ABCs of principal–agent interactions: Accurate predictions, biased processes, and contrasts between working and delegating

Katherine A. Burson a,*, David Faro b, Yuval Rottenstreich c

* Ross School of Business, University of Michigan, United States
b London Business School, United Kingdom
c Stern School of Business, New York University, United States

A B S T R A C T

We experimentally investigate people's evaluations of incentive pay contracts and people's predictions of others' evaluations of incentive pay contracts. We emphasize that the construction of evaluations and predictions often includes two substeps, involving likelihood judgment and likelihood weighting. Predictors appear to be biased at both substeps but in opposing directions. Accurate overall predictions thus sometimes reflect two errors that are of the same magnitude and thereby offset. Moreover, predictions can become more inaccurate if one step is debiased but the other is left untouched. Importantly, principals deciding whether to delegate a task are susceptible to just one of the biases. Delegation assessments are thus often flawed, reflecting a single error that is not offset.

© 2010 Elsevier Inc. All rights reserved.

Introduction

A foundational observation of agency theory is that pay-for-performance regimes involve an unavoidable tradeoff: They can motivate employees to work hard, but they do so by exposing risk-averse employees to uncertainty (see, e.g., Eisenhardt, 1989; Holmstrom, 1979; Jensen & Meckling, 1976; Levinthal, 1988). For instance, suppose a salesperson will receive a commission only if he or she completes a sale. The uncertainty of compensation imposes risk on the salesperson—no sale means no pay—and may thus motivate him or her to work hard. However, any resulting motivation does not come for free. The salesperson must be offered an “upside” large enough to compensate him or her for bearing risk. A profit-maximizing principal will presumably strive to structure compensation in a way that induces agents to work hard by imposing as little risk as possible and offering as little upside as possible. By implication, principals must predict potential agents’ perceptions and reactions to uncertainty.

In this paper, we experimentally investigate people’s evaluations of incentive pay contracts and people's predictions of others' evaluations of incentive pay contracts. Our points of departure are motivational theories such as expectancy theory (Vroom, 1964) and models of decision making under uncertainty such as prospect theory (e.g., Kahneman & Tversky, 1979). Expectancy theory stresses the role of likelihood perceptions in the evaluation of incentive pay. For example, in considering an incentive to complete a particular sale, a salesperson will assess how likely it is that he or she will successfully complete the sale. Prospect theory stresses likelihood perceptions as well and in addition highlights a second factor: reactions to likelihood perceptions. For example, if two salespeople believe the probability of a successful sale is 10%, one salesperson may find this likelihood appealing, react positively to it, and therefore find an associated incentive scheme attractive. On the other hand, the other salesperson may find this likelihood dispiriting, react negatively, and find an associated incentive scheme unattractive.

More formally, let $U(x, T) = r(x)pw(T)$ be the utility or attractiveness of an incentive scheme that pays an amount $x$ for successful completion of task $T$. Here, the function $r$ indexes the subjective value of monetary payments, $p$ quantifies beliefs about the likelihood of successfully completing tasks, and $w$ reflects the weight or impact of beliefs (see, e.g., Fox & Tversky, 1998; Tversky & Fox, 1995). We distinguish between $v_w$, $w_p$, and $p_w$ vs. $v_o$, $w_o$, and $p_o$, where the subscript $s$ marks an individual's own evaluations and the subscript $o$ marks either predictions of others' evaluations or predictions about others (depending on the context).
individual assesses the attractiveness of personally performing some task for a potential payment, \( v, w, \) and \( p_0 \) are implicated, and we write \( U_s = v(x)w_1[p_{o}(T)] \). The idea is that any evaluation of an incentive scheme can be stitched together from subevaluations of value, likelihood, and weight. Note also that because the evaluation in question is of potential benefits, it will be compared to any disutility of effortfully working on the relevant task, opportunity costs, and the like. In respect to predictions, when an individual forecasts how attractive someone else will find working on some task for a potential payment, \( v_0, w_0, \) and \( p_0 \) are implicated, and we write \( U_0 = v_0(x)w_0[p_{o}(T)] \). The idea is that just as evaluations can be stitched together from subevaluations, predictions can be stitched together from subpredictions.

An interesting mix of subevaluations and subpredictions arises when, rather than personally working on some task, a principal considers delegating the task to an agent who will be paid via an incentive contract. The principal will receive a payoff if the agent successfully completes the task (it is presumably from this payoff that the principal will remunerate the agent). The principal must thus predict the agent’s chances of success. So \( p_0 \) is implicated (rather than \( p_0 \)). However, the principal must consider the value of the potential payoff to him or herself and the impact of the relevant uncertainty on him or herself. So \( v_0 \) and \( w_0 \) are implicated (rather than \( v_1 \) and \( w_1 \)). The utility of delegating rather than personally performing some task is thus stitched together from \( p_0, v_0, \) and \( w_0 \). We find that in the context we study, forecasts of value tend to be on target, but forecasts of likelihood and weight are often biased. Interestingly, the bias in likelihood forecasts tends to be in the opposite direction of the bias in forecasts of weight. For difficult tasks, where \( p_0 \) and \( p_0 \) are relatively small, we find that people tend to think others are more optimistic about the likelihood of success than they actually are, so \( p_0 > p_0 \), but people also forecast that others will weigh beliefs more negatively than they actually do, so \( w_0 < w_0 \). Put differently, at low probabilities, we observe that \( p_0 \) tends to be too large and \( w_0 \) tends to be too small. For easy tasks, where \( p_0 \) and \( p_0 \) are relatively large, people think others are more pessimistic about likelihood than they actually are, \( p_0 < p_0 \), but people also forecast that others will weigh beliefs more positively than they actually do, \( w_0 > w_0 \). In other words, at high probabilities, \( p_0 \) tends to be too large and \( w_0 \) tends to be too small. Because the biases in \( p_0 \) and \( w_0 \) are in opposite directions, our results suggest that in any given context, the overall accuracy of predictions, whether \( U_0 \) tends to equal \( U_0 \), will depend on whether these biases are of about the same magnitude. Even if \( p_0 \) and \( w_0 \) are both off target, they may offset to yield overall predictions that are on target. Interestingly, then, interventions that attempt to debias just one substep, either \( p \) or \( w \), may improve accuracy at that individual substep yet damage overall accuracy.

Second, we contrast evaluations of delegating a task with evaluations of personally working on a task and predictions of another person’s evaluation of working on a task. We find that people’s evaluation of the utility of delegating reveals systematic biases. Recall the formula for the utility of delegating, \( U_0(x, T) = v(x)w_1[p_{o}(T)] \). Because \( U_0 \) includes \( p_o \) but not \( w_0 \), it implicates only a likelihood prediction bias and no weighting prediction bias. When likelihood predictions err, \( U_0 \) will diverge from the utility of personally working on the same task, \( U_s = v(x)w_1[p_{o}(T)] \), which reflects neither prediction bias. When weighting predictions err, \( U_0 \) can diverge from predictions of the utility of another person’s working on the same task, \( U_0 = v(x)w_0[p_{o}(T)] \), which reflects both predictive biases.

