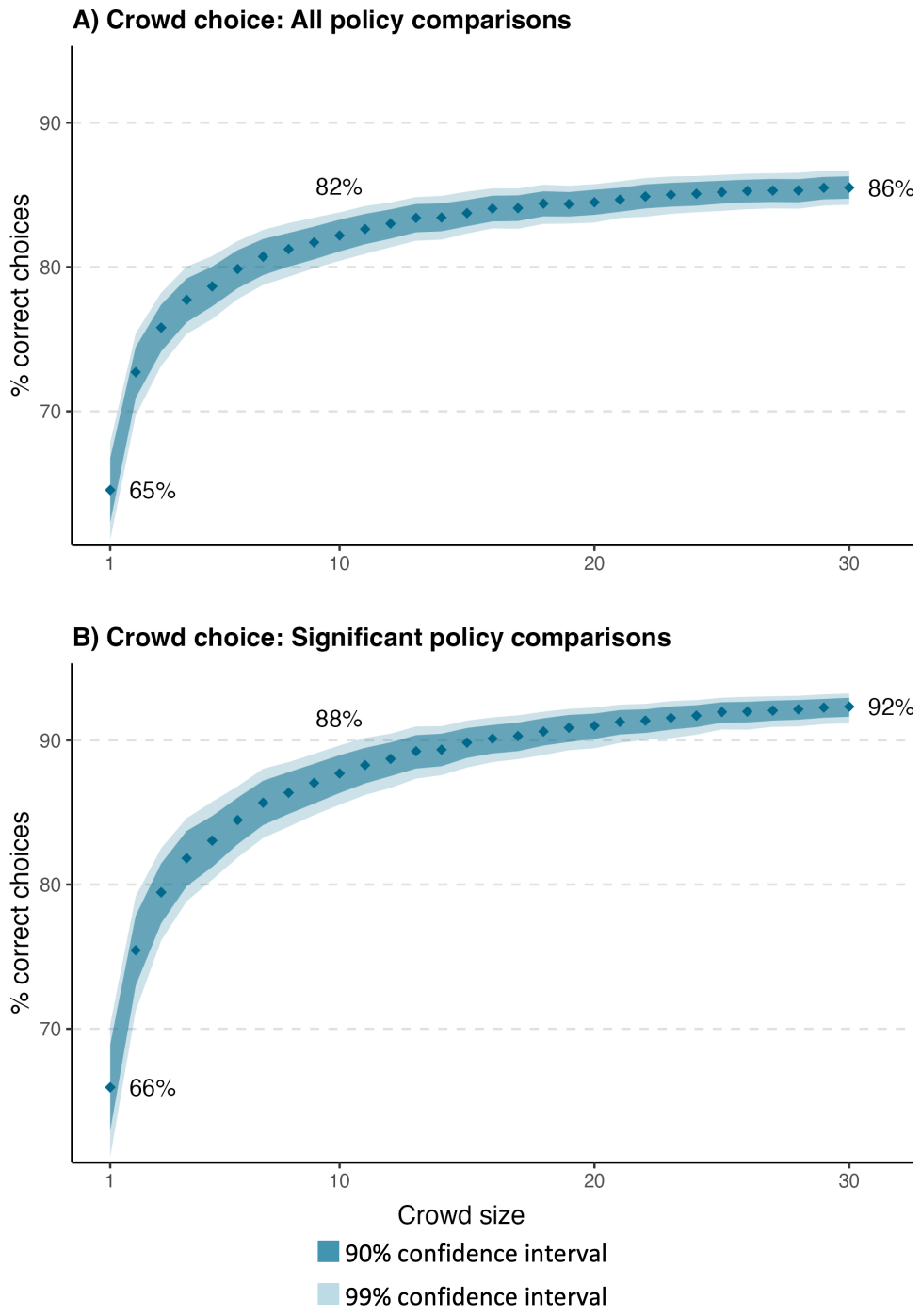
Thomas, C. C., Otis, N. G., Abraham, J. R., Markus, H. R., and Walton, G. M. (2020). Toward a science of delivering aid with dignity: Experimental evidence and local forecasts from kenya. *Proceedings of the National Academy of Sciences*, 117(27):15546–15553.

Table 1: Comparison of experimental features

Experiment	Number of				
	Policies	Outcomes	Policy comparisons	Forecasters	Forecasts
	(1)	(2)	(3)	(4)	(5)
ASPIRATIONS	2	3	3	39	234
EFFORT	15	1	105	355	5325
MASK	7	2	42	199	2786
NEWS	2	1	1	65	130
PSYCHOTHERAPY	2	2	2	50	200
SOFTSKILLS	2	2	2	103	412
VACCINATION	4	1	6	52	208
Total	34	12	161	863	9295

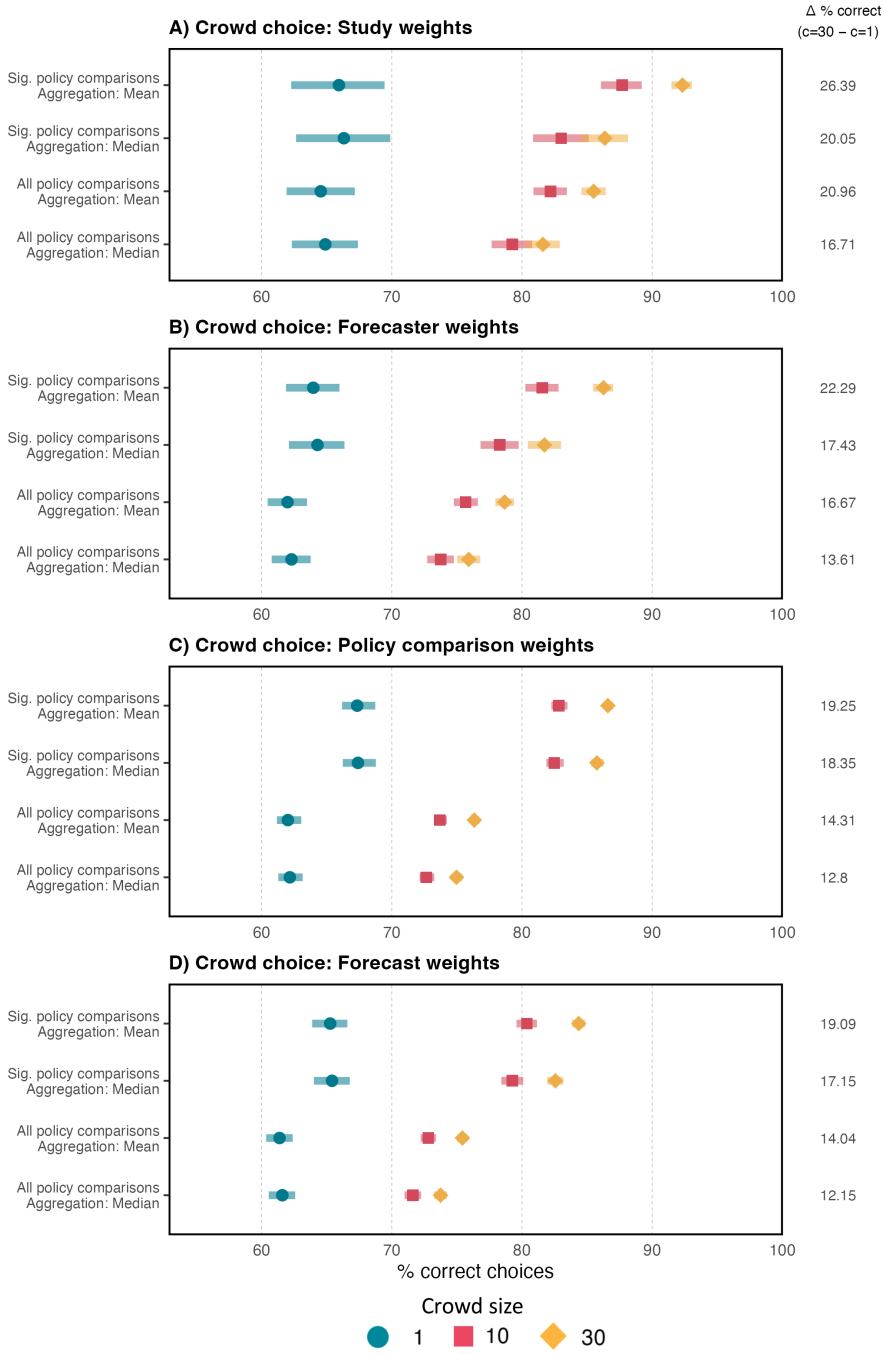
Cols. 1 and 2 present the number of non-control policies and behavioral outcomes in each study. Col. 3 reports the number of policy combinations × the number of outcomes. Col. 4 presents the number of academic experts in each study. Col. 5 reports the number of forecasts (policies × outcomes × forecasters) in each study.

Figure 1: Policy choice by crowd size



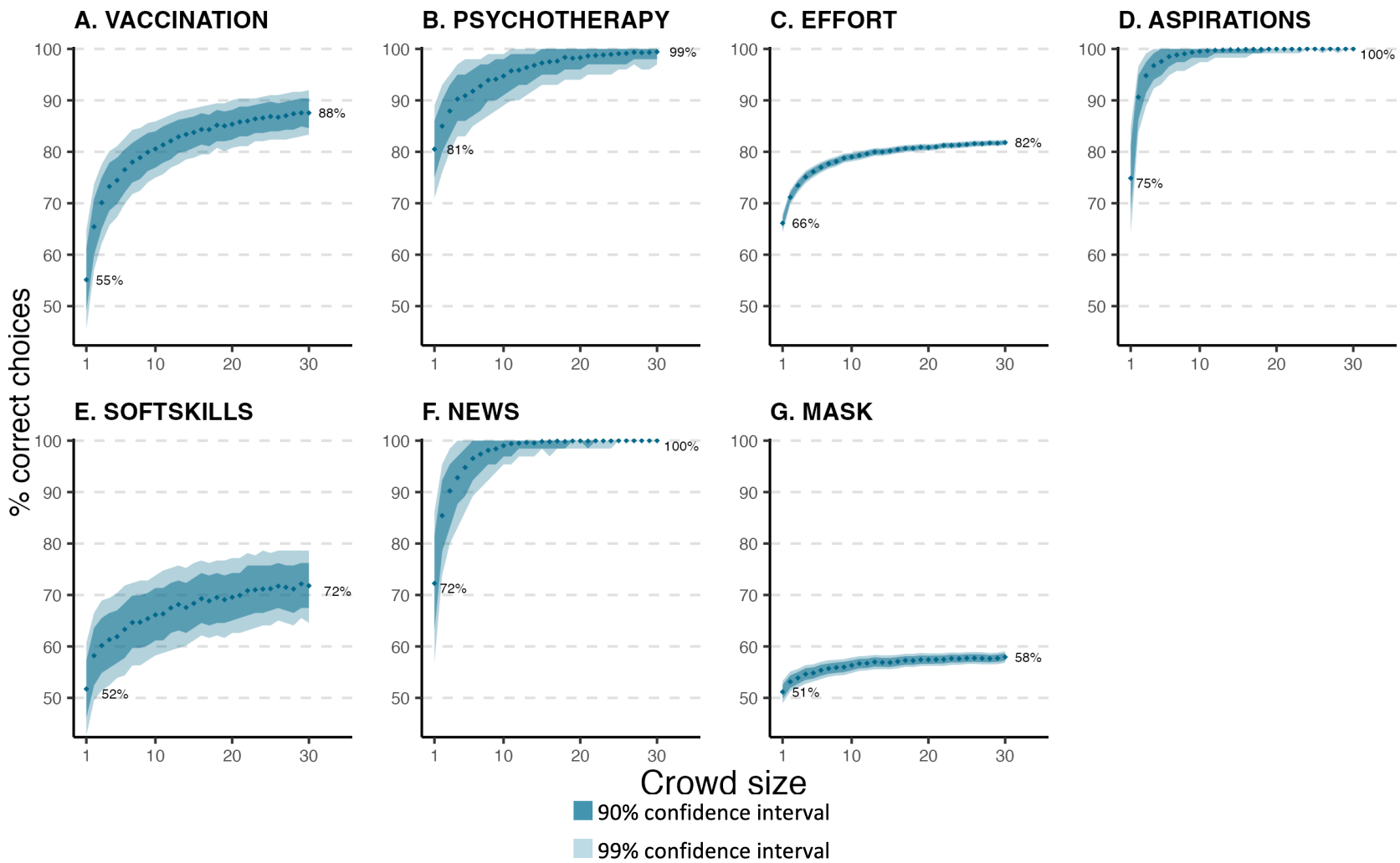
This figure plots the percent of mean crowd forecasts that correctly rank pairs of policies by their experimentally estimated causal effects. For each of the seven studies, crowds are generated by taking 5,000 bootstrapped samples of $c = 1, \ldots, 30$ forecasters. Each of the seven studies is given equal weight in calculating the percent of correct choices, and policy comparisons (a pair of policies and an outcome) are weighted equally within studies. Panel A includes all policies and outcomes. Panel B provides a robustness check excluding policy comparisons that are not significantly different at the $p < 0.10$ level. 90% and 99% bootstrapped confidence intervals are presented using dark and light bands.

Figure 2: Crowd choice robustness checks: alternative weighting and aggregation strategies



This figure plots the percent of crowd forecasts that correctly rank pairs of policies by efficacy for crowds of size 1, 10, and 30. Each panel provides results aggregating forecasts using the mean and median crowd prediction, for the full set of policy comparisons, and for the subset of policy comparisons where one treatment has a significantly larger effect at the $p < 0.10$ level. Panel A weights each study equally and weights policy comparisons (a pair of policies and an outcome) equally within studies. Panel B weights studies by the number of forecasters. Panel C weights studies by the number of policy comparisons. Panel D weights studies by the number of forecasts (the number of forecasts is equal to the number of forecasters $\times$ the number of policies $\times$ the number of outcomes). Crowds are generated by taking 5,000 bootstrapped samples of size c of forecasters for each crowd size and experiment. Bands represent 95% bootstrapped confidence intervals. The column to the right of the figures presents the change in the percent of correct policy choices from crowds of size 30 vs 1.

Figure 3: Policy choice by crowd size and study



This figure plots the percent of mean crowd forecasts that correctly rank pairs of policies by their experimentally estimated causal effects by study. For each of the seven studies, crowds are generated by taking 5,000 bootstrapped samples of $c = 1, \ldots, 30$ forecasters. Policy comparisons (a pair of policies and an outcome) are weighted equally within studies. 90% and 99% bootstrapped confidence intervals are presented using dark and light shading.

APPENDIX

Policy Choice and the Wisdom of Crowds

A Framework and Empirical Strategy

Data structure and notation. Define $\theta_{y(t_e)}$ as the estimated causal effect of policy $t = 1, \ldots, T$ on outcome $y = 1, \ldots, Y$ in experiment $e = 1, \ldots, E$, and $f^i_{y(t_e)}$ as forecaster i's prediction of this effect. Experiment e with Y_e outcomes and T_e policies has $\binom{T_e}{2}$ unique pairs of policies indexed by $\tau_e = 1, \ldots, \mathcal{T}_e$, and $\mathcal{T}_e \times Y_e$ policy comparisons. For example, VACCINATION evaluates 4 policies and 1 outcome, producing $\binom{4}{2} \times 1 = 6$ policy combinations, while PSYCHOTHERAPY has 2 policies and 2 outcomes (4 policy combinations). For a given pair of policies $\tau_e = (t, t')$ and an outcome y, define t as the more effective policy such that $\theta_{y(t_e)} - \theta_{y(t'_e)} \equiv \theta_{y(\tau_e)} > 0$.

Policy choice by individuals. Given that each policy comparison is ordered such that $\theta_{y(\tau_e)} > 0$, forecaster i makes the correct policy choice if $f^i_{y(t_e)} - f^i_{y(t'_e)} \equiv f^i_{y(\tau_e)} > 0$. The average percent of correct policy choices made by N individuals is:

$$\mathrm{PC}^i_{y(\tau_e)} = \frac{100}{N} \sum_{i=1}^{N} \mathbb{1}\Big[f^i_{y(\tau_e)} > 0\Big] \tag{1}$$

where *PC* stands for *Percent Correct* (or *Percent Concordant*).

Policy choice by crowds of independent experts. Define g as a function mapping c forecasters' predictions to a *crowd forecast*. For example, g could be the mean or median crowd forecast. The total number of policy comparisons across all studies is $(\boldsymbol{\mathcal{T}} \cdot \mathbf{Y}) = 161$, where $\boldsymbol{\mathcal{T}}$ is a vector of counts of policy combinations by study (Column 3 in Table 1), $\boldsymbol{\mathcal{Y}}$ is the corresponding vector of counts of outcomes by study (Column 2 in Table 1), and "$\cdot$" is the dot product. If N crowds provide predictions for each study, there are a total of $N(\boldsymbol{\mathcal{T}} \cdot \mathbf{Y})$ crowd policy choices. Finally, define $j = 1, ..., J$ with $J = N(\boldsymbol{\mathcal{T}} \cdot \mathbf{Y})$ as an index of these crowds' policy choices, with $i \in c_j$ denoting the set of forecasters involved in the policy choice. This crowd has made the correct choice if $g\big(f^{i \in c_j}_{y(\tau_e)}\big) > 0$. For N crowds of size

c, the percent of correct policy choices is:

$$\mathrm{PC}^{c}_{y(\tau_e)} = \frac{100}{N} \sum_{j=1}^{J} \pi_j \, \mathbb{1}\Big[g\big(f^{i \in c_j}_{y(\tau_e)} \big) > 0 \Big] \tag{2}$$

Where π_j is the weight placed on a policy comparison from crowd c_j.

Design decisions. Estimation involves three main design decisions, which I use to organize a series of robustness checks.

1. **Forecasting weights**: I consider four weights:
   (a) **Study weights** $\pi_j^{Study} = \big(\mathcal{T}_e Y_e E\big)^{-1}$ which give equal weight to each of the seven studies and within each study give equal weight to each policy comparison (a pair of policies and an outcome).
   (b) **Forecaster weights** $\pi_j^{Forecaster} = I_e \big(\mathcal{T}_e Y_e I_{All}\big)^{-1}$ which weight studies by the number of forecasters where I is the total number of forecasters in each study (I_e) or across studies (I_{All})
   (c) **Forecast weights** $\pi_j^{Forecast} = F_e \big(\mathcal{T}_e Y_e F_{All}\big)^{-1}$ which weights studies by the number of forecasts (the number of forecasts is equal to the number of forecasters $\times$ the number of policies $\times$ the number of outcomes), where F is the total number of forecasts in each study (F_e) or across studies (F_{All}).
   (4) Policy comparison weights $\pi_j^{Policy\ Comparison} = \big(\boldsymbol{\mathcal{T}} \cdot \mathbf{Y}\big)$ which weights studies by the number of policy comparisons.
2. **Choice of function** g which maps the predictions of c forecasters to a single crowd forecast. I examine crowd prediction using both the mean and median forecast, though more sophisticated aggregation strategies (e.g., those that weight forecasters based on performance) could likely improve forecast accuracy (Tetlock and Gardner, 2015; DellaVigna and Pope, 2018a).
3. **Restrictions on policy comparisons**: Some policies perform trivially better than others. I present results restricting analysis to policy comparisons (a pair of policies and an outcome) that are significantly different at the $p < 0.10$ level. This results in some small notational changes in Equation 2 and the weights above. For example, the number of policy comparisons in study e is not necessarily ($\mathcal{T}_e Y_e$), as a pair of policies may be significantly different for some outcomes but not others.

My primary analysis (1) gives each study equal weight, and weights policy comparisons equally within studies, (2) uses the mean as the crowd aggregation function, and (3) does not exclude any policy comparisons.

Generating crowds and measuring accuracy. The accuracy of a single crowd of size c

is calculated by (1) sampling c forecasters with replacement; (2) for each pair of policies and each outcome, calculating the crowd forecast using method g; and (3) calculating whether the crowd correctly forecasts which policy has a larger causal effect for each policy comparison. For each crowd size $c = 1, \ldots, C$ and for each experiment I repeat this procedure 5,000 times. I then calculate the percent of correct policy choices by crowds of size c using Equation 2.

Bootstrapped confidence intervals. Bootstrapped confidence intervals for experiment e and crowd size c are generated by (1) taking draws $5,000$ draws of I_e crowds, where the number of crowds is equal to the total number of unique forecasters I_e in experiment e; (2) for each sample of I_e crowds, calculating the percent of correct choices from the crowds' aggregate forecasts using Equation 2; (3) generating the empirical distribution of correct choices for the $5,000$ samples (for results pooled across studies I combine empirical distributions using the same weights that crowd accuracy was calculated with); (4) calculating the percentile bootstrap confidence intervals from the simulated empirical distribution.

B Screening and Exclusion Criteria

Study screening

1. **Number of interventions.** To allow for policy ranking, each study must include at least two discrete policies (excluding the control condition).
2. **Number of forecasters.** Studies must have a sufficiently large number of forecasters to generate crowd-wisdom. Following previous research (DellaVigna et al., 2020), I require that at least 30 forecasters provide predictions for each pair of policies, and that these forecasters have predicted effects of the same interventions. This is to ensure that I can generate crowds of forecasters who have seen the same interventions and excludes studies like Milkman et al. (2021) and DellaVigna and Linos (2022) that–despite having many forecasters–have relatively few who provided predictions for the same pair of interventions.
3. **Type of forecaster.** I focus on studies where *academic experts* provide predictions of experimental results. This excludes studies collecting predictions from laypeople (Thomas et al., 2020; Chadimová et al., 2022; Otis, 2022), or studies that collect predictions from policymakers (Casey et al., 2021). Because I draw data from a diverse and decentralized set of studies, I have different information on the forecasters in each study. Academic experts are defined broadly as academic researchers, faculty, and graduate students. I also omit publications where only the *authors* provided predic-

tions. Evidence suggests that those directly involved in a study may be overoptimistic about the causal effects of their interventions (Milkman et al., 2022), though a more detailed analysis is beyond the scope of this paper.

4. **Type of forecast.** Forecasters must predict point estimates of the effects of policy interventions. This excludes papers that elicit only the sign of a treatment effect or which ask for predictions in wide bins (Abaluck et al., 2021).

I include to my knowledge every published study meeting these criteria. I complement these studies (which are mostly based in Western countries) with predictions I collected for two large, pre-registered field experiments from Kenya, a lower-income country (Otis, 2021). Note that in Otis (2021) I elicit predictions for a third study whose main intervention varies the saturation of cash transfers in clusters of villages (Egger et al., 2020). Forecasters in this case would only have to predict, for example, whether higher levels of cash transfers increase consumption, (they do). Including this study would lead to even more accurate crowd policy choices.

Exclusion criteria

1. **Outcome exclusion criteria.** For each of the studies, I focus only on predictions of behavioral outcomes (including self-reported behavior) but excluding subjective outcomes like beliefs or intentions. To motivate this decision, Campos-Mercade et al. (2021) measure both intentions to vaccinate (non-behavioral) and actual vaccination and find large differences between vaccination intentions and behavior.
2. **Policy exclusion criteria.** Several studies include cross-randomized *combined* interventions. For example, Orkin et al. (2020) evaluate a combined cash transfer and aspirations intervention and Haushofer et al. (2021) evaluate a combined cash transfer and therapy intervention. I exclude these combined interventions to focus on the more common policy problem of selecting which policy (as opposed to combinations of policies) to implement. I also exclude the control conditions from the policy comparison, as well as a *pure* control group in Campos-Mercade et al. (2021) that does not include a reminder to get vaccinated (all other conditions, including the main control, have a reminder). Including these conditions results in even more accurate policy choices, and I choose to omit them and present more conservative estimates.
3. **Forecaster exclusion criteria.** I exclude any forecaster who is missing predictions for a given outcome to avoid creating imbalance in the number of forecasts among crowds of the same size. Because of this, my analytic sample of forecasters is sometimes smaller than the sample used in the original papers. Additionally, I follow any pre-registered forecaster exclusion criteria specified by study authors. Wisdom-of-crowds effects are

even stronger when these participants are included (intuitively, noisy responses from inattentive respondents leave more space for improvements from aggregation), and I again choose to focus on the more conservative estimate.

C Overview of Studies

MASK (Dimant et al., 2022; Gelfand et al., 2022)

Sample: Participants were drawn from an online panel run by Qualtrics and completed the experiment online.

Outcomes:

1. **Signing.** Digitally signing a pledge to wear a mask.
2. **Sharing.** Sharing this pledge over social media.

Interventions: The interventions from Gelfand et al. (2022) are described verbatim:

1. **Control**. *This was the baseline condition and included the standard message with no additional justification.*
2. **Protection from Harm (Self).** *This condition highlighted the liberal moral value harm as justification for engaging in prevention behaviors.*
3. **Protection from Harm (Community).** *This condition focused on preventing harm to others as the justification for wearing a mask or face covering.*
4. **Patriotic Duty.** *This condition was designed to tap into the moral foundation of ingroup-loyalty at a broader level. i.e., making patriotic sacrifices for one's country.*
5. **Purity.** *This condition employed the conservative moral value purity, which is based in the psychological desire to avoid contamination.*
6. **Reviving the Economy.** *This condition highlighted the importance of following health guidelines for a successful reopening of the economy.*
7. **Threat.** *This condition emphasized the threat that COVID-19 continues to pose to Americans and the severity of the potential consequences of contracting the virus.*
8. **Scientific evidence.** *...this condition emphasized that there is clear scientific evidence showing that masks effectively reduce the spread of the virus...*

Forecasters: Forecasters were recruited through academic mailing lists and from a consortium of social science researchers. Included participants self-classified as academics. 91% held graduate degrees, 43% were economists, 24% were psychologists, and the remaining participants self-classified as "other" or had degrees from multiple fields.

NEWS (Chopra et al., 2022)

Sample: Participants were drawn from an online pool run by Lucid and completed the experiment online.

Outcome: Newsletter sign-up. The proportion of participants who sign up for a weekly newsletter on an economic relief plan.

Interventions:

1. **Control**. Participants are offered weekly newsletters on an economic relief plan.
2. **Fox with fact checking**. Participants are offered weekly newsletters on an economic relief plan with articles *from Fox News.* They are told all articles will be fact checked.
3. **MSNBC with fact checking**. Participants are offered weekly newsletters on an economic relief plan with articles *from MSNBC.* They are told all articles will be fact checked.

Forecasters: Forecasters were leading academic experts who had attended one of several economics conferences. 74% were (full, associate, or assistant) professors, and 24% were PhD students or postdoctoral researchers.

VACCINATION (Campos-Mercade et al., 2021)

Sample: Swedish participants were recruited by the research firm Norstat.

Outcome: Vaccination. Vaccination was measured using administrative records from the Swedish Public Health Agency.

Interventions:

1. **Control**. Participants receive two reminders to get vaccinated.
2. **Monetary incentives**. Participants who get vaccinated are paid $24 (SEK 200) (and receive the control condition reminders).
3. **Social impact**. Participants list people who would benefit from them receiving vaccination (and receive the control condition reminders).
4. **Arguments**. Participants develop arguments to persuade other people to be vaccinated (and receive the control condition reminders).
5. **Information**. Participants are asked to complete a quiz with information on the COVID-19 vaccine (and receive the control condition reminders).

I exclude a *pure* control group that did not receive reminders to get vaccinated.

Forecasters: Forecasters were recruited from the Social Science Prediction Platform, an online platform for collecting predictions of social science results. 81% of forecasters were economists, 30% were (Full or Assistant) professors, 49% were PhD students or postdoctoral researchers, and 21% were other researchers. Two forecasters with incomplete predictions were excluded.

EFFORT (DellaVigna and Pope, 2018a,b)

Sample: Participants were recruited from the online platform Amazon Mechanical Turk and completed the experiment online.

Outcome: Effort. Effort is measured through a real-effort task where participants repeatedly type "A" then "B".

Interventions: The interventions from DellaVigna and Pope (2018a,b) are described verbatim:

1. **Control**. *Your score will not affect your payment in any way.*
2. **Piece rate, 4 cent**. *As a bonus, you will be paid an extra 4 cents for every 100 points that you score.*
3. **Very low pay**. *As a bonus, you will be paid an extra 1 cent for every 1,000 points that you score.*
4. **Red Cross, 1 cent**. *As a bonus, the Red Cross charitable fund will be given 1 cent for every 100 points that you score.*
5. **Red Cross, 10 cents**. *As a bonus, the Red Cross charitable fund will be given 10 cents for every 100 points that you score.*
6. **40 Cent Bonus**. *In appreciation to you for performing this task, you will be paid a bonus of 40 cents. Your score will not affect your payment in any way.*
7. **Discounting: 2 weeks**. *As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account two weeks from today.*
8. **Discounting: 4 weeks**. *As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account four weeks from today.*
9. **40 cent threshold bonus**. *As a bonus, you will be paid an extra 40 cents if you score at least 2,000 points.*
10. **40 cent threshold bonus - loss**. *As a bonus, you will be paid an extra 40 cents. However, you will lose this bonus (it will not be placed in your account) unless you score at least 2,000 points.*
11. **80 cent threshold bonus**. *As a bonus, you will be paid an extra 80 cents if you score at least 2,000 points.*
12. **1% chance of $1**. *As a bonus, you will have a 1% chance of being paid an extra $1 for every 100 points that you score. One out of every 100 participants who perform this task will be randomly chosen to be paid this reward.*
13. **50% chance of 2 cents**. *As a bonus, you will have a 50% chance of being paid an extra 2 cents for every 100 points that you score. One out of two participants who perform this task will be randomly chosen to be paid this reward.*
14. **Social comparisons**. *Your score will not affect your payment in any way. In a previous version of this task, many participants were able to score more than 2,000 points.*
15. **Ranking**. *Your score will not affect your payment in any way. After you play, we will show you how well you did relative to other participants who have previously done this*

task.

16. **Task significance**. *Your score will not affect your payment in any way. We are interested in how fast people choose to press digits and we would like you to do your very best. So please try as hard as you can.*

Forecasters: Forecasters were behavioral experts recruited primarily from conferences and academic research organizations. 87% were economists, 56% were (full, associate, or assistant) professors, 41% were PhD students, and 2% were other researchers. Five forecasters with incomplete predictions were excluded.

PSYCHOTHERAPY (Haushofer et al., 2021)

Sample: Participants are low-income individuals living in rural Kenya.

Outcomes:

1. **Intimate partner violence.** The proportion of women reporting physical violence from their male partners.
2. **Consumption.** The sum of monthly household consumption.

Interventions:

1. **Control**. Control participants received no intervention.
2. **Unconditional cash transfer**. Participants received an unconditional cash transfer of $500.
3. **Five-week psychotherapy treatment**. Participants received five weeks of 90-minute sessions of a mental health intervention called Problem Management Plus.

I exclude a combined intervention that delivers both (2) and (3).

Forecasters: Forecasters were PhD students or academic researchers who had recently published papers on cash transfers or mental health interventions. 86% were economists, and the remaining forecasters were psychologists or epidemiologists. 54% were (full, associate, or assistant) professors, 31% were PhD students, and 15% were other researchers.

ASPIRATIONS (Orkin et al., 2020)

Sample: Participants are low-income individuals living in rural Kenya.

Outcomes:

1. **Assets.** The total value of non-land household assets.
2. **Education expenditure.** The total annual per-child expenditure on education.
3. **Consumption.** Total household consumption expenditure in the last 30 days.

Interventions:

1. **Aspirations and Goal-Setting Intervention**. Participants received an intervention involving short videos starring role models, facilitated exercises, a calendar, and stickers to encourage goal achievement.

2. **Control**. A placebo condition where participants received a psychologically inactive version of intervention (1).
3. **Unconditional Cash Transfer**. Participants received an unconditional cash transfer of $1,100. This group also received the psychologically inactive intervention from (2).

I exclude a combined intervention that delivers both (1) and (3).

Forecasters: Forecasters were PhD students or academic researchers who had recently published papers on cash transfers, aspirations, or goal setting. 97% of forecasters were economists. 48% were (full, associate, or assistant) professors, 38% were PhD students or postdoctoral researchers, and 14% were other researchers.

SOFTSKILLS (Groh et al., 2016)

Sample: Participants were young Jordanian women.

Outcomes:

1. **Short-run employment.** Short-run employment is measured after 6 months.
2. **Long-run employment.** Long-run employment is measured after 18 months.

Interventions:

1. **Control**. Control participants receive no intervention.
2. **Wage subsidies**. Participants receive a wage subsidy voucher worth $210 per month that they could present to firms during their job search.
3. **Soft-skills training**. Participants were invited to receive a free 45-hour training on interpersonal skills.

Forecasters: Forecasts were collected during academic presentations and on a popular academic blog. Six forecasters with incomplete predictions were excluded.

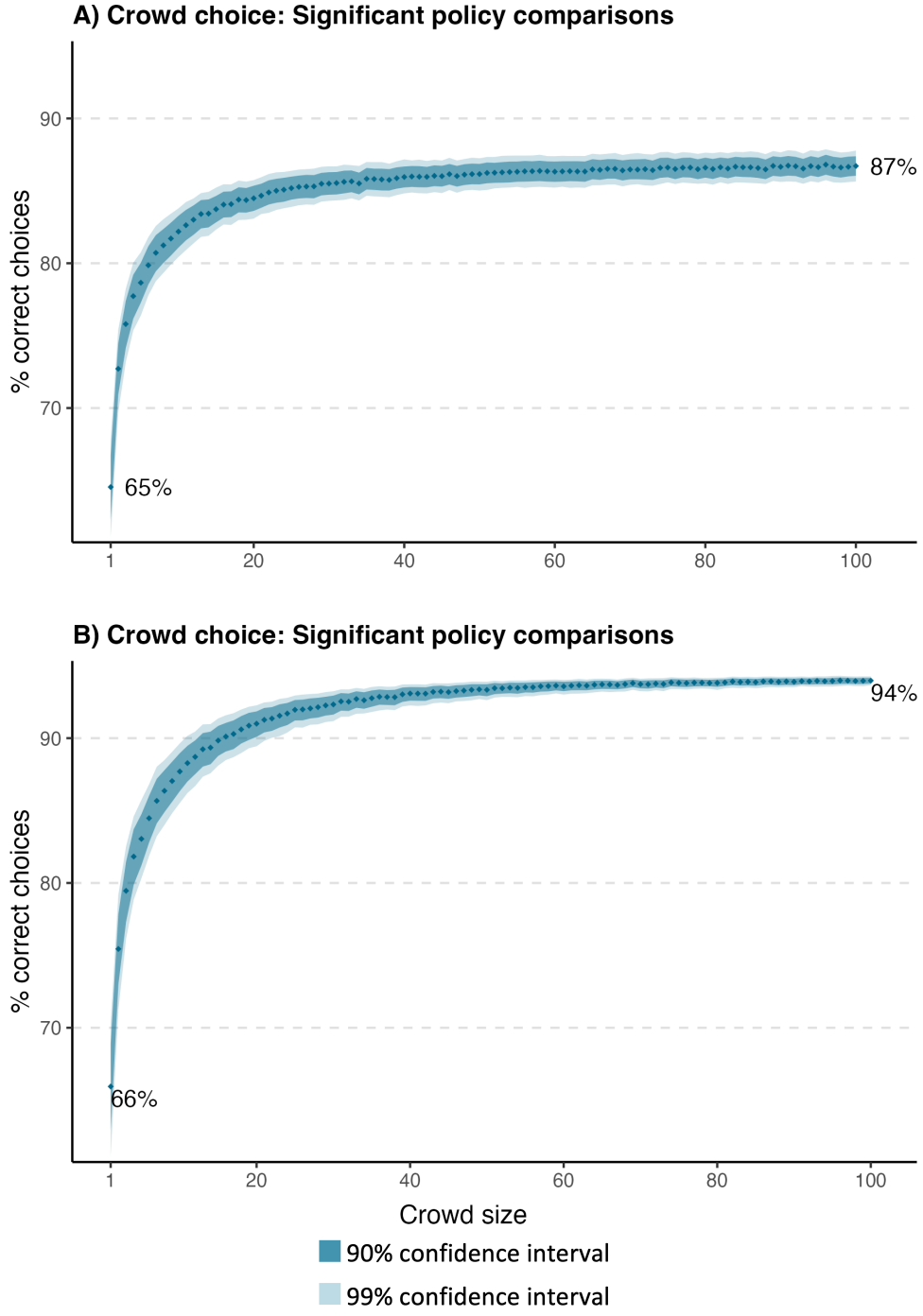
D Additional Tables and Figures

Table A1: Comparison of experimental features (significant policy comparisons)

Experiment	Number of				
	Policies	Outcomes	Policy comparisons	Forecasters	Forecasts
	(1)	(2)	(3)	(4)	(5)
ASPIRATIONS	2	1	2	39	156
EFFORT	15	1	105	355	5325
MASK	5	2	5	199	1194
NEWS	2	1	1	65	130
PSYCHOTHERAPY	2	2	2	50	200
SOFTSKILLS	2	1	1	103	206
VACCINATION	3	1	2	52	156
Total	31	9	118	863	7367

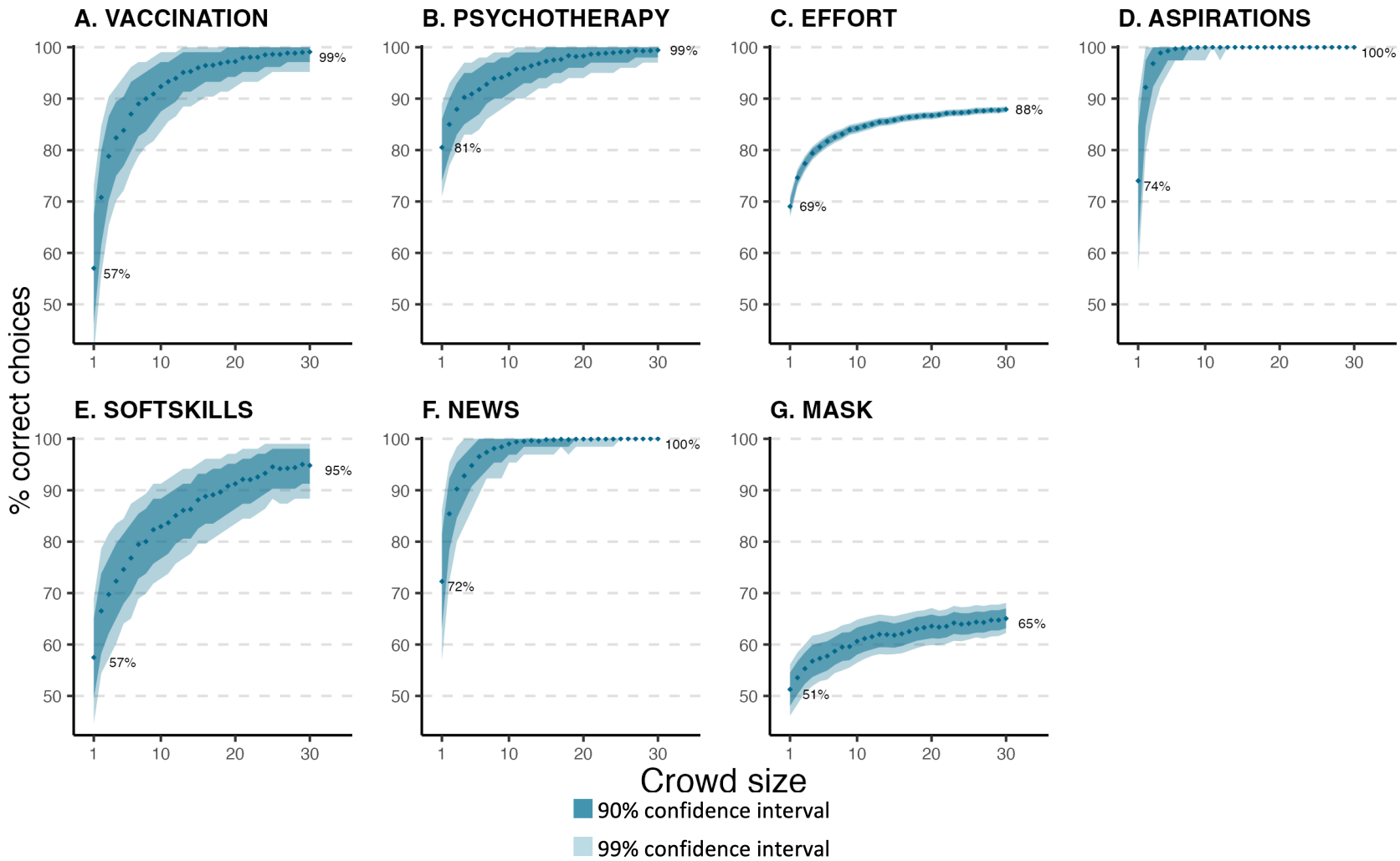
This table presents experimental features excluding policy comparisons that are not significant at the $p < 0.10$ level. Cols. 1 and 2 present the number of non-control policies and behavioral outcomes in each study. Col. 3 reports the number of policy combinations $\times$ the number of outcomes. Col. 4 presents the number of academic experts in each study. Col. 5 reports the number of forecasts (policies $\times$ outcomes $\times$ forecasters) in each study.

Figure A1: Policy choice by crowd size (larger crowds)



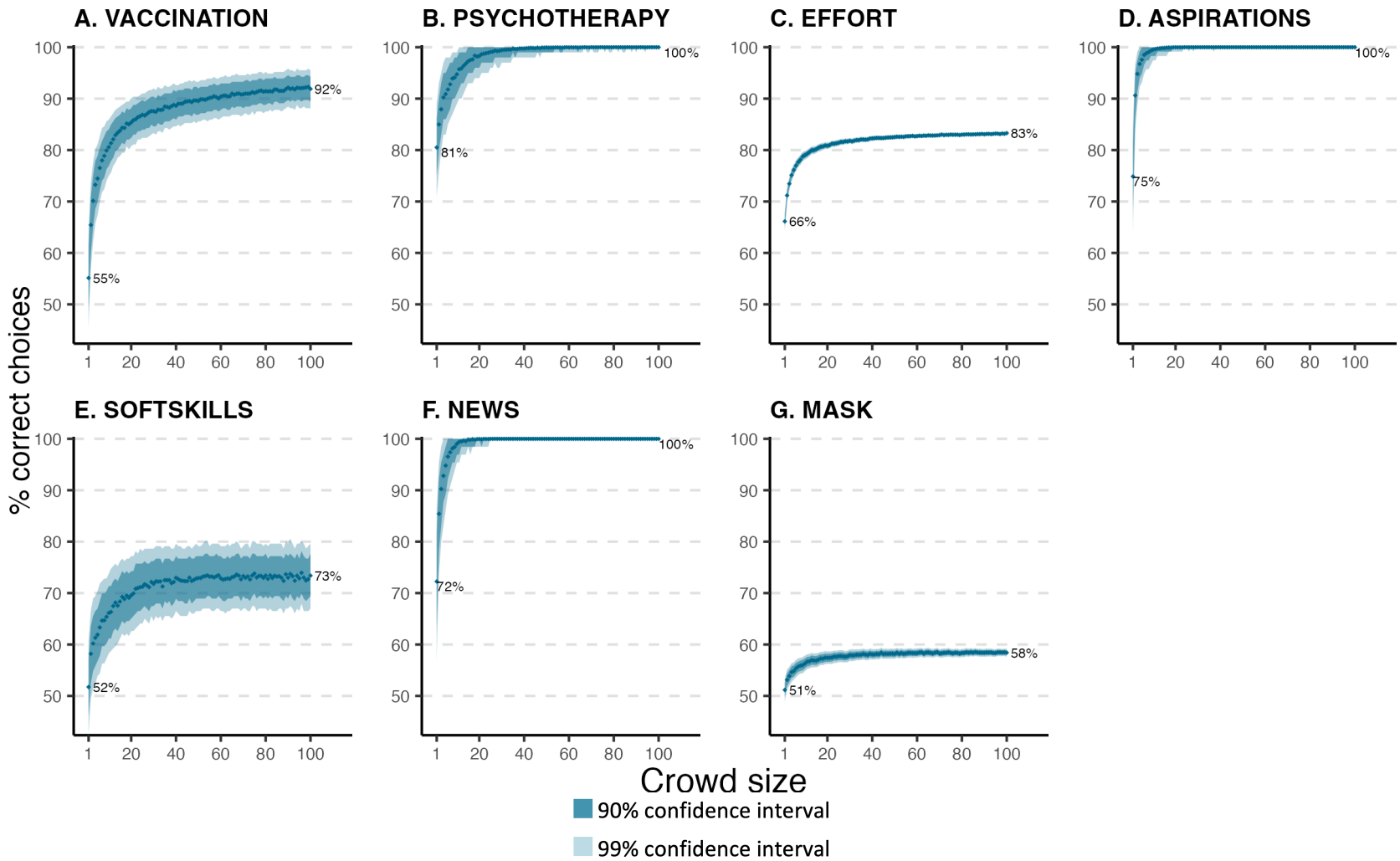
This figure plots the percent of mean crowd forecasts that correctly rank pairs of policies by their experimentally estimated causal effects. For each of the seven studies, crowds are generated by taking 5,000 bootstrapped samples of $c = 1, \ldots, 100$ forecasters. Each of the seven studies is given equal weight in calculating the percent of correct choices, and policy comparisons (a pair of policies and an outcome) are weighted equally within studies. Panel A includes all policies and outcomes. Panel B provides a robustness check excluding policy comparisons that are not significantly different at the $p < 0.10$ level. 90% and 99% bootstrapped confidence intervals are presented using dark and light bands.

Figure A2: Policy choice by crowd size and study (significant policy comparisons)



This figure plots the percent of mean crowd forecasts that correctly rank pairs of policies by their experimentally estimated causal effects by study and restricting analysis to policy comparisons that are significantly different at the $p < 0.10$ level. For each of the seven studies, crowds are generated by taking 5,000 bootstrapped samples of $c = 1, \ldots, 30$ forecasters. Policy comparisons (a pair of policies and an outcome) are weighted equally within studies. Panel A includes all policies and outcomes. Panel B provides a robustness check excluding policy comparisons that are not significantly different at the $p < 0.10$ level. 90% and 99% bootstrapped confidence intervals are presented using dark and light shading.

Figure A3: Policy choice by crowd size and study (larger crowds)



This figure plots the percent of mean crowd forecasts that correctly rank pairs of policies by their experimentally estimated causal effects by study. For each of the seven studies, crowds are generated by taking 5,000 bootstrapped samples of $c = 1, \ldots, 100$ forecasters. Policy comparisons (a pair of policies and an outcome) are weighted equally within studies. 90% and 99% bootstrapped confidence intervals are presented using dark and light shading.