Evaluating Managerial Expectations

Nicholas G. Otis Mathijs de Vaan*

UC Berkeley

Abstract

Managers play a crucial role in shaping firm performance. However, we know little about how managers perform on one of their key responsibilities: managing strategic decisions. The challenge of studying the quality of these decisions is that we often only observe the outcome associated with the chosen decision, not the counterfactual. We address this challenge using data from six recent studies that elicited predictions from expert managers about the direction and size of strategic interventions in domains such as hiring, workplace incentives, layoffs, marketing, and fundraising, and then estimated the causal effects of these interventions in large randomized controlled trials. We evaluate managerial expectations by comparing predicted to experimentally estimated intervention effects under a unified set of performance metrics. Our findings reveal that managers often struggle to accurately predict the causal effects of common strategic interventions. First, when comparing interventions, managers are only slightly better than chance at determining which intervention will have a greater impact. Second, their estimates of an intervention's magnitude often deviate substantially from the true impact. Finally, despite low levels of average accuracy, some managers perform better than others, and better performing managers can be consistently identified using signals of revealed ability. These findings add to the body of research on the impact of managers on firms and demonstrate that organizational experimentation can reveal not only which interventions are effective but also who has knowledge of what works, which can be used to inform the allocation of decision-making authority within a firm.

*Otis: notis@berkeley.edu. de Vaan: mdevaan@berkeley.edu. The authors would like to thank Stefano Caria, Solène Delecourt, Stefano DellaVigna, Douglas Guilbeault, Ambar La Forgia, and Joseph Reiff for helpful comments.

Introduction

A growing literature details the significant influence that managers have on the firms they work for (Bertrand and Schoar, 2003; Bloom and Van Reenen, 2007; Demerjian et al., 2012; Hoffman and Tadelis, 2021; Camuffo et al., 2020, 2021). Managers play a critical role in high-stakes decisions related to managing human capital (Abebe et al., 2021; Del Carpio and Guadalupe, 2022; Heinz et al., 2020; Friebel et al., 2022), fundraising (Samek and Longfield, 2023; Adena and Huck, 2020; Rau et al., 2022), marketing (McKenzie et al., 2023), and directing organizational experimentation more broadly (Sorenson, 2003; Levinthal, 2021; Camuffo et al., 2022; Koning et al., 2022). In making such decisions, managers typically must choose between multiple alternatives – which we refer to as *strategic interventions* – and determine how much to invest in the chosen option. The consequences of these decisions can have a far-reaching impact on the direction, performance, and success of the firm.

Despite growing interest in theorizing and measuring how managers affect firm performance, there is little research that examines whether the decisions that managers make are the ones that are most likely to benefit the firm and whether managers differ in their ability to anticipate the consequences of the interventions under consideration (Csaszar and Laureiro-Martínez, 2018; Ryall and Sorenson, 2022; Kapoor and Wilde, 2022). The challenge in evaluating managerial performance in making strategic decisions is that it is difficult to isolate the effects of a decision from confounding factors and establish a credible counterfactual. Finding a good instrument for the intervention the manager ends up pursuing is often challenging or impossible. Moreover, (quasi-)random assignment of interventions across managers makes it difficult to identify performance variations across managers, which, if present, could be used to allocate decision-making authority within a firm.

In this paper we address these challenges by leveraging data from six recent randomized controlled trials (RCTs) evaluating strategic interventions that were recently conducted by other researchers. In each study, expert managers familiar with the experimental setting and intervention domain provided forecasts of the causal effects of the interventions. These data provide two key pieces of information – the causal effects of multiple strategic interventions and a manager's estimates of the impacts of the interventions – that allow us to estimate (1)

the prediction errors of a large sample of managers in a range of decision-making domains and (2) the extent to which prediction quality varies between managers. Jointly, these statistics allow us to characterize a manager's ability to accurately anticipate the consequences of strategic interventions.

The six studies in our sample address diverse managerial domains: hiring (Abebe et al., 2021), workplace incentives (DellaVigna and Pope, 2018b), layoffs (Heinz et al., 2020), customer feedback (Reiff et al., 2021), and fundraising (Samek and Longfield, 2023; Rau et al., 2022). In each study, a separate sample of expert managers familiar with the substantive focus and experimental context predicted the causal effects of different strategic interventions. For instance, Abebe et al. (2021) assessed the impact of application incentives and wage increases on applicant quality and application rates and solicited predictions from HR department heads and CEOs of Ethiopian firms currently hiring for similar positions. Another study by Rau et al. (2022) evaluated different door-to-door fundraising strategies, eliciting predictions from fundraising heads of leading charities. Overall, our data include nearly 700 managers who provide over 4,500 predictions (3-15 per manager). In five of the six studies, managers received financial incentives for prediction accuracy, analogous to performance-based bonuses in managerial roles.

In our analysis, we pool findings from the six studies using a cohesive set of managerial performance metrics. We find that managers forecasting the impact of a range of strategic interventions perform only slightly better than chance at identifying which strategic interventions will be more effective: across all six experiments in our sample, managers on average fail to identify which of two interventions will have a larger causal effect 42% of the time. Furthermore, managers' impact estimates for individual interventions often significantly deviate from the true effects. On average, the prediction errors are over 2.5 times larger than the intervention effects themselves. Finally, we demonstrate that there are "good" and "bad" managers: randomly selecting one of their forecasts allows us to predict their error in other forecasts. These correlated errors imply that our results are not driven by noise in either the experimental estimates or managers' forecasts and underscore the value of organizational experimentation to identify which managers have consistently accurate beliefs.

This work contributes to a growing literature on the impact of managers on firm performance. While much of this research broadly examines variation in manager quality (Bloom and Van Reenen, 2007, 2010), some work has started to scrutinize the specific tasks that managers are typically responsible for. For example, Hoffman and Tadelis (2021) find that managers with superior people management skills reduce employee turnover and have higher promotion rates and larger salary increases. Likewise, Liebscher and Mählmann (2017) find that high-performing mentors positively influence subordinates' early promotions, with promotion probabilities rising as mentorship duration increases. Recent studies have also begun exploring manager's skills in making specific decisions. For example, Camuffo et al. (2020, 2021) show that adopting a scientific approach to management improves managers' ability to identify and pursue projects with false negative returns as opposed to false positive returns. Our work complements these findings by providing empirical evidence on a crucial, yet unexplored aspect of strategic management: managers' capacity to anticipate the outcomes of different strategic interventions. In doing so, we contribute to the growing literature on the role of managers in shaping firm outcomes and respond to recent calls to bring managers back into management research (Aguinis et al., 2022).

Overview of data

In this paper, we use data from six recent studies that share three critical characteristics: (1) they include a randomized controlled trial (RCT), (2) they examine one or more strategic interventions relevant to managerial decision-making, and (3) they involve a sample of managers with contextual knowledge who predict the interventions' impacts.

Sample of randomized experiments and managers. The six studies in our sample cover a broad range of settings, countries, and managerial domains.[1] For each study, predictions were collected from a separate group of managers who had, on average, considerable experience with the study context and were responsible for making decisions in the domains of the strategic interventions they were asked to predict.[2] We briefly review the six samples

[1] Appendix Section C describes our study sampling procedure.

[2] In the majority of studies in our sample (Abebe et al., 2021; Samek and Longfield, 2023; Rau et al.,

of managers below.

Abebe et al. (2021) investigated how application incentives and increased wages affect application rates and applicant quality for clerical positions in Ethiopia. Predictions were obtained from Ethiopian firms actively hiring clerical workers from their head of HR or CEO. Heinz et al. (2020) tested how different forms of layoffs influence employee motivation at a call center in Germany and collected forecasts from experienced HR professionals at medium-to-large-sized German companies. Rau et al. (2022) evaluated different fundraising strategies' impact on charitable donations in the United States and elicit predictions from nonprofit managers responsible for fundraising programs. Reiff et al. (2021) evaluate the effects of four different types of customer feedback solicitations. Marketing experts recruited from a major professional organization predicted the effects of these interventions. Samek and Longfield (2023) test the effects of thank you calls on charitable donations in the U.S. Forecasts were provided by the heads of fundraising at large charities in the United States.

The last study in our sample (DellaVigna and Pope, 2018a) tested the impact of different incentives on performance on a simple clerical task performed online. Unlike the previous five studies, the forecasters in this study were full-time MBA students. While they were not managers when they were making the predictions, many MBAs hold managerial positions before starting their degree and will hold such roles after completing their program.

We summarize the features of the studies in our sample in Table 1. In all but one study (Reiff et al. (2021)), incentives based on prediction accuracy were supplied to managers (Column 3 provides details). The median study offered up to $100, and two offered over $300. These accuracy incentives help ensure that the managers provide thoughtful predictions.

2022; DellaVigna and Pope, 2018a; Reiff et al., 2021), managers are not asked to provide predictions about the effects of an intervention in a specific firm but rather the effects for a typical firm.