The divergence of \( U_0 \) from \( U_0 \) and \( U_0 \) is normatively inappropriate. In the model we present, the issue is clear: Evaluations of delegation include just one predictive bias that is not offset and thus inevitably rears its head. Moreover, even at an observational level that is agnostic about underlying models, the divergence of \( U_0 \) from \( U_0 \) and \( U_0 \) is problematic. Consider the hypothetical twins Jack and Jill. Suppose Jack considers delegating some task to Jill, whom Jack knows is identical to him in every relevant way. Then, a correct evaluation of delegation along with accuracy of predictions would yield, for any fixed reward \( x \), \( U_0 = U_0 = U_0 \). Our analysis indicates that the left-most equality will not obtain. The very act of considering an incentive scheme as a delegation prospect rather than a work prospect will inappropriately change the prospective utility of the available benefits. This is a framing effect: Why should the appraisal of potential benefits depend on whether one performs the work or delegates it?

An ABC mnemonic summarizes our analysis: Accurate predictions of the attractiveness of work to someone else can arise from offsetting Biases, but a single, unopposed bias underlies delegation and thus creates a Contrast between the attractiveness of work and delegation.

**Applicability of our framework**

Because we examine evaluations and predictions of utility, our analysis applies to situations in which people’s behavior is shaped by the calculations of benefits and costs reflected by such assessments. As we have mentioned, such behavior is consistent with important treatments of motivation such as expectancy theory and agency theory and with theories of individual choice such as prospect theory. Nevertheless, there are clearly many circumstances in which people’s behavior is shaped by other considerations beyond benefits and costs.

We next turn to a review of past work that helps delineate the boundary conditions of our investigation by pointing out situational variables that indeed orient people away from relying on evaluations and predictions of utility as the basis of behavior. We also consider extant research that is consistent with evaluations and predictions of utility and that therefore offers findings that inform our hypotheses.

**Relevance of social factors**

Experiments that instantiate principal–agent interactions tend to find marked deviations from the predictions of agency theory (Conlon & Parks, 1990; Falk, 2007; Fehr, Klein, & Schmidt, 2007; Fehr & Schmidt, 2004; Hannan, Kagel, & Moser, 2002; Miller & Whitford, 2002; Parks & Conlon, 1995; see also Bottom, Holloway, Miller, Mislin, & Whitford, 2006). Principals frequently offer high, guaranteed salaries and uncertain pay that is insufficient, at least theoretically, to induce agents to work hard. Nevertheless, agents frequently provide very high levels of effort. The conclusion commonly drawn from such studies is that fairness, trust, reciprocity, and similar factors connect principals and agents in at least some minimal way and that these connections prompt agents to work hard, thereby obviating any need to rely on the imposition of risk as a motivational tool (see, e.g., Akerlof & Kranton, 2000).

In this view, behavior in many principal–agent interactions is not shaped by evaluations and prediction of utility per se, but by consideration of roles and identity, relationships and bonds, norms
and rules, and the like. In other words, behavior is shaped by social factors rather than, say, perceptions of likelihood and weighting. For instance, as March and Heath (1994, pp. 57–58) put it: “When individuals follow rules . . . they see as appropriate. Neither preferences nor expectations [of likelihood] enter directly into the calculus. Decision makers ask: What kind of situation is this? What does a person such as I do in a situation such as this?” (see also Stone & Allgaier, 2008).

Accordingly, the experimental methodology we instantiate is set up to minimize the relevance of social factors in ways we detail later. The types of situations our experiments are meant to mimic are thus ones in which connections between parties are not primary drivers of behavior, so that evaluations and predictions of utility may predominate.

**Availability of information about payoffs**

Beyond the specific setting of principal–agent interactions, accurate predictions of others’ reactions to uncertainty are critical in any strategic interaction. Bottom (1998), for instance, emphasized the role that forecasts of others’ risk attitudes can play in determining the outcome of bilateral negotiations (see also Bottom & Studt, 1993; Murnighan, Roth, & Schoumaker, 1988; Schurr, 1987). Consider two individuals negotiating the distribution of a fixed resource. Suppose it is common knowledge that in the absence of a negotiated agreement, both bargainers will be consigned to a relatively unattractive outcome. Then, to avoid the unappealing outcome associated with a breakdown in negotiation, highly risk-averse bargainers may be flexible in their demands. Thus, if a wise bargainer correctly anticipates the risk attitudes of a very risk-averse counterpart, he or she will recognize the opportunity to push for splits of the available resource that are particularly favorable to the bargainer.

Importantly, it is often the case that the payoffs available to negotiators, or to participants in other strategic interactions, are private information rather than common knowledge. In such circumstances, the parties involved may guess at the payoffs available to one another and the values that these payoffs hold for each party. For example, a bargainer may form an estimate of his counterpart’s reservation price (Bottom & Pease, 1997). Likewise, a bargainer may estimate the value to his counterpart of various integrative-deal structures. A foundational observation in the literature on negotiations holds that in constructing such estimates, negotiators form beliefs about others’ payoffs and values that hew too closely to their own payoffs and values (Carnevale & Isen, 1986; Kimmel, Pruitt, Magneau, Konar-Goldband, & Carnevale, 1980; Pinkley, Griffith, & Northcraft, 1995). Many critical patterns, such as inappropriate assumptions of a fixed pie in integrative settings (Bazerman, Magliozi, & Neale, 1985; Thompson & Hastie, 1990) or negotiator overconfidence in purely distributive settings (Larrick & Wu, 2007) follow from such beliefs.

An appreciation of this foundational insight is useful in understanding our experimental method. In our experiments, the payoffs are simple monetary amounts that are common knowledge. Thus, our experiments minimize the relevance of perceptions and predictions of others’ payoffs and values. Put in terms of the framework we have presented, because our experiments minimize potential variance in predictions of \( v \), they apply most directly to settings in which what is most central is indeed uncertainty about whether or not a work task will be successfully completed. In essence, our experiments put the onus on predictions of \( p \) and \( w \). This accords with the nature of principal–agent interactions: By virtue of the fact that an agent’s payoffs are typically paid by the principal, in most principal–agent settings, the agent’s payoffs are common knowledge between the principal and agent.

**Specific vs. generic knowledge**

In some strategic interactions, the parties involved have “specific knowledge” of one another: They know each others’ identities and may even be highly familiar with each other. In other strategic interactions, the parties have only “generic knowledge” of one another: They may not know each others’ individual identities or may not be familiar with each other. For instance, a manager may offer incentive pay to one specific individual or to thousands of employees, most of whom he has never personally met and many of whom work in far-flung geographic locations.

Considering the risk attitudes and reactions to uncertainty of a pool of people whom one can only generically characterize is a common task in many management settings and in many marketing domains as well (Dowling & Staelin, 1994; Kahn & Luce, 2003; Kunreuther et al., 2002). Numerous product categories include goods that are intentionally marketed to different consumer segments (e.g., single people vs. married people) in large part because these goods differ in their risk profile (e.g., low-premium, minimal-coverage insurance policies vs. high-premium, expansive-coverage policies; sports coupes that are relatively dangerous vs. sedans and SUVs that are perceived as safer).

Our experiments concern situations that tap generic knowledge. For example, we have participants from a large pool indicate either their own reactions to uncertainty (i.e., provide \( u_p \)) or predict the reactions of a representative participant from the pool (i.e., forecast \( u_0 \)). The patterns we observe may or may not generalize to situations in which people have specific knowledge (cf. Menon, Kyung, & Agrawal, 2009).

**Task difficulty and biased likelihood perceptions**

We hypothesize that for difficult tasks, \( p_b \) tends to be greater than \( p_b \), whereas, for easy tasks, \( p_e \) tends to be less than \( p_b \). That is, for difficult tasks, people tend to think others are more optimistic about the likelihood of success than they actually are, but for easy tasks, people tend to think others are more pessimistic than they actually are.

Research on better-than-average and worse-than-average effects underlies this hypothesis. Many studies in this area (see, for instance, Kruger, 1999; Kruger & Burs, 2004; Kruger, Windschitl, Burs, Fessel, & Chambers, 2008) have asked participants to predict either their own or a representative person’s absolute performance at some task (e.g., the number of questions answered correctly on a 20-question quiz). In some studies, however, participants estimated either their own or a representative person’s chances of achieving some benchmark (e.g., the likelihood of answering 16 or more questions correctly). These latter studies indeed indicate that when participants estimate their own and peers’ likelihood of success at easy tasks, they estimate their own chances to be high but others’ chances not as high. When considering difficult tasks, participants estimate their own chances of success to be low but others’ chances not as low (see, for instance, Moore & Kim, 2003; Moore & Small, 2007; Windschitl, Kruger, & Simms, 2003).

There is a significant hurdle in generalizing from the studies just mentioned to the settings with which we are concerned. These studies asked participants to estimate other people’s chances of success—a parallel would be asking a manager to estimate how likely it is that employees will complete some task. In contrast, we are often interested in people’s estimates of other people’s beliefs about their chances of success—a parallel would be asking a manager to estimate how likely employees think it is that they will complete some task. In other words, earlier studies asked participants to evaluate others’ chances, whereas we are also interested in asking participants to assess others’ thinking about others’
chances. That the two entities yield qualitatively equivalent patterns cannot simply be assumed. Our studies test this assumption and explore its implications.

Task difficulty and biased weighting perceptions

We hypothesize that the direction of bias evinced by forecasts of the weighting of likelihood may be the opposite direction of bias evinced by forecasts of likelihood. For difficult tasks, \( p_s \) may tend to be greater than \( p_w \), but \( w_s \) will tend to be less than \( w_w \). On the other hand, for easy tasks, \( p_s \) may tend to be less than \( p_w \), but \( w_s \) will tend to be greater than \( w_w \).

To develop this hypothesis, we again draw on extant research. Much work reveals that people react in a pronounced way to differences between small probabilities and impossibility and to differences between large probabilities and certainty (Camerer & Ho, 1994; Gonzalez & Wu, 1999; Kilka & Weber, 2001; Tversky & Fox, 1995; Wu & Gonzalez, 1996). For instance, Tversky and Kahneman (1992) found that their median participant was indifferent between receiving (a) a 1% chance at $200 or (b) $10 for sure and also indifferent between receiving (c) a 99% chance at $200 or (d) $188 for sure. Thus, the impact of the first upward hundredth of movement away from impossibility was substantial, worth $10, and the impact of the first downward hundredth of movement away from certainty was also substantial, worth $12. In contrast, the impact of each of the 98 intermediate hundredths was negligible, worth $178 total, or $1.78 per hundredth. Such behavior is consistent with \( w_s(p) > p \) for difficult tasks of small \( p \) and with \( w_w(p) < p \) for easy tasks of large \( p \). “Overweighting” of small probabilities and “underweighting” of large probabilities captures the marked impact of departures from impossibility and certainty.

Faro and Rottenstreich (2006) found that people do not adequately anticipate others’ overweighting and underweighting (see also Hsee & Weber, 1997). Their median participant was indifferent between a 1 in 1000 chance of winning $4000 or $10 for sure but thought a typical peer would ask for only $5 for sure. Evidently, for small likelihoods, \( w_s < w_w \). On the other hand, the median participant was indifferent between a 99 in 100 chance at $4000 or $3250 for sure but thought a peer would ask for $3900. Evidently, for large likelihoods, \( w_s > w_w \). Thus, forecasts of others’ weighting may be regressive. One explanation for the pattern found by Faro and Rottenstreich is that people react emotionally to risk themselves but also have an “empathy gap.” Because of this empathy gap, people fail to recognize that others have emotional reactions to the same events that they experience (Loewenstein, 1996). Predictions of others therefore seem to take a more rational view than is warranted. In fact, manipulations of empathy and measurements of empathy appear to moderate mispredictions of \( w \) (Faro & Rottenstreich, 2006; Hsee & Weber, 1997).

As before, there is a significant hurdle in generalizing from Faro and Rottenstreich’s studies to our setting. Critically, Faro and Rottenstreich provided participants with numerical chances (e.g., 1 in 1000) and examined the impact of these specified likelihoods. However, employees are almost never provided with a numerical chance of reaching relevant performance benchmarks. Instead, they must form their own estimate of likelihood (as captured by the function \( p \)) and then weight or react to their estimate. Nevertheless, despite the substantial difference involved, we believe the pattern uncovered by Faro and Rottenstreich may translate to our setting. Our studies test this assumption and explore its implications.

Summary

As we have seen, people do not always rely on evaluations and predictions of utility to guide their behavior. In what follows, we present experiments that minimize the impact of social factors, also minimize uncertainty about payoffs and values, and tap generic knowledge. Situations defined by this set of characteristics are common; for instance, managers in large firms must often consider the ideal compensation package for their average employee. These are conditions under which people may be expected to draw on evaluations and predictions of utility and to do so in a way that puts the onus on potential predictive biases in \( p \) and \( w \).

If the predictive biases in \( p \) and \( w \) tend to be of opposite direction and of about the same magnitude, overall predictions of utility, \( U_o \), may frequently equal overall evaluations of utility, \( U_e \). Moreover, interventions that attenuate one predictive bias but not the other may harm overall predictive accuracy. Evaluations of delegation implicate only one predictive bias, concerning \( p \). The predictive bias in \( p \) is such that, relative to themselves, people expect others to be more optimistic about difficult tasks and less optimistic about easy tasks. Thus, principals may offer agents too little money to perform difficult tasks and too much money to perform easy tasks.

We test for all these possibilities in the studies below. Study 1 takes a first look at the accuracy or inaccuracy of predictions. Study 2 considers an emotionality intervention that may impact just one predictive bias, concerning \( w \), and looks at the implications of this intervention for overall accuracy. Study 3 focuses on the contrast between work and delegation.

Study 1: word puzzles

To create a setting in which people form evaluations or predict others’ evaluations of incentive pay, we presented participants with a puzzle game called Word Prospector (Burson, Larrick, & Klayman, 2006). Players in this game are shown a “source word” (e.g., troglodyte) and must generate new words (e.g., gold, rode, drool) using the letters of the source word. A goal is set for each puzzle that consists of a specific number of new words to generate, a minimum acceptable length for those words, and a time limit. Players “solve” (i.e., succeed at) a puzzle if they meet the goal. Because the goal can vary across puzzles, and because extracting new words is easier from some source words than others, the puzzles vary in difficulty.

We asked participants to imagine that they would receive $1000 for successful completion of each puzzle. This contingency allowed us to assess both the overall accuracy of participants’ predictions of other participants’ evaluations and the accuracy of each predictive substep. In particular, we had participants answer two questions about each puzzle. The first question concerned the likelihood of reaching the goal; participants either estimated their own chances or what they believed a randomly chosen participant would believe his or her own chances were. The second question required participants to indicate either a cash amount that would make them indifferent or a cash amount that would make a randomly chosen participant indifferent between (a) receiving that cash amount for sure or (b) taking the chance that they would receive $1000 for successful completion of the puzzle. In the framework of the model we have offered, these “cash equivalents” (CEs) can be assumed to satisfy \( v(CE) = v(x)w[p(T)] \).

Clearly, the likelihood judgments obtained from our first question correspond to one of the predictive substeps we have highlighted. The CEs obtained from our second question are positively correlated with either a participant’s own overall evaluations, \( U_o \), or a prediction of others’ overall evaluations, \( U_e \). Finally, using the CEs, we can infer a participant’s weighting or prediction of weighting (in a manner that we detail fully below).

We view the Word Prospector setup as capturing fundamental aspects of incentive pay environments. First, a prize is awarded for attainment of a goal. Second, whether the goal will be attained
is uncertain and will in part depend on the individuals who are performing the task and whose evaluations are predicted (though luck and other factors may also play a role). Thus, although we do not instantiate principal–agent relationships in this study, we believe critical dynamics relevant to principal–agent contexts are instantiated by our experimental setup.

We should also mention one methodological nuance. Extant studies of worse-than-average and better-than-average effects made a distinction between “direct” comparisons, in which people answered questions that explicitly contrasted themselves with others (e.g., “Compared to others, how likely are you to succeed on this task?”) and “indirect” comparisons, in which people estimated only their own or only someone else’s chances (Menon et al., 2009). In direct comparisons, both worse-than-average and better-than-average effects were typically robust. On the other hand, some studies using indirect comparisons reported robust better-than-average effects but attenuated worse-than-average effects (e.g., Kruger, 1999), whereas other studies reported the reverse (e.g., Moore & Small, 2007). Evidently, in indirect comparisons, specific features of different settings accentuate one pattern and diminish the other (cf. Moore & Healy, 2008). It is indirect comparisons that we pursue in Study 1 and our later studies.

Method

The study employed a 3 (assessment type: evaluation, prediction-of-other’s-likelihood-opinion, prediction-of-other’s-likelihood) × 9 (puzzle) mixed design. Assessment type was manipulated between participants. Participants in the evaluation condition estimated their own likelihood of success and chose CEs for themselves (to reiterate, CEs in this condition are a measure of U0). Participants in the prediction-of-other’s-likelihood-opinion condition provided the likelihood of success they thought would be estimated by a random other participant for him- or herself and also forecast this random other’s choice of CE. Participants in the prediction-of-other’s-likelihood condition provided their own estimate of the likelihood of success that a random other participant would have and predicted the random other’s choice of CE. Note that participants in the latter two conditions encountered exactly the same CE question (to reiterate, responses to this question are a measure of U0). Only the likelihood question differed for participants in these conditions: In one condition, participants estimated a representative participant’s perception of that participant’s chances of success; in the other condition, participants evaluated the chances of success of the representative participant. As we mentioned, we assumed that answers to these questions would be approximately equal, but we wanted to check this assumption. The nine different puzzles, which formed a within-subjects factor, were pretested to yield a wide variety of difficulty levels.

One hundred and ten University of Michigan undergraduates participated for course credit. For each puzzle, participants initially had a preliminary thirty seconds to work on the source word without knowing the goal set for it. We included a real rather than hypothetical incentive for this work: The participant with the best performance across these nine preliminary exposures received $50. After thirty seconds with each puzzle, we informed participants of the goal. At this juncture, they answered, first, the likelihood question about the puzzle and, second, the CE question. After answering these questions, participants did not continue working on the puzzle. Thus, participants never actually completed any of the puzzles.

We presented the CE question as a multipart query. Participants began by selecting a CE range ($1–$25, $30–$100, $110–$200, $210–$300, . . . , $910–$970, $975–$999). Then, for each of several cash amounts in this range, participants indicated whether they preferred receiving that amount for sure or taking the chance that they would receive $1000 for successfully completing the puzzle. The CE we attributed to each participant was the midpoint between the highest sure amount he or she preferred less than taking the chance at the puzzle and the lowest sure amount he or she preferred more than taking the chance. We excluded two participants from our analyses because their CEs across the different puzzles were entirely constant.

Results

We divide our presentation of results into three parts corresponding to likelihood, weighting, and a contrast of overall evaluations for the self with overall predictions of others.

Likelihood

A repeated measures ANOVA indicated that, as anticipated, participants provided very similar estimates in the prediction-of-other’s-likelihood-opinion condition and prediction-of-other’s-likelihood condition, producing no significant effect of this manipulation (Ps < .1, ps’ > .32). We thus collapsed these conditions in subsequent analyses.

Recall that for difficult puzzles of small p, we expected p < p; for easy puzzles of large p, we expected p > p; and we were likely to observe one of these patterns to be stronger than the other. Indeed, for the four most difficult puzzles, for which participants estimated that both they and others would have less than a 50% chance of completing, Table 1a shows a consistent trend: p is less than p for petroglyph, gargantuan, and cummervund, t’s > 2.44 and p’s < .009; for troglodyte, t = 1.23, and p = .110.1 The next four puzzles in Table 1a, overthrown through fearlessly, were easier, and all elicited likelihood estimates that were essentially identical for self and other (t’s < .23, p’s > .41). Finally, the easiest puzzle of all, stoplights, showed a trend of lower likelihood estimates for other than for self (t(1, 106) = 1.53, p = .065). In sum, it appears that even for the indirect judgments we considered, p diverged from p in the manner predicted.

Weighting

We surmised that forecasting of weighting would be biased in a manner opposite that of likelihood. To assess this possibility, we fit our CE data using the model U0(x, T) = U0(x)w(p, [T]) and U0(x, T) = U0(x)w(p0, [T]). To introduce our fitting procedure (note that because we are free to assume w(0) = 0 and w(1) = 1): The definition of CEs implies that the following equality will hold for every puzzle for every participant: v(CE) = v(1000)[w(estimated probability of success)]. Rearranging terms to isolate w, we have:

w = v(CE)/v(1000).

Moreover, because each participant provided responses for nine different puzzles, forming specific assumptions about the functional forms of v and w allows us to estimate every participant’s weighting function, w, or his or her prediction of a random other’s weighting function, w. We can then examine whether for different puzzles, in which p, and p are both small (and p < p), our estimation yields w > w, and whether for easy puzzles, in which p, and p are both large (and p > p), our estimation yields w < w.

We selected forms for v and w laid out in a seminal paper by Prelec (1998). Prelec suggested a power value function for money, v(x) = xβ and offered w(p) = exp{−(−ln p)γ} as a one-parameter specification of w. A weighting parameter, β, close to 1 reflects relatively linear weighting of probability; when β is exactly equal to 1,
w = p. On the other hand, β < 1 induces a curvilinear function that overweights small probabilities and underweights large probabilities; this over- and underweighting becomes more pronounced as β draws closer to 0. The hypothesis of ws > w.o for difficult tasks of small probability and ws < w.o for easy tasks of large probability thus reduces to the notion that β will be higher for ws than for w.o.

Substituting Prelec’s functional forms into Eq. (1) and rearranging terms yields:

$$-\ln(-\ln(CE/\$1000)) = \ln(x) + \beta(-\ln(-\ln(estimated\ probability\ of\ success)))$$

Eq. (2) is handy because it allows us to use OLS regression to estimate each participant’s x and β. If w.o is less extreme than w.s, the OLS coefficient for (what may be called) β.s should be larger than that of β.w.

We obtained OLS estimates of x and β for each participant and then used a MANOVA to examine the impact of self vs. other on these parameters. Importantly, we found no effect on the value parameter x (F5 < .04, p > .85), indicating that participants did not view the attractiveness of $1000 differently than they predicted others would (x.s = .91 and x.w = .89). However, there was a significant main effect on β (F(1,106) = 8.76, p < .004): Participants predicted less curvilinear weighting for others than they showed themselves; mean β.s was equal to .68, whereas mean β.w was only .39. In short, predictions of w.s appeared to be biased, and in a manner opposite to that of p.o. To reiterate, this finding is of value in and of itself and apart from our opposing biases analysis in that it reveals that prediction errors known to occur with specified, numerical chances generalize to uncertain events.

Cash equivalents

Because we observed opposing biases for p and w, overall predictions may be either on target or off target, depending on whether both errors occurred and whether they happened to reconcile. Table 1a reveals a number of interesting patterns.

The four hardest puzzles, which participants estimated both they and others would have less than a 50% chance of successfully solving, are essentially exactly balanced the extent to which the distinct biases, if they are nonzero, match in magnitude and offset. Thus, to the extent that the distinct biases, if they are nonzero, match in magnitude and offset.

It is interesting that the two puzzles of intermediate difficulty engender little error in either step of the prediction process. This observation is consistent with the notion that the bias in each of p and w (e.g., the value of the quantity p.s – p.o) changes direction as we move from difficult to easy tasks. Such biases will almost always be more pronounced at the endpoints of the relevant scale and diminutive at its center. Thus, to the extent that predictive biases follow the form outlined by our analysis, it follows that predictions of others’ beliefs, weighting, and overall evaluations may all be more accurate for events of middling likelihood than for improbable or highly probable events.

The two next most difficult puzzles, overthrown and management, both of which elicit likelihood estimates suggesting an intermediate level of difficulty, yield predictions of CE measures that appear relatively accurate (t’s < .65, ns). Furthermore, rechecking Table 1a, note that these two puzzles yield little error in predictions of likelihood either; for both puzzles, p.s and p.o are virtually identical. Evidently, these puzzles of intermediate difficulty engender accuracy at both steps and therefore relative accuracy of overall predictions.

The next two puzzles, typewriter and fearlessly, also yield essentially identical p.s and p.o; for these puzzles, CE.s and CE.o do not differ significantly. Finally, stoplights appears to reveal accuracy via offsetting biases. Here, there is some evidence that p.o exceeds p.s (t(1,106) = 1.53, p = .065), but CE.s and CE.o are essentially equivalent (t(1,106) = -.07, p = .477). Evidently, for this single puzzle we have precisely offsetting errors, as w.s > w.o to a degree that essentially exactly balances the extent to which p.s > p.o.

Discussion

The results corroborate our analysis by showing that predictions can be decomposed into substeps involving likelihood judgment and weighting and that these substeps may yield opposing biases. The accuracy of overall predictions thus depends on the extent to which the distinct biases, if they are nonzero, match in magnitude and offset.

It is noteworthy that Study 1 reveals a divergence between difficult and easy puzzles. Overall inaccuracy arises in both, despite the fact that errors in p.o are more pronounced for the difficult puzzles of small probability. The co-occurrence of diverse error magnitudes for p.s with consistent error magnitudes for overall predictions highlights the importance of understanding the interplay between distinct predictive substeps. To illustrate, suppose that a manager vastly overestimates employees’ confidence in their ability to successfully complete some project, estimating that it is 25% when it is actually 10%. A straightforward mapping from beliefs to overall predictions would suggest that the manager will also overestimate employees’ taste for a risky compensation program.
scheme tying rewards to performance. However, our analysis suggests that because the manager will tend to underestimate the weight that employees will place on 25%, his or her overall prediction of employees’ taste for risky compensation will be less biased than the 25% likelihood judgment suggests.

So far we have taken for granted that people construct evaluations and predictions via processes that include the particular substeps of likelihood judgment and weighting. The method of Study 1 essentially forces participants into these substeps. Accordingly, we ran a follow-up to Study 1 designed to ascertain whether, in the contexts we studied, people left to their own devices do or do not respond in a manner consistent with the substeps of likelihood judgment and weighting.

We had 102 New York University undergraduates get acquainted with Word Prospector for one minute by extracting new words from the source word management. As before, a between-subjects manipulation placed each participant in either an evaluation or prediction condition. Participants provided responses concerning a puzzle based on the source word troglodyte, one of the difficult source words of Study 1, and then provided responses concerning a puzzle based on fearlessly, one of the easier source words of Study 1. For each puzzle, participants spent 1 min attempting to extract new words from the source word. As in Study 1, we asked participants to imagine that successful completion of the puzzle would result in a reward of $1000 and that their CE responses should correspond to a sure amount that would be just as attractive.

The key to this follow-up was that in both the evaluation and prediction conditions, only about half of the participants provided a likelihood response followed by a CE response for each puzzle, as in Study 1. In contrast, the remaining participants provided their CE response for a puzzle without having previously provided a likelihood response for that puzzle. In other words, some participants were, as in Study 1, essentially forced into two steps of the form we assumed. In contrast, half of the participants were left to their own devices.

As Table 1b shows, in both the evaluation and prediction conditions, CE measures are about the same whether or not participants were funneled into a two-step process. Troglodyte elicits a mean CE of $225 after an explicit likelihood judgment and a mean CE of $224 in the absence of an explicit likelihood judgment (t(1, 51) = .02, p = .494). The CE responses for troglodyte are also quite similar, $164 and $178 (t(1, 45) = .27, p = .396). A qualitatively identical pattern emerges for fearlessly (t’s < .73, p’s > .23). These observations buttress the notion that, at least in the context we study, people left to their own devices construct evaluations and predictions in a manner consistent with the model we have invoked.

Table 1b
Mean likelihoods and cash equivalents by source word and condition for the follow-up to Study 1.

<table>
<thead>
<tr>
<th>Source word</th>
<th>Condition</th>
<th>Likelihood responses</th>
<th>Difference (self-prediction) (%)</th>
<th>CE responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Self (%)</td>
<td>Prediction (%)</td>
<td></td>
</tr>
<tr>
<td>Troglodyte</td>
<td>Probability then CE only</td>
<td>22</td>
<td>23</td>
<td>$225</td>
</tr>
<tr>
<td>Fearlessly</td>
<td>Probability then CE only</td>
<td>76</td>
<td>69</td>
<td>$660</td>
</tr>
</tbody>
</table>

Note that though neither evaluations nor predictions vary much with the presence or absence of two explicit steps, CE responses are largely inaccurate in either case. Yet predictions were relatively accurate for both troglodyte and fearlessly in Study 1. If accuracy vs. inaccuracy is in some sense the by-product of two biased processes that may either match or mismatch, we would indeed expect that seemingly slight changes in procedure (the amount of practice time provided for each puzzle, the number of puzzles, etc.), participants, and the like could frequently impact the extent to which predictions happen to be on or off target.

Study 2: rational and emotional decisions

We have suggested that predictions are accurate when two opposing predictive biases are either both absent or both present to the same degree so that they offset. This analysis implies that an intervention that influences just one of the substeps susceptible to bias can sometimes improve but also sometimes worsen overall accuracy. The impact of the intervention will depend on the balance that obtains prior to the intervention. For instance, if both substeps engender errors, but these errors offset absent intervention, then debiasing at just one step will transform overall accuracy into overall inaccuracy. This observation allows us to further test our analysis by manipulating the choices that participants make in ways that can be expected to either diminish or enlarge errors in forecasts of weighting.

In particular, in a manner similar to past research (Kopelman, Rosette, & Thompson, 2006; Susskind & Cruikshank, 1987; Thompson, Medvec, Seiden, & Kopelman, 2001; see also Pham, 2007), we asked some participants to base their evaluations on rational (i.e., nonemotional) considerations and other participants to base their evaluations on emotional considerations. Past work has suggested that, compared to a control condition, evaluators following rational considerations will reveal less pronounced over- and underweighting (Faro & Rottenstreich, 2006; Hsee & Weber, 1997; Rottenstreich & Hsee, 2001). On the other hand, evaluators following emotional considerations will reveal even more pronounced over- and underweighting. Furthermore, research on empathy gaps suggests that predictors will not be as sensitive to emotional or rational cues experienced by others to nearly the degree those others are (Loewenstein, 1996). Predictors in the emotional and rational conditions may thus be expected to form predictions that are highly similar to predictions made by control participants.

As a result, given rationalistic evaluations (because predictors’ inability to foresee overweighting and underweighting by others will not be as large an error as before), the weighting prediction bias should be diminished. In contrast, given emotional evaluations (because predictors’ inability to adequately foresee overweighting and underweighting by others will now be an even larger error), the weighting prediction bias should be increased.

Overall accuracy should then vary in predictable, though intricate, ways across conditions. Suppose that in the control condition $CE_S - CE_0$ is equal to some number $k$. The rational condition should show $CES - CE_0 < k$ for small probabilities and $CES - CE_0 > k$ for large probabilities, mirroring the bias revealed by likelihood forecasts. Put another way, a reduced weighting error in the rational condition implies that overall predictions in the rational condition will largely reflect the likelihood error. The emotional condition, on the other hand, should show the reverse pattern, $CE_S - CE_0 > k$ for
small probabilities and $CE_3 - CE_2 < k$ for large probabilities. In other words, an increased weighting error in the emotional condition may overcorrect for the extant likelihood error, so the inaccuracy of overall predictions in the emotional condition will be highly consistent with the bias revealed by predictions of weighting.

Directly comparing the emotional and rational conditions can neatly summarize much of the above. For difficult puzzles of small probability, the difference $CE_3 - CE_2$ should be larger (e.g., positive) in the emotional condition and smaller (e.g., less positive or negative) in the rational condition. On the other hand, for easy puzzles of large probability, $CE_3 - CE_2$ should be smaller (e.g., negative) in the emotional condition and larger (e.g., less negative or positive) in the rational condition.

Method

The study followed a 2 (assessment type: evaluation vs. prediction) × 4 (puzzle) × 3 (assessment basis: control vs. emotional vs. rational) mixed design. In a complement to Study 1 and its follow-up, which manipulated target between-participants, we this time manipulated assessment type within-participants; each individual in the study made likelihood and CE judgments for his- or herself and also predicted these responses for a randomly selected participant. Each participant encountered four puzzles; two were relatively easy, and two were relatively difficult. Finally, decision basis was varied between participants in a manner detailed below.

One hundred University of Michigan undergraduates took part for course credit. They began by spending thirty seconds attempting to complete a practice puzzle involving the source word man-agement. Thirty seconds was not nearly enough time to solve the puzzle—the practice was intended only to familiarize participants with Word Prospector. Next, participants encountered a randomly ordered sequence of four puzzles. The four puzzles, concerning the source words petroglyph, troglodyte, fearlessly, and typewriter, were the two most difficult and two easiest words from Study 1. Unlike in Study 1 and its follow-up, we did not ask participants to spend time on any of these four puzzles; they merely answered the likelihood and CE questions (the hypothetical prize for solving each puzzle was again $1000). All participants first indicated their own responses to both questions for all four puzzles, then engaged in a 10-min filler task, and finally predicted the responses of a “randomly selected participant.”

Participants in the control condition completed the experiment without additional instructions, as in Study 1. Participants in the rational condition read the following instructions after providing their likelihood judgment and before choosing a CE:

<table>
<thead>
<tr>
<th>Source word</th>
<th>Condition</th>
<th>Likelihood responses</th>
<th>CE responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Self (%)</td>
<td>Prediction (%)</td>
</tr>
<tr>
<td>Petroglyph</td>
<td>Control</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Emotional</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Rational</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Troglodyte</td>
<td>Control</td>
<td>24</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Emotional</td>
<td>26</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Rational</td>
<td>27</td>
<td>35</td>
</tr>
<tr>
<td>Fearlessly</td>
<td>Control</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Emotional</td>
<td>79</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Rational</td>
<td>76</td>
<td>81</td>
</tr>
<tr>
<td>Typewriter</td>
<td>Control</td>
<td>81</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>Emotional</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>Rational</td>
<td>82</td>
<td>82</td>
</tr>
</tbody>
</table>
not significantly different from zero for the two easy puzzles ($t's < 1.04$, $p's > .15$). Indeed, when the two difficult puzzles and the two easy puzzles were taken together, a repeated measures ANOVA on the within-participant measure $p_i - p_o$ revealed a main effect of difficulty; difficult puzzles yielded greater likelihoods for predictions than evaluations ($M = -.53$), but easy words did not ($M = 0.00$, $F(1, 97) = 7.66$, $p = .007$). Moreover, the correlation of $p_i$ and $p_i - p_o$ was significant in all three conditions, equal to .40 ($p = .006$), .27 ($p = .061$), and .40 ($p = .013$) in the control, emotional, and rational conditions, respectively.

In sum, small probabilities tended to show $p_i < p_o$, whereas large probabilities tended to show either $p_i > p_o$ or $p_i = p_o$ just as in Study 1.

### Weighting

Because each participant provided only four CES responses and CEo responses, we did not have sufficient power to fit these responses using Prelec's functional forms. However, the within-participants method of this study provides an alternative approach for examining weighting.

If errors in $p_o$ and $w_o$ are of opposite directions, then the within-participant difference $p_i - p_o$ and the within-participant difference $w_i - w_o$ should be negatively correlated. The more a participant believes that another's likelihood of success is smaller than his or her own, the more he or she should overestimate the other's weighting relative to the participant's own. To assess this possibility, we constructed a crude measure of weight by dividing CE responses by their corresponding likelihoods. Collapsing across all four words and all three assessment-basis conditions, the resulting correlation between $p_i - p_o$ and the constructed measure of weighting difference was negative $r(351) = -.22$, $p < .001$. Of course, we additionally expected this correlation to differ by assessment basis. Indeed, the correlation was strong in the control condition ($r(130) = -.23$, $p = .009$), nonsignificant in the rational condition ($r(102) = -.16$, $p = .100$), and strongest in the emotional condition ($r(119) = -.34$, $p < .001$).

### Cash equivalents

To test the predicted pattern of CES, we examined the difference $CE - CEo$ for each participant. We observed an interaction of difficulty and assessment basis for this measure ($F(2, 97) = 5.10$, $p = .008$). As anticipated, for difficult puzzles, $CE - CEo$ tended to be smallest (negative) in the rational condition, intermediate in the control condition, and largest (positive) in the emotional condition. On the other hand, for easy puzzles, $CE - CEo$ tended to be smallest in the emotional condition, intermediate in the control condition, and largest in the rational condition. Indeed, for difficult puzzles, the difference $CE - CEo$ was significantly greater given rational rather than emotional instructions ($F(1, 97) = 6.26$, $p = .005$), but the reverse was true for easy puzzles ($F(1, 97) = 2.60$, $p = .055$).

This is precisely the trend one would anticipate if (a) rationalistic instructions attenuated the weighting forecast error and (b) emotional instructions enhanced the weighting forecast error. Relative to the control condition, our rationalistic intervention yielded an overall bias of the same direction as the error that prevailed in likelihood judgment. Put differently, the weighting forecast error may no longer override the likelihood forecast error as effectively as it did in the control condition, so that overall predictions themselves reflected the likelihood forecast error. Paradoxically, debiasing the weighting forecast error led to a systematic overall bias. Furthermore, relative to the control condition, our emotional intervention yielded an overall bias that was highly directionally consistent with the error only in weighting. In other words, magnifying the bias in forecasts of weighting may have engendered overall predictions that overcorrected for the likelihood forecast error. Magnifying the weighting forecast bias, just like debiasing the weighting forecast bias, could lead to a systematic overall bias.

Note that though the control condition was relatively close to achieving accuracy, accuracy did not arise consistently in the other two conditions. To some extent, any real-world situation (and our control condition, too) may be viewed as tapping some mix of rational and emotional cues (or lack thereof) that combine to engender some particular degree of likelihood forecast bias and weighting forecast bias. We speculate, then, that accuracy of predictions may tend to arise in the presence of specific combinations of rationality and emotion that happen to produce offsetting biases. Any situation that engenders mixes of rationality and emotion lying outside these specific combinations may frequently yield inaccuracy of overall predictions.

### Study 3: contrasting assessments of work and delegation in principal–agent settings

We now shift focus from the determinants of predictive accuracy to a contrast of work and delegation. Recall that the evaluation of delegation, $U_d = n(x)w|p_o/T|$, implicates $p_o$ but $w_o$ and is thus susceptible to a single predictive bias that will not be offset. As a result, even in an environment where $U_d = U_o$, $U_d$ can be expected to diverge from both these measures. In particular, with easy tasks characterized by $p_o > p_i$ and $w_i > w_o$, we should observe $U_d < U_o$ ($U_d < U_o$ because $p_o < p_i$; $U_d < U_o$ because $w_i > w_o$). With difficult tasks characterized by $p_o > p_i$ and $w_i < w_o$, we should observe the reverse profile, $U_d > U_o$. We tested for this possibility in Study 3, again using Word Prospector puzzles.

### Method

The study followed a 3 (assessment type: principal-working vs. prediction-of-principal-working vs. principal-delegating) × 4 (puzzle) mixed design. Participants were 106 undergraduate students from the University of Michigan, who completed the experiment for course credit. We largely followed the methodology of Study 2, providing all participants with four puzzles (i.e., puzzle was a within-participants factor), two of which were hard and two of which were easy, using a practice word, etc. However, we departed from the methodology of Study 2 in one respect, by allowing participants to spend 30 s working on each of the four puzzles, as in Study 1.

Assessment type was varied between participants. About one third of the participants provided likelihood and CE judgments for working for themselves on the puzzles (the hypothetical reward they were asked to consider was again $1000 per puzzle). Another third predicted the responses of a randomly selected participant who was working for him or herself on the puzzles. The remaining third were placed in the role of a principal delegating the work to a randomly chosen participant as their agent. They predicted a random other’s likelihood of success and provided a cash amount that made them indifferent between (a) receiving that amount for sure and (b) having the chance to receive $1000 for each puzzle their agent completed successfully.

Note that the likelihood questions posed to predictors and delegators were not identical. Predictors were asked what a random other would deem his or her own likelihood of success to be, whereas delegators were asked what they deemed a random others’ likelihood of success to be. As Study 1 indicated, we can expect these different likelihood questions to elicit essentially equivalent responses.

### Results

#### Likelihood

As anticipated, likelihood judgments were largely equivalent across predictors and delegators ($F's < .26$, $p's > .85$); we thus
collapsed these responses in subsequent analyses. To obtain further statistical tests, we lumped together the two easy puzzles and also lumped together the two difficult puzzles. A repeated measures ANOVA then showed a significant interaction between assessment type and puzzle difficulty ($F(1, 104) = 6.47, p = .012$) of the form we previously observed (see Table 3). Contrasts revealed significant differences for both difficult puzzles ($F(1, 104) = 3.37, p = .035$) and easy puzzles ($F(1, 104) = 2.78, p = .05$).

**Cash equivalents**

A repeated measures ANOVA confirmed that difficult puzzles produced lower CE responses than easy puzzles ($F(1, 103) = 184.72, p < .001$) and also yielded a significant assessment type by difficulty interaction ($F(2, 103) = 5.47, p = .006$). More importantly, we observed two critical patterns. First, $CE_p$ and $CE_o$ were very similar ($F(1, 103) = 4.09, p = .023$) but significantly smaller than the remaining CE measures for difficult puzzles ($F(1, 103) = 7.95, p = .003$).

**Discussion**

Consistent with Studies 1 and 2, the results corroborate the notion that the combination of two oppositely biased, predictive substeps determines whether a prediction is on target. Predictors and delegators show the same pattern of likelihood judgment. Nevertheless, predictors and delegators show different patterns for the CE measure. By inference, the different CE responses must be driven by differences in weighting. In particular, it appears that predictors’ biased forecasts of weighting correct for their biased forecasts of likelihood but that delegators’ weighting does not counteract their biased forecasts of likelihood. Indeed, participants who considered working on the puzzles themselves provided different CE responses than delegators. Moreover, the difference in the CE responses of principals working for themselves and delegators is consistent with these participants’ distinct patterns of likelihood judgment. By inference, delegators’ weighting must be equivalent to that of principals working for themselves.

A consequence of delegators’ mixing likelihood judgment of others with weighting for the self is that even though we observe accurate overall predictions (of the evaluation of someone working for him or herself), we nevertheless observe evaluations of delegation that reflect a likelihood bias. In some sense, then, predictions incorporate two flaws that may cancel out, whereas assessments of delegation incorporate a single flaw that is not canceled out and instead rears its head.

In our experimental context, delegators and those working for them are drawn from the same population, and thus by construction possess the same average level of ability. Moreover, although principals often hire agents because of the agent’s greater expertise (as when homeowners hire an interior decorator), cases in which principals and agents are equally skilled are also common. For instance, a manager and employees at a large pharmaceutical company might have similar skills with respect to some research problem at hand. Indeed, in many circumstances, individuals may differ in tenure or seniority, with these factors determining who is the principal and who is the agent, yet possess largely equivalent skills and abilities.

Our findings suggest that when principals and agents are on average equally able, principals will show the following pattern. First, when principals choose whether to perform an easy task on their own or to hire an agent, their likelihood forecasting error will lead them to believe that the agent’s chances of success are lower than their own, and their weighting will not correct for this misperception. As a result, principals may end up underinvesting in agents. Second, when principals decide about a difficult task, their likelihood forecasting error will lead them to believe that the agent’s chances of success are higher than their own, and their weighting will not correct for this misperception either. As a result, principals may end up overinvesting in agents. In sum, if delegation decisions are biased in the way our experiments suggest, principals will underinvest in agents for easy tasks and overinvest in agents for difficult tasks.

Relatedly, recall that our results can be interpreted as revealing a framing effect: Viewing potential benefits as stemming from work that an individual performs vs. work that he or she delegates evidently changes the perceived utility of these benefits. Such a change is inappropriate; the appraisal of potential benefits should not depend on whether an individual performs the relevant work or delegates it.

Importantly, a straightforward debiasing technique is available. Suppose the principal will receive reward $x$ if the task is successfully completed. When deciding whether or not to use an agent, the principal should not ask the $UB$ question, “What is my utility for using an agent to try to complete the task and thus receive $x$?” Instead, the principal should ask the $UB$ question, “What would the agent’s utility be for working on the task if I offered him a payment of $x$, given success?” The latter question concerns work rather than delegation and thus forces predictions at both substeps.

**General discussion**

We have examined evaluations of incentive pay and predictions of such evaluations, arguing that two critical predictive substeps, involving likelihood judgment and weighting, tend to show opposing biases. The accuracy of overall predictions may thus depend on whether the biases at the two substeps happen to match in magnitude so that they offset. In addition, we noted that evaluations of delegation include a single predictive bias, at the likelihood step, that will not be offset. As a result, even settings that yield accurate predictions may yield flawed evaluations of delegation.

An *ABC* mnemonic summarizes our analysis: Accurate predictions of work prospects can arise from two offsetting *Biases*, but a single, unopposed bias underlies delegation prospects and thus creates a Contrast between assessments of work and delegation.

We would like to reiterate that our experiments have focused on situations bearing a critical set of characteristics. In particular, we studied situations that minimize the impact of social factors, essentially eliminate uncertainty about payoffs and values, and tap generic knowledge. Such situations are common—for instance, managers in large firms must often consider the ideal compensation package for their average employee. Alternative empirical patterns may emerge in situations bearing other characteristics, for instance, where people make predictions concerning others with whom they are highly familiar (cf. Menon et al., 2009).

Variability in