Table 1: Overview of experiments

Title	Managers	Accuracy incentives (max)	Sample size		
			Managers	Forecasts/ manager	Total forecasts
(1)	(2)	(3)	(4)	(5)	(6)
Measuring the Indirect Effects of Adverse Employer Behaviour on Worker Productivity (Heinz et al., 2020)	Managers from a professional German HR organization and HR magazine readership.	€30	43	3	129
The Selection of Talent: Experimental and Structural Evidence from Ethiopia (Abebe et al., 2021)	Firm hiring managers in the same industry who are actively hiring.	About $335 (PPP adjusted)	195	3	585
When Impact Appeals Backfire (Reiff et al., 2021)	Managers from the Customer Experience Professionals Association.	None	42	4	168
Thank-You Calls Increase Charitable Giving?(Samek and Longfield, 2023)	Fundraising managers who oversaw their charity's fundraising program.	$100	141	6	846
Do I Care if You Are Paid? Field Experiments and Expert Forecasts in Charitable Giving (Rau et al., 2022)	Fundraising experts and managers who oversaw their charity's fundraising program.	$100	100	4	400
What Motivates Effort? Evidence and Expert Forecasts (DellaVigna and Pope, 2018a)	MBAs from Booth and Haas.	$1000	160	15	2400
Total			681	35	4528

Notes: This table summarizes key details of the six studies in our sample, including study title (Col. 1), the sample of managers (Col. 2), accuracy incentives (Col. 3), and the number of managers, forecasts per manager, and total forecasts (Cols. 4-6) in each study.

Unlike the six original studies which mainly used the average prediction across all managers as a benchmark for their experimental estimates, we use the experimental estimates to measure the performance of individual managers. Specifically, we introduce a consistent set of performance measures that can be applied across the heterogeneous studies, resulting in reliable and transferable evidence on managerial accuracy. Table A1 provides an overview of how our analyses differ from the six papers whose data we use.

Measurement and Analytic Strategy

Measurement

This section presents a framework for identifying managerial performance by pairing managers' predictions of intervention impact with estimates of the intervention's causal effects. Specifically, we operationalize managerial performance by constructing error measures that capture the level of alignment between the predicted and estimated effects.

Rank error. First, managers must decide which intervention is likely to be most effective. For example, a manager may have to decide which contract to offer or which hiring strategy to pursue. Define $\mathbb{E}[Y|a]$ as the expected level of outcome Y given that a firm pursues intervention a, and $\widehat{\mathbb{E}}[Y|a]$ as the sample analogue. For example, this could be productivity (Y) under different contracts (a) or applicant quality (Y) under different hiring procedures (a). Define $f_i(Y|a)$ as manager i's prediction of this effect. Given any two interventions a and a', we say that manager i has made the correct decision if $f_i(Y|a) > f_i(Y|a')$ and $\widehat{\mathbb{E}}[Y|a] > \widehat{\mathbb{E}}[Y|a']$. In other words, a manager makes the correct *rank* decision when she correctly anticipates which intervention will have a larger effect. Formally, for each pair of interventions a and a' where the estimated experimental results indicate that $\mathbb{E}[Y|a] > \mathbb{E}[Y|a']$, we calculate:

$$Rank\ Error_i = \Big(1 - \mathbb{1}\big[f_i(Y|a) > f_i(Y|a')\big]\Big) \times 100 \tag{1}$$

where $\mathbb{1}\big[f_i(Y|a) > f_i(Y|a')\big]$ is an indicator function taking a value of 1 if the manager i correctly predicts that a will be more effective than a'. We subtract this value from 1 and multiply the resulting term by 100 to create a rank error measure that gives us the percent of managers who incorrectly predict which intervention will be more effective.

Point error. Our second measure of managerial performance involves assessing a manager's ability to accurately anticipate the *magnitude* of an intervention's effect. Correctly estimating the magnitude of the effect of an intervention is especially important if the allocation of resources depends on effect size. For instance, a manager who expects large productivity

improvements from layoffs may lay off more people than a manager who predicts only a minimal impact. We measure the magnitude of managerial error as the absolute difference between a manager's prediction and the actual experimental effect:

$$Absolute\ Error_i = \big|f_i(Y|a) - \mathbb{E}[Y|a]\big|/\sigma_Y, \tag{2}$$

where we substitute the sample analogue of the conditional expectation using estimates from the randomized experiments. We divide each prediction by the standard deviation (s.d.) of the experimental control group (σ_y) to measure accuracy in the same units in each study.

Combining causal estimates and manager forecasts. A key challenge to evaluating managerial performance in decision making comes from the fact that managerial decisions are generally endogenous, leading to classic omitted variable concerns; a firm laying off workers may simultaneously implement other cost-cutting strategies, and a firm adopting one hiring strategy may allocate fewer resources to alternative approaches. The consequence of this endogeneity for assessments of managerial performance is that the benchmark against which managers are evaluated can easily be biased. The crucial feature of our analytic strategy is that we observe unbiased estimates of $\mathbb{E}[Y|a]$ because our benchmark experimental results are based on random assignment of different interventions (Rubin, 1974). This provides a causal benchmark that managerial expectations can be evaluated against and yields a simple and transparent strategy for estimating managerial performance for both "rank" and "point" expectations.

Analytic Strategy

Building on our two measures of managerial performance, we conduct two main types of analyses. First, we estimate average managerial performance by calculating both the percentage of managers making the wrong choice and the average absolute error across managers, weighting each experiment equally. We also provide disaggregated results for each of the six studies.

Our second set of analyses tests whether "experts" who consistently make better choices

can be reliably identified. This clarifies whether our results are simply due to noise. If poor managerial performance is a result of noise (e.g., by calculating error relative to estimates from under-powered experiments), we would expect uncorrelated errors across managers. However, if there is variation in managerial expertise, we would expect a manager's error for predicting one intervention to correlate with their error for predicting the effect of another intervention. Furthermore, the ability to identify "good" and "bad" managers (as defined by their ability to correctly assess the causal effects of different interventions), highlights the value of combining organizational experimentation and managerial forecasts: not only can organizations learn "what works," but also "who knows what works" which may be used to allocate decision-making authority within a firm (DellaVigna and Pope, 2018a).

Our analytic strategy for this second question leverages the fact that we observe multiple predictions from each manager, meaning we can assess a manager's performance conditional on observing a signal of their ability. As an example, say we have two observations of managerial accuracy e_1 and e_2 for each manager (these could be point- or rank-based measures). To test whether we can recover managerial ability, imagine that we first observe a signal of managerial performance (error) e_1. For expositional purposes, assume that error e_1 is categorized using a binary split such that $d(e_1) = 1$ indicates a manager has performed well and $d(e_1) = 0$ indicates poor performance. We can then calculate the average level of e_2 conditional on $d(e_1)$. If e_2 does not vary by $d(e_1)$, it suggests that there is no reliable pattern of errors among managers. However, if e_2 varies with $d(e_1)$ it would indicate that managerial performance is correlated across interventions and that a single signal of managerial ability can help to separate accurate from inaccurate managers. Differences in our estimation strategy for conditional-rank and conditional point-based error are described in the next section.

Results

Errors in Managerial Forecasts

The first set of results provide point estimates of our two managerial performance measures: *Rank Error* which measures the percent of managers who fail to identify which of two interventions will be most effective and *Absolute Error* which measures the absolute difference (in s.d) between managers' predictions and the observed effects of interventions. Figure 1 displays the percent of managers who incorrectly predict which of two interventions will be more effective, excluding comparisons where the difference in experimental effects is less than 0.01 s.d. Figure A2 provides robustness checks testing different exclusion criteria and weak as opposed to strict rankings of interventions to measure accuracy.

Figure 1: Managerial rank errors

Notes: This figure displays the percent of managers who failed to strictly rank which intervention would be more effective based on randomized experimental findings. Experimental comparisons with effect differences less than 0.01 s.d. are excluded. Figure A2 provides robustness checks. The first row presents the average error across studies, giving each study equal weight, and the next six display study-level results. Error bars present 95% confidence intervals clustered at the manager level.

Across all six studies we observe that managers on average perform only slightly better than chance, identifying the better performing policy only 58.22% of the time (95% confidence interval (CI) = $[55.40, 61.05]$).[3] In other words, if one asks ten managers to predict which

[3]This result is not driven by managers predicting trivial differences between interventions. Figure A3

of two interventions will be more effective, four of them will give the wrong answer.[4] These results mask considerable variation across studies. For example, in the "fundraising calls" study, only 8.51% of managers fail to predict that calls increase donations. Viewed in isolation, this error measure indicates that managers in this particular study are making accurate judgements. The performance of their point predictions, however, shows that this is not the case.

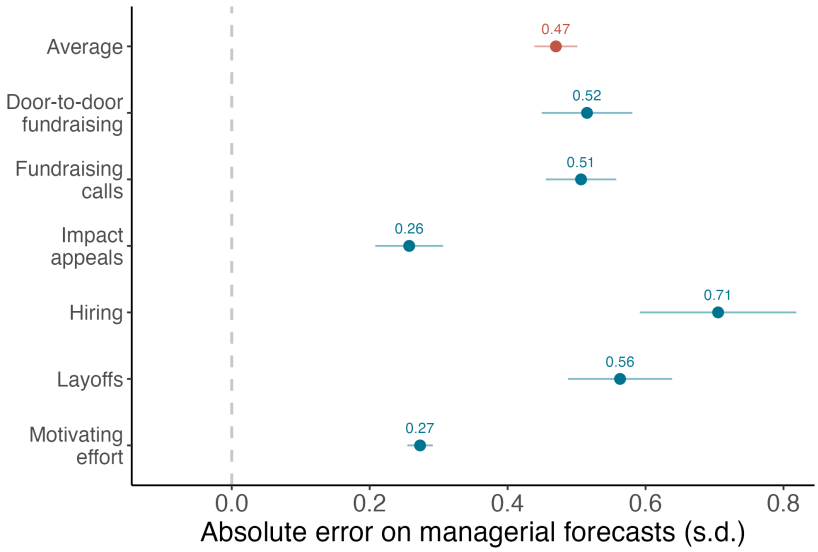
Figure 2 presents the extent to which managers have accurate beliefs about the magnitude of different interventions. In the fundraising study, managers' predictions deviate from the experimental results by 0.51 s.d. To put these magnitudes into perspective, the average experimental effect in the "Fundraising Calls" study is 0.02 s.d., meaning that beliefs deviate from the experimentally estimated results by a factor of 25.[5] Looking across all comparisons we see that the average absolute error is 0.47 s.d., (95% CI = $[0.44, 0.50]$), which is 2.57 times as large as the average effect magnitude.

shows that the average predicted effect size is large, 0.59 s.d., and less than 10% of managers predict an effect below 0.05 s.d. Furthermore, we observe little variation in ability to rank interventions correctly across a range of predicted effect sizes, with the average percent of incorrect rankings ranging from 42– 43% for modest to large predicted effect sizes (>0.05 to 0.30 s.d.).

[4]In the RCTs included in our paper, managers were asked to make predictions individually. When making decisions collectively accuracy is likely to be impacted. See, for example, Almaatouq et al. (2020).

[5]Note that managers' predictions of the conditional mean $\mathbb{E}[Y|a]$ will, in most cases, be equivalent to their prediction of the causal effect of an intervention relative to a status quo "control" condition. Also note that the average treatment effect of an intervention a relative to control a_c is $\mathbb{E}[Y|a] - \mathbb{E}[Y|a_c]$. Managers who are provided with an estimate of the control mean $\mathbb{E}[Y|a_c]$ (which they are in five of the six studies in our sample) provide a prediction of the causal effect of intervention a of $f(Y|a) - \mathbb{E}[Y|a_c]$. Their forecast error is thus $\big(f(Y|a) - \mathbb{E}[Y|a_c]\big) - \big(\mathbb{E}[Y|a] - \mathbb{E}[Y|a_c]\big) = f(Y|a) - \mathbb{E}[Y|a]$. An exception is the "Layoffs" study, where managers are provided with a pre-trend for the control and treatment groups and then predict the treatment and control outcomes in the post-treatment period.

Figure 2: Managerial point errors



Notes: This figure depicts the average absolute distance between managers' forecasts and observed experimental results measured in standard deviations. The first row presents the average error across studies, weighting each study equally. The next six rows display the average error across all outcomes within each study, weighting each forecast equally. Error bars depict 95% confidence intervals with standard errors clustered at the manager level.

Additional results presented in Appendix D examine both the direction and magnitude of error for the set of interventions managers anticipate will be most effective overall. We find that managers overestimate the effects of their most preferred interventions by 0.34 s.d. (95%CI = $[0.28, 0.40]$), with 80.79% of managers predicting that interventions will have larger effects than they actually do.

Conditional Managerial Performance

The results presented so far show that managers consistently fail to identify which interventions will be more effective, and how effective different interventions will be. In this section, we test whether consistently accurate "experts" can be identified.

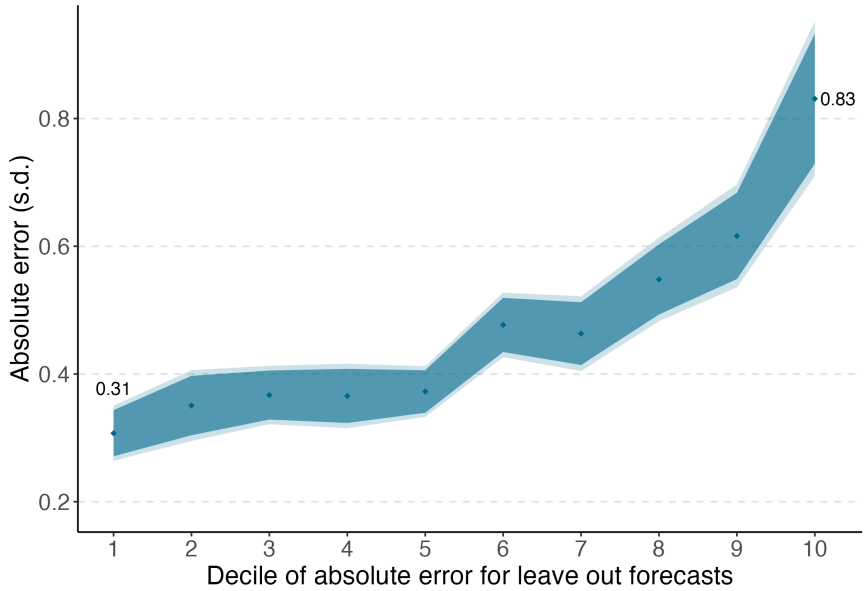
Conditional point decisions. When looking at errors for point predictions, consider a study where managers each provide J predictions. We start by omitting prediction $j = 1$, and then calculate the average absolute error on this prediction for each manager. Based on a comparison with other managers in the study, we give the prediction a decile rank which puts the most accurate managers in the lowest decile. We then calculate the average absolute

error (in s.d.) for the other $J-1$ predicted effects conditional on their decile of performance on $j = 1$. We then rotate through the other $j = 2, 3, \ldots, J$ predictions, omitting each one and calculating the conditional absolute point error. For each study we then calculate the average conditional absolute error across all rotations.

Conditional rank decisions. For rank error measures we instead condition on whether a manager correctly identified which of two interventions would have a larger effect. This conditional measure imposes additional requirements on our data. Consider a study with three conditions (a_1, a_2, a_3). We are not able to create two unique rankings of interventions, because any two rankings would end up re-using the same prediction. This restricts our analysis to four of the six studies with sufficient unique rankings. Our main empirical specification conditions on two arbitrarily selected rankings, but our results are consistent with a wide range of robustness checks in the appendix.

We start by looking at a manager's ability to predict the causal effects of an intervention conditional on their error on an arbitrary "leave out" forecast, which we assigned a decile rank (the highest decile has the largest absolute error). Figure 3 shows that there is a strong correlation in absolute error across managers. Pooling results across studies and giving each study equal weight, we find that those within the top decile for the omitted prediction in each study have an average absolute error of 0.83 s.d. (95% CI = $[0.71, 0.96]$) compared to just 0.31 s.d. for the bottom decile (95% CI = $[0.26, 0.35]$). Appendix Figure A4 presents study-level results which reveal a consistent improvement in all but one of the studies. Appendix Figure A5 provides robustness checks varying the number of "leave out" predictions used when calculating conditional absolute error. It also provides an "out of sample" test following (DellaVigna and Pope, 2018a) using a k-fold procedure to avoid overfitting to our data. Our most conservative estimate of the difference between the top and bottom decile is 0.30 s.d., which is 64% larger than the average magnitude of the effects across all our experiments, while our largest estimated difference is 0.67 s.d. This implies that firms allocating resources based on the perceived efficacy of an intervention could make considerably better allocations if they placed additional weight on managers who had a track record of accurate expectations.

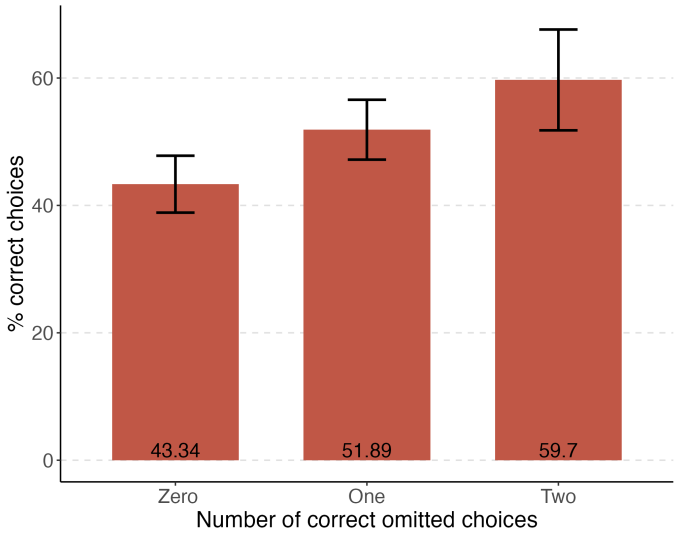
Figure 3: Conditional point error



Notes: This figure presents the average absolute error on managers' forecasts (y-axis) conditional on managers' decile of absolute error on one omitted prediction (x-axis). Conditional absolute error is generated by creating all permutations of two effects in each study, denoted generically as j and j'. For each pair, we regress the absolute error for j on the decile of absolute error for j' across managers. We repeat this procedure across all pairs of effects in each study, and then calculate the average conditional absolute error across studies giving each study equal weight. Light and dark bars present 90 and 95% confidence intervals with standard errors clustered at the manager level.

Turning to our rank-based error measure, we now look at the probability that a manager correctly identifies which intervention will be more effective conditional on their ability to correctly rank other interventions. Because we condition on two omitted choices, we observe performance based on whether they provided the correct ranking 0, 1, or 2 times across their two omitted choices. Comparisons where the difference in experimental effects is less than 0.01 s.d are excluded, as in the previous analysis. Figure 4 presents the findings pooled across the four studies that satisfy the stricter data requirements for this analysis.

Figure 4: Conditional rank decisions



Notes: This figure displays the percentage of managers who incorrectly rank which intervention will be more effective based on their performance on two omitted leave-out choices. Rank-error is measured using causal estimates from the randomized experiments. Conditional error is calculated by omitting each pair of managerial choice and rotating through all pairs of predicted effects in each study. Experimental comparisons with effect differences less than 0.01 s.d. are excluded, as are comparisons where the same forecast is used in the conditioning variables and outcome. Figure A7 provides robustness checks. Results are pooled across studies giving each study equal weight. For the "Effort" study, the analysis is based on a random sample of 5,000 out of over 750,000 combinations of conditional rank-variables. Standard errors are clustered at the manager level.

Managers who provided two correct rankings identified the better performing intervention 59.7% of the time (95% CI = $[51.79, 67.62]$), compared to just 43.3% percent for those making two incorrect decisions (95%CI = $[38.87, 47.80]$). Figure A6 presents results at the study level, and Figure A7 provides a series of robustness checks including whether strict or weak inequalities are used to evaluate whether managers rank interventions correctly, and the minimum difference in experimental effects for a pair of strategic interventions to be included in our analytic sample. Across these robustness checks, managers who rank two interventions correctly are 16-17 percentage points more likely to rank another arbitrary pair of interventions correctly.

When combined with our conditional point-accuracy results, these findings underscore the value of performance signals in allocating decision-making authority in a firm. Importantly, we would not observe this result if different strategic interventions were equally impactful, or if observed top performers had low error due to noise alone. Instead, our results stem from correlated error levels within managers, with some managers being consistently better

able to anticipate both how effective interventions are, and which interventions are more effective.

Conclusion

In this paper we examine strategic decision making by managers. We demonstrate that managers frequently make mistakes regarding (1) which of two strategic interventions will be more effective, and (2) how effective different interventions will be. We then show that variation in managerial ability to predict the impact of strategic interventions cannot be attributed solely to noise in our measures: managerial performance on an arbitrary choice is predictive of their performance for other choices. This implies that there are both "low-" and "high-" ability managers, and that a small signal of managerial ability is useful for identifying whose advice should be followed when making strategic decisions.

Our research exploits a novel combination of data sources. Given the endogenous nature of most estimates of the impact of strategic interventions, we leverage six published studies that report both the causal estimate of strategic interventions as well as manager's forecasts of these effects. By implementing a consistent set of performance measures in each study and sample, we can make "apples-to-apples" comparisons across contexts that prior work has not been able to establish. Furthermore, because our results draw from a diverse set of experiments, contexts, and samples of managers, we are able to provide aggregate results that suggest considerable generalizability to our findings.

Although this study provides several novel insights about a manager's ability to make strategic decisions, it has some limitations worth noting. First, we rely on a set of strategic interventions designed by other scholars and our sample of interventions is unlikely to be representative of the full set of strategic decisions that the average manager makes. Second, managerial beliefs about intervention effectiveness are only one aspect of strategy, with other factors such as implementation costs and inertia also playing a role. Third, managers in the studies we rely on do not predict strategic interventions in their own workplace which means that they do not have access to private information that they might otherwise have access

to. However, in five of the six studies in our sample, managers were asked to predict the effect of strategic interventions that were tested in competitive marketplaces, and not strictly within the organization (the exception is Heinz et al. (2020), who test the effect of layoffs on worker productivity). In these five studies, managers were purposively recruited based on their experience with the decisions and settings where the interventions took place. However, further research should investigate the role of private information on the ability of managers to predict the effects of strategic decisions, especially as some research suggests that beliefs about one's own firm may be especially biased (McKenzie, 2018). Finally, managers were only asked to make forecasts in one domain, and additional work is needed to understand the extent to which forecasting ability in one area translates to other domains.

Our work contributes to a growing literature that details the impact of managers in the firms they work for. While future research should build more robust evidence for our findings, we emphasize three directions for future research with important implications for firms: First, hiring managers may consider forecasting ability as a new dimension to assess when making hiring decisions, especially as we demonstrate that the average manager struggles to anticipate effectiveness of strategic interventions. Second, firms may work to develop the capabilities of managers to have more accurate managerial expectations. Finally, firms could benefit from recognizing an underappreciated feature of organizational experimentation, which is that it shows not only which interventions are effective, but also whose opinion should be trusted when making strategic decisions.

References

Abebe, G., Caria, A. S., and Ortiz-Ospina, E. (2021). The selection of talent: Experimental and structural evidence from ethiopia. *American Economic Review*, 111(6):1757–1806.

Adena, M. and Huck, S. (2020). Online fundraising, self-image, and the long-term impact of ask avoidance. *Management Science*, 66(2):722–743.

Aguinis, H., Audretsch, D. B., Flammer, C., Meyer, K. E., Peng, M. W., and Teece, D. J. (2022). Bringing the manager back into management scholarship. *Journal of Management*, page 01492063221082555.

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P., Moussaid, M., and Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 117(21):11379–11386.

Bertrand, M. and Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *The Quarterly journal of economics*, 118(4):1169–1208.

Bessone, P., Rao, G., Schilbach, F., Schofield, H., and Toma, M. (2021). The economic consequences of increasing sleep among the urban poor. *The Quarterly Journal of Economics*, 136(3):1887–1941.

Bloom, N. and Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries. *The quarterly journal of Economics*, 122(4):1351–1408.

Bloom, N. and Van Reenen, J. (2010). Why do management practices differ across firms and countries? *Journal of economic perspectives*, 24(1):203–224.

Camuffo, A., Cordova, A., Gambardella, A., and Spina, C. (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. *Management Science*, 66(2):564–586.

Camuffo, A., Gambardella, A., Maccheroni, F., Marinacci, M., and Pignataro, A. (2022). Microfoundations of low-frequency high-impact decisions.

Camuffo, A., Gambardella, A., Messinese, D., Novelli, E., Paolucci, E., and Spina, C. (2021). A scientific approach to innovation management: theory and evidence from four field experiments.

Csaszar, F. A. and Laureiro-Martínez, D. (2018). Individual and organizational antecedents of strategic foresight: A representational approach. *Strategy Science*, 3(3):513–532.

Del Carpio, L. and Guadalupe, M. (2022). More women in tech? evidence from a field experiment addressing social identity. *Management Science*, 68(5):3196–3218.

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. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

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.

Demerjian, P., Lev, B., and McVay, S. (2012). Quantifying managerial ability: A new measure and validity tests. *Management science*, 58(7):1229–1248.

Friebel, G., Heinz, M., and Zubanov, N. (2022). Middle managers, personnel turnover, and performance: A long-term field experiment in a retail chain. *Management Science*, 68(1):211–229.

Heinz, M., Jeworrek, S., Mertins, V., Schumacher, H., and Sutter, M. (2020). Measuring the indirect effects of adverse employer behaviour on worker productivity: A field experiment. *The Economic Journal*, 130(632):2546–2568.

Hoffman, M. and Tadelis, S. (2021). People management skills, employee attrition, and manager rewards: An empirical analysis. *Journal of Political Economy*, 129(1):243–285.

Kapoor, R. and Wilde, D. (2022). Peering into a crystal ball: Forecasting behavior and industry foresight. *Strategic Management Journal*.

Koning, R., Hasan, S., and Chatterji, A. (2022). Experimentation and start-up performance: Evidence from a/b testing. *Management Science*, 68(9):6434–6453.

Levinthal, D. A. (2021). *Evolutionary processes and organizational adaptation: a Mendelian perspective on strategic management*. Oxford University Press.

Liebscher, R. and Mählmann, T. (2017). Are professional investment managers skilled? evidence from syndicated loan portfolios. *Management science*, 63(6):1892–1918.

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.

McKenzie, D., Osman, A., and Rahman, A. (2023). Training and subsidies vs pay-for-results in spurring digital marketing take-up and small firm growth.

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.

Otis, N. (2022a). Policy choice and the wisdom of crowds. *Available at SSRN 4200841*.

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

Rau, H., Samek, A., and Zhurakhovska, L. (2022). Do i care if you are paid? field experiments and expert forecasts in charitable giving. *Journal of Economic Behavior & Organization*, 195:42–51.

Reiff, J., Dai, H., Gallus, J., McClough, A., Eitniear, S., Slick, M., and Blank, C. (2021). When impact appeals backfire: Evidence from a multinational field experiment and the lab. *Available at SSRN 3946685*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Ryall, M. and Sorenson, O. (2022). Causal inference as an organizational problem.

Samek, A. and Longfield, C. (2023). Do thank-you calls increase charitable giving? expert forecasts and field experimental evidence. *American Economic Journal: Applied Economics*.

Sorenson, O. (2003). Interdependence and adaptability: organizational learning and the long–term effect of integration. *Management Science*, 49(4):446–463.

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.

Online Appendix

Managerial Expectations

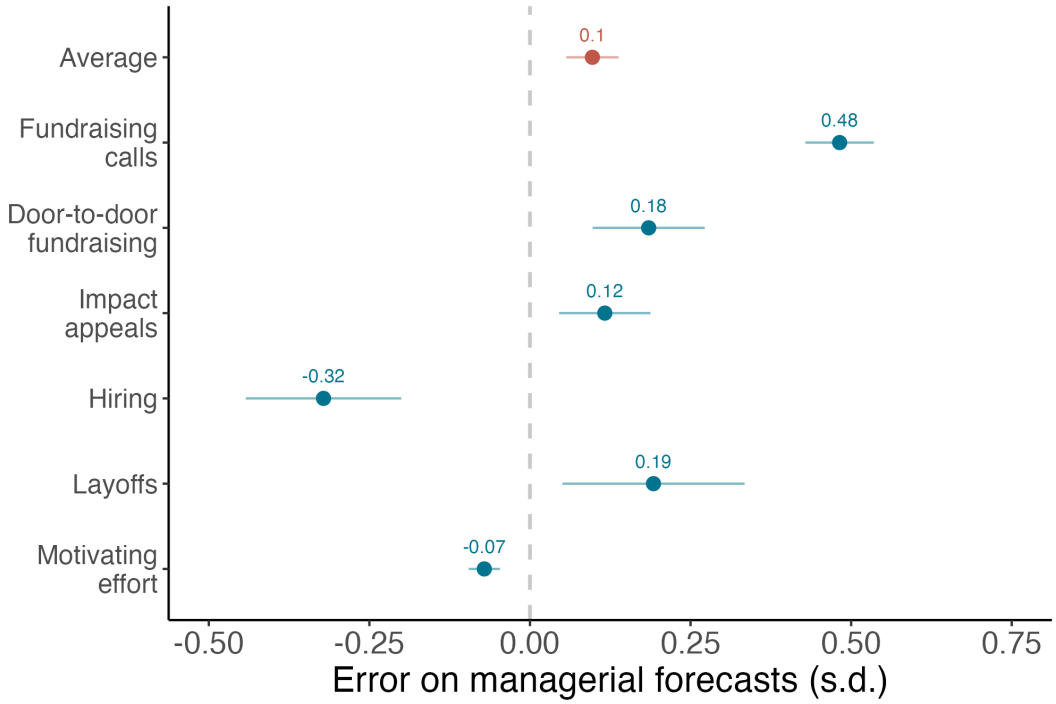
Nicholas G. Otis
UC Berkeley

Mathijs de Vaan
UC Berkeley

Table of Contents

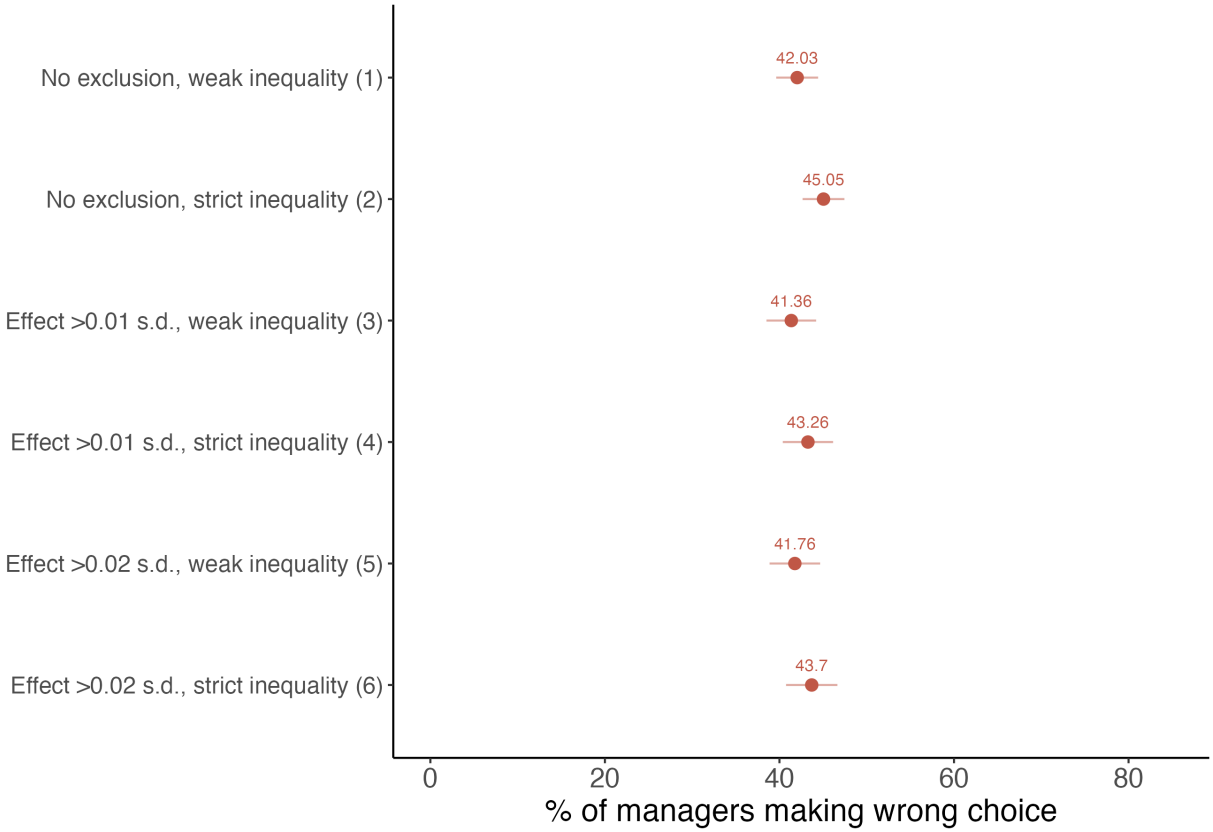
A Appendix Figures

Figure A1: Managerial point errors (directional)



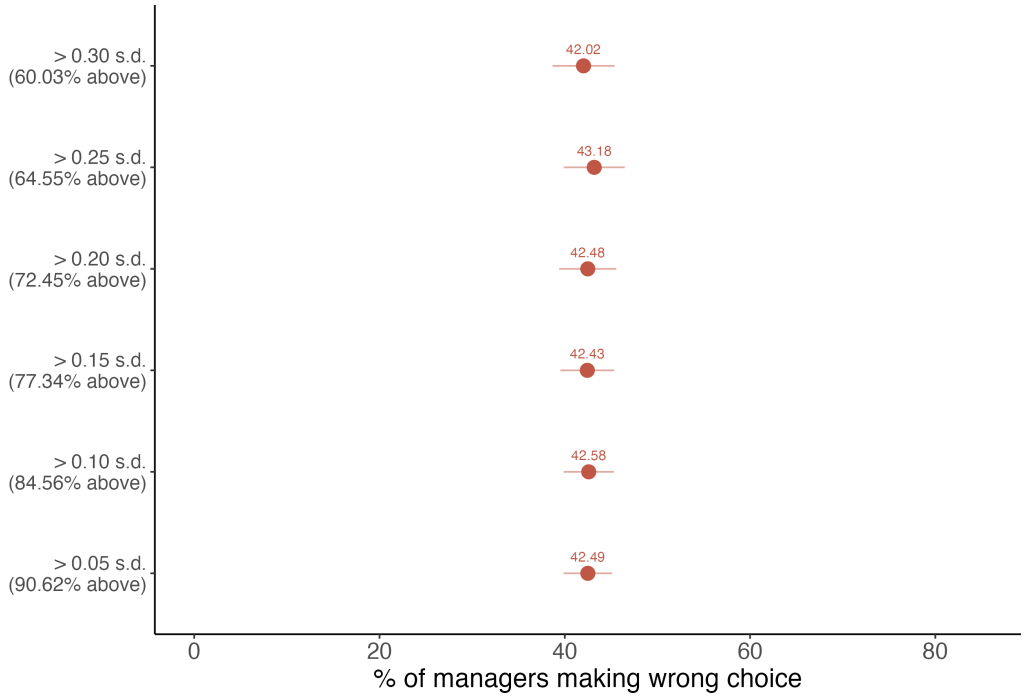
Notes: This figure depicts the average distance between managers' forecasts and observed experimental results measured in standard deviations. The first presents the average error across studies, weighting each study equally. The next six rows display the average error across all outcomes within each study, weighting each forecast equally. Error bars depict 95% confidence intervals with standard errors clustered at the manager level.

Figure A2: Incorrect managerial decisions:
Robustness checks



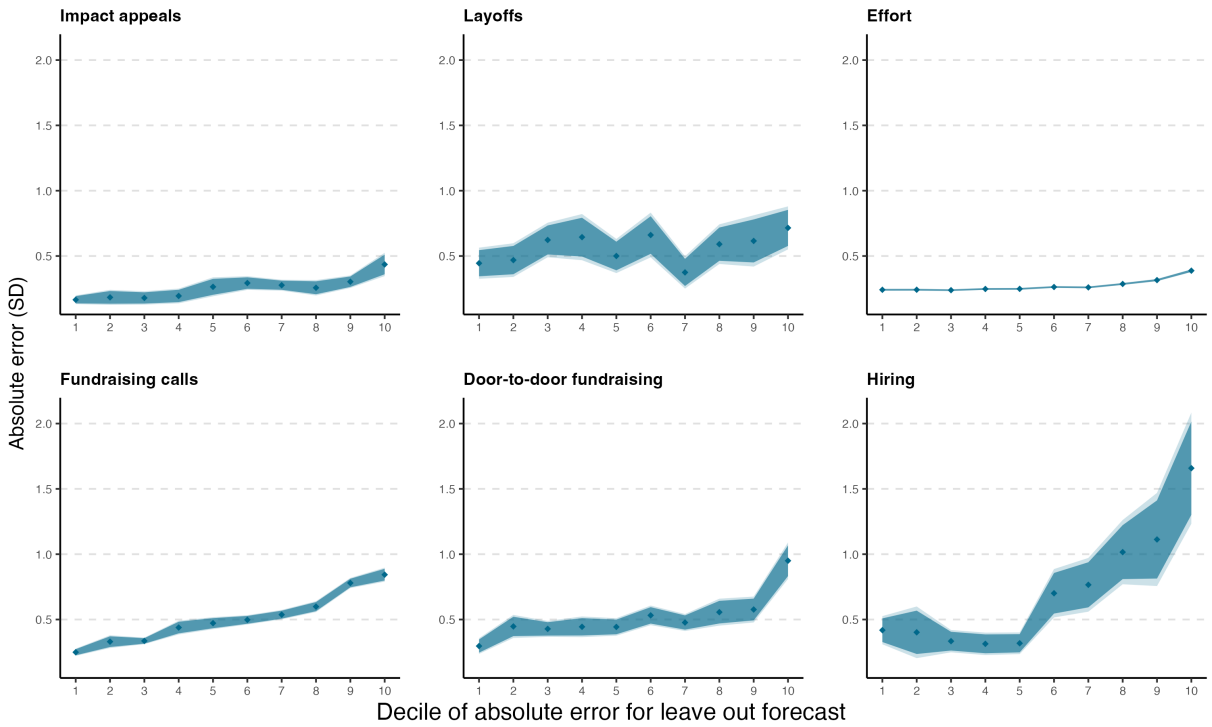
Notes: This figure depicts the percent of manager's making the wrong decision about which of two interventions will be more effective. Points depict averages across all six studies pooled giving each study equal weight. Managers are said to have made the wrong choice if they incorrectly predict which of two interventions will be more effective relative to causal estimates from the randomized experiments. Rows 1 and 2 include comparisons of all interventions. Rows 3 and 4 exclude comparisons that are below 0.01 s.d., and rows 5 and 6 exclude comparisons below 0.02. In rows 1, 3, and 5 managers are said to have made the correct choice if they weakly identify which intervention was more effective, and in rows 2, 4, and 6 they are required to have strictly ranked interventions correctly. Error bars denote 95% confidence intervals around the mean clustered at the manager level.

Figure A3: Incorrect managerial decisions by effect magnitude



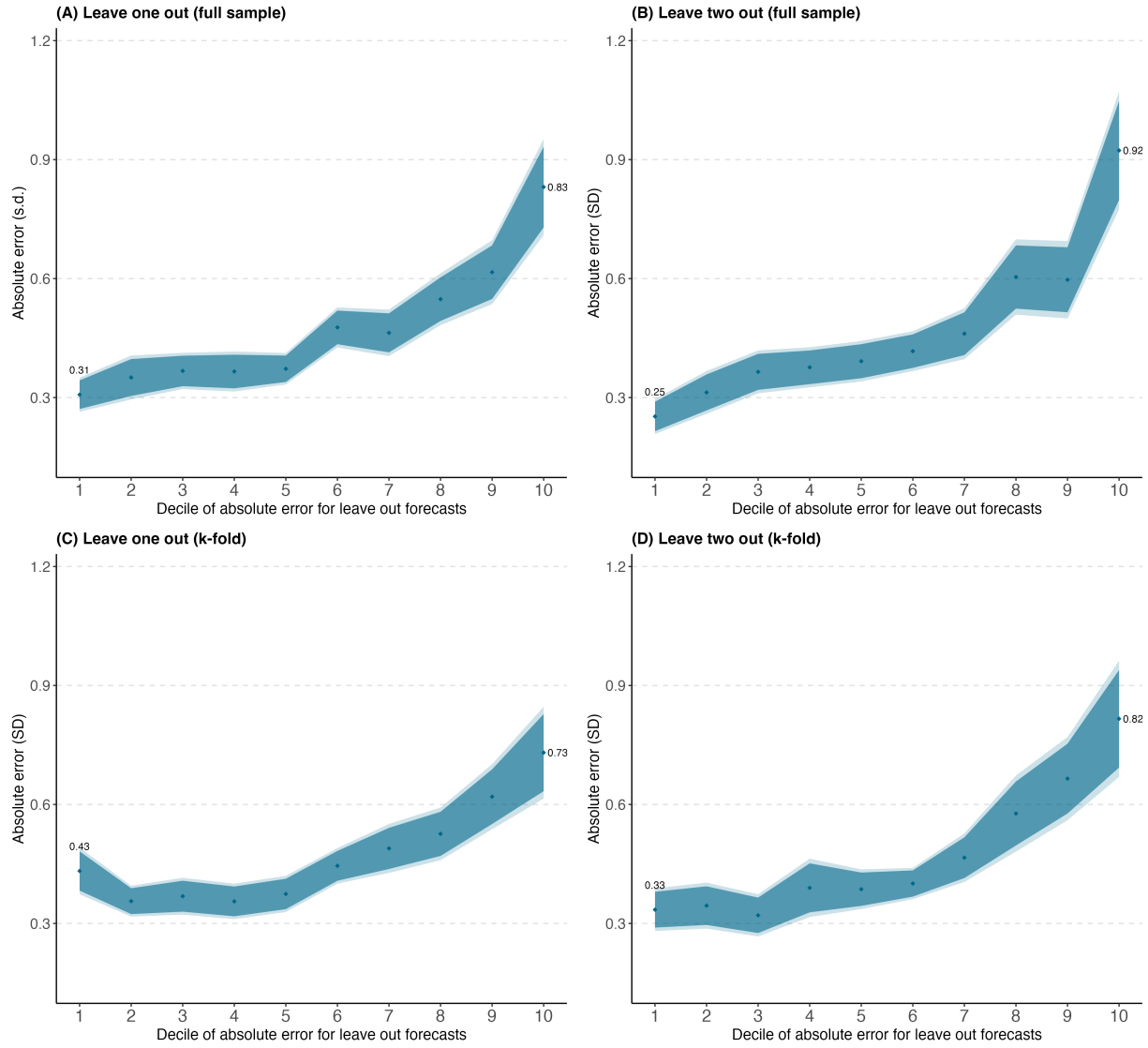
Notes: This figure depicts the percent of manager's making the wrong decision about which of two interventions will be more effective. Points depict averages across all six studies pooled giving each study equal weight. Managers are said to have made the wrong choice if they incorrectly predict which of two interventions will be more effective relative to causal estimates from the randomized experiments. Rows capture the percent of correct choices by whether the difference in predicted effects for a given pair of effects is above $x \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$ s.d. The number of forecasts that are above the x s.d. threshold are presented in parentheses. Error bars denote 95% confidence intervals around the mean clustered at the manager level.

Figure A4: Conditional managerial error:
Study-level effects



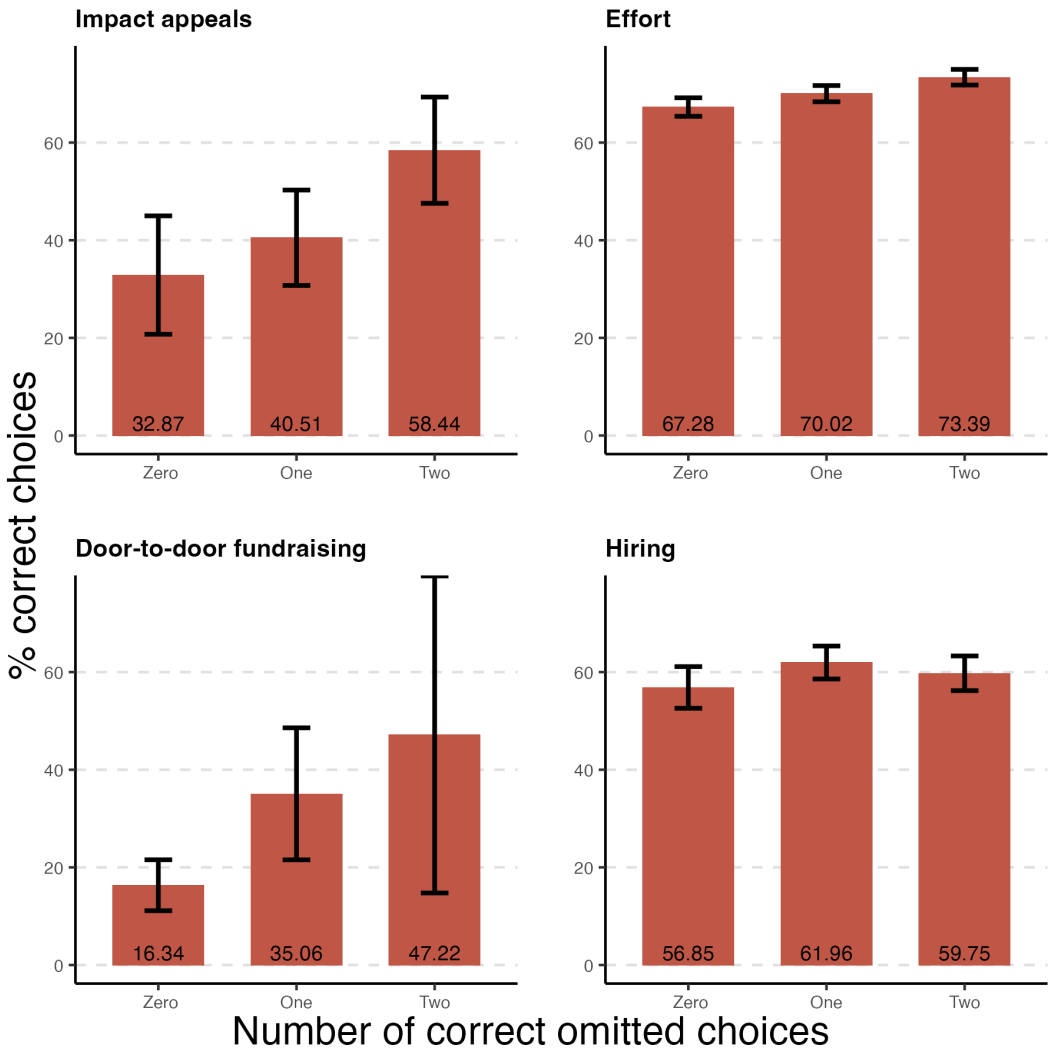
Notes: This figure presents the average absolute error on managers' forecasts (y-axis) conditional on managers' error on one omitted prediction, pooled by decile of absolute error (x-axis). Conditional error is calculated separately for each omitted prediction, rotating through all the predicted effects in each study, and then pooling across all omitted effects. Standard errors are clustered at the manager level. Light and dark bars represent 90 and 95% confidence intervals.

Figure A5: Conditional managerial error:
Robustness checks



Notes: In Panel (A), conditional absolute error is generated by creating all permutations of two effects in each study, denoted generically as j and j'. For each pair, we regress the absolute error for j on the decile of absolute error for j'. We repeat this procedure across all pairs of effects in each study. Panel (B) presents results conditional on the quantile of the sum of absolute error from two omitted effects instead of one. In Panels (C) and (D) we conduct a k-fold procedure to avoid overfitting. For each two-effect pair, we: (i) Randomly split the sample into 10 folds; (ii) Omit fold $k = 1$, and regress forecasters' absolute error for effect j on their absolute error for effect j' using the data from folds $k = 2, ..., 10$; (iii) Generate fitted values of absolute error for the omitted $k = 1$ fold from this regression; (iv) Rotate through the data applying steps 1-3 omitting each fold separately, and calculating fitted values of absolute error for each. We repeat steps $1 - 4$ for each permutation of two effects. Then we calculate the average predicted out of sample absolute error for each forecaster across all effect permutations, which we separate into deciles. Panel (D) uses the same procedure but uses two omitted effects when calculating out of sample fitted values of absolute error. In each panel, we calculate the average conditional absolute error across studies giving each equal weight. Light and dark bars present 90 and 95% confidence intervals with robust standard error clustered at the manager level.

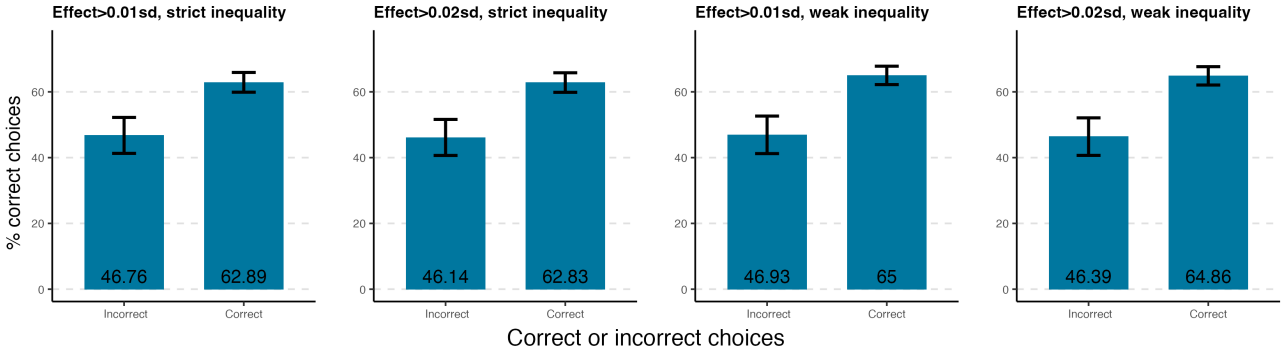
Figure A6: Conditional rank decisions:
Study-level effects



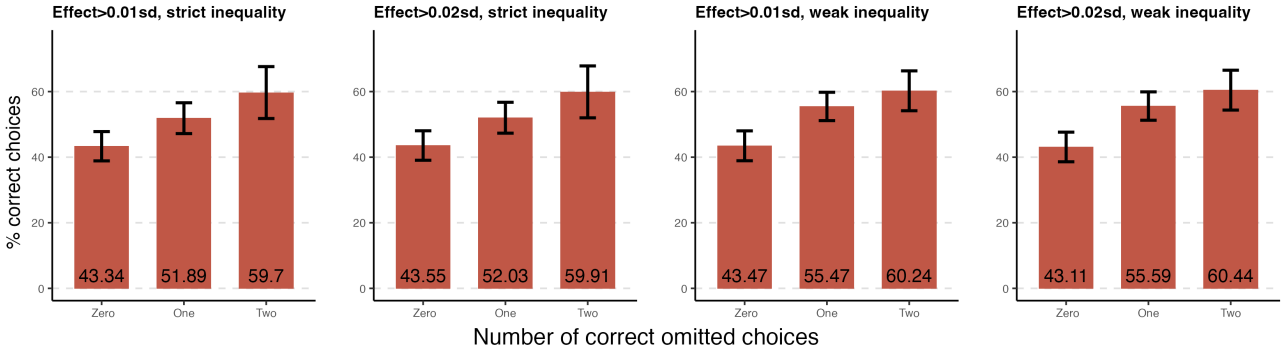
Notes: This figure displays the percentage of managers who incorrectly rank which intervention will be more effective based on their performance on two omitted leave-out choices. Rank-error is measured using causal estimates from the randomized experiments. Conditional error is calculated by omitting each pair of managerial choice and rotating through all pairs of predicted effects in each study. Experimental comparisons with effect differences less than 0.01 s.d. are excluded, as are comparisons where the same forecast is used in the conditioning variables and outcome. For the "Effort" study, the analysis is based on a random sample of 5,000 out of over 750,000 combinations of conditional rank-variables. Standard errors are clustered at the manager level.

Figure A7: Conditional managerial decisions:
Robustness checks

(A) Conditioning on one choice



(B) Conditioning on two choices



Notes: Both panels depict percent of manager decisions whose forecasts incorrectly rank which intervention will be more effective conditional on their performance on two omitted leave-out choices. Managers are said to have made the right choice if they correctly (strictly) predict which of two interventions will be more effective, relative to the causal estimates from the randomized experiments. Conditional error is calculated omitting every managerial choice, rotating through all the pairs of predicted effects in each study. Comparisons where the difference in experimental effects are less than 0.01 s.d. are omitted, as are conditional comparisons where the same forecast is used in conditioned outcome and in the conditioning variables. Panel (A) pools results across studies, giving each study equal weight. Panel (B) presents results at the study level. Both panels cluster standard errors at the manager level.

B Appendix Tables

Table A1: Review of previous research

Title	Average point	Individual error			
		Point	Rank	Cond. point	Cond. rank
(1)	(2)	(3)	(4)	(5)	(6)
Measuring the Indirect Effects of Adverse Employer Behaviour on Worker Productivity (Heinz et al., 2020)	✓	×	×	×	×
The Selection of Talent: Experimental and Structural Evidence from Ethiopia (Abebe et al., 2021)	✓	×	×	×	×
When Impact Appeals Backfire (Reiff et al., 2021)	✓	×	×	×	×
Do Thank-You Calls Increase Charitable Giving? (Samek and Longfield, 2023)	✓	×	×	×	×
Do I Care if You Are Paid? Field Experiments and Expert Forecasts in Charitable Giving (Rau et al., 2022)	✓	×	×	×	×
What Motivates Effort? Evidence and Expert Forecasts (DellaVigna and Pope, 2018a)	✓	✓	✓	✓	×

Notes: This table presents a summary of the main analyses used in the six papers included in our data. ✓ and × whether a particular type of analysis or error measure was or was not included in the paper. Col. 2 captures error measured using the average prediction across all managers. Col. 3 examines at individual manager-level error, as measured through an absolute or a quadratic loss function. Col. 4 displays at manager-level rank error. Col. 5 captures point error (as in Col. 3) conditional on performance on leave-out point predictions. Col. 6 examines rank error (as in Col. 4) conditional on performance on leave-out rank predictions. Our paper implements all 5 measures/analyses for each study.

Table A2: Overview of experimental details

Study	Outcome	Treatment
Calls		
	Donation rate	Call (Experiment 1)
	Donation rate	Call (Experiment 2)
	Donation rate	Call (Experiment 3)
	Donation rate	Reference (not forecast): Control
	Donation amount	Call (Experiment 1)
	Donation amount	Call (Experiment 2)
	Donation amount	Call (Experiment 3)
	Donation amount	Reference (not forecast): Control
Fundraising		
	Donation rate	Volunteer fundraiser
	Donation rate	Paid fundraiser
	Donation rate	Reference (not forecast): Control
	Donation amount	Volunteer fundraiser
	Donation amount	Paid fundraiser
	Donation amount	Reference (not forecast): Control
Impact		
	Feedback rate	Time
	Feedback rate	Voice
	Feedback rate	Help
	Feedback rate	Expert
	Feedback rate	Reference (not forecast): Control
Layoffs		
	Average calls made (by non-laid off workers)	No layoffs
	Average calls made (by non-laid off workers)	Layoffs (remaining workers aren't informed about layoffs)
	Average calls made (by non-laid off workers)	Layoffs (remaining workers are informed about layoffs)
Effort		
	Effort in a simple clerical task	Piece rate, 4 cent
	Effort in a simple clerical task	Very low pay
	Effort in a simple clerical task	Red Cross, 1 cent
	Effort in a simple clerical task	Red Cross, 10 cents
	Effort in a simple clerical task	40 Cent Bonus
	Effort in a simple clerical task	Discounting: 2 weeks
	Effort in a simple clerical task	Discounting: 4 weeks
	Effort in a simple clerical task	40 cent threshold bonus
	Effort in a simple clerical task	40 cent threshold bonus - loss
	Effort in a simple clerical task	80 cent threshold bonus
	Effort in a simple clerical task	1% chance of $1
	Effort in a simple clerical task	50% chance of 2 cents
	Effort in a simple clerical task	Social comparisons
	Effort in a simple clerical task	Ranking
	Effort in a simple clerical task	Reference (not forecast): Control
	Effort in a simple clerical task	Reference (not forecast): Piece rate, 1 cent
	Effort in a simple clerical task	Reference (not forecast): piece rate, 10 cent
Hiring		
	Application rate	Application incentives
	Application rate	Reference (not forecast): Control
	Application rate	Reference (not forecast): High wage
	Applicant quality (average)	Application incentives
	Applicant quality (average)	Reference (not forecast): Control
	Applicant quality (average)	Reference (not forecast): High wage
	Applicant quality (top applicants)	Application incentives
	Applicant quality (top applicants)	Reference (not forecast): Control
	Applicant quality (top applicants)	Reference (not forecast): High wage

Notes: This table lists experimental outcomes and interventions in each of the six studies. It also provides details on reference conditions that were provided to the managers but which they did not forecast.

Table A3: Details on forecasts

				Number of interventions			Number of ranked pairs		
	Forecasters	Experiments	Outcomes	All	Predicted	Reference	All	>0.01 s.d.	>0.02 s.d.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Layoffs	43	1	1	3	3	0	3	3	3
Hiring	195	1	3	3	1	1	6	6	6
Impact Appeals	42	1	1	5	1	1	10	8	7
Thank you calls	141	3	2	2	2	1	6	2	2
Door-to-door fundraising	100	1	2	3	2	1	6	5	5
Effort	160	1	1	18	15	3	150	148	144

Notes: This table provides an overview of the number of forecasts and experimental features across the six studies. Cols. 1–3 present the number of managers, randomized experiments, and outcomes in each study. Cols. 4 presents the number of interventions evaluated in each study, and Cols. 5 and 6 separate this into the number of interventions that managers provided predictions of and the number of reference interventions that managers were given as a benchmark. Col. 7 presents the number of unique combinations of two interventions and an outcome, where managers provided predictions on at least one of the interventions. This excludes pairs of two reference interventions. Cols. 8 and 9 exclude pairs where the difference in experimental effects is less than 0.01 s.d. and 0.02 s.d., respectively.

C Screening and Exclusion Criteria

Study identification

Studies were identified using the following procedure. First, a research assistant reviewed all papers citing three prominent papers that collected predictions of social science results (DellaVigna and Pope, 2018a,b; DellaVigna et al., 2019). We also search for papers on google scholar using combinations of the search terms ("manager")×("forecast","expectation", "prediction", "beliefs")×("experiment","RCT","randomized").

Within papers that collected predictions of experimental results, we identified those that (i) collected point predictions from managers, consultants, or MBAs of the causal effects of interventions, which excludes a recent set of studies that have looked at predictions from academic researchers (DellaVigna et al., 2020; Bessone et al., 2021; Milkman et al., 2021) or laypeople (e.g., Thomas et al. (2020); Otis (2022a,b); (ii) were published or had made their data publicly available, (iii) contained less than five predictions from managers.

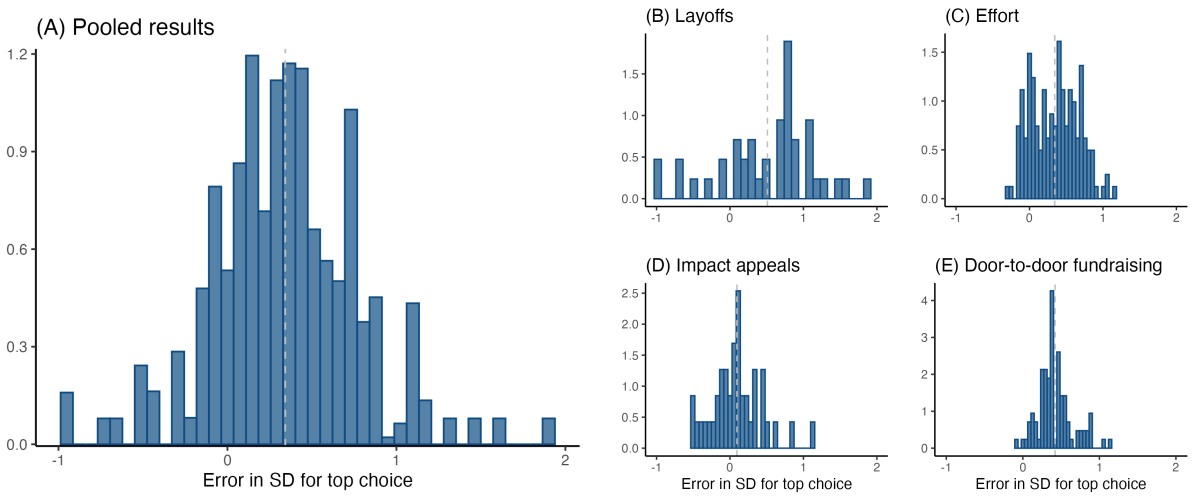
Manager exclusion criteria

Our analytic sample of managers follows the criteria used in each of the six studies. Additionally, we exclude a small number of managers who did not complete the forecasting surveys. This results in the exclusion of 4 of 145 managers in (Samek and Longfield, 2023) and one of the 101 managers in (Rau et al., 2022).

D Error on Top Choices

How well are managers able to predict the effect of their most preferred option? In Panel A of Figure A8 we present a weighted distribution plot of managers' predictions of their most-preferred intervention relative to control. Values above 0 indicate that managers overestimate the effects of their top choice, and values below 0 represent forecasts that are below the experimentally estimated effect.

Figure A8: Managerial error for top choices



Notes: This figure presents the error on the intervention that managers predict will be most effective relative to control. Panel (A) presents pooled results across the four studies which have at least two non-control interventions, giving each study equal weight. Panels B-E present results for each of the four studies. Vertical dashed line denotes the average error in each study.

On average, managers overestimate the experimental effect of their top choice by 0.34 s.d. (95% CI = $[0.28, 0.40]$), which is indicated by the gray dashed vertical line. Note that this tendency is found only for the managers' top choices, and not across their full set of predictions. For example, in DellaVigna and Pope (2018a) the average error on the top prediction was 0.34 s.d. (95% CI = $[0.29, 0.39]$). However, across all forecasts the average error was actually negative (-0.07 s.d., 95% CI = $[-0.05, 0.10]$). Panels B-E of Figure A8 depict results by study and show that in each experiment managers overestimate the effects of their top choice. Together these results indicate that managers on average substantially overestimate the effects of the interventions that they believe will be most effective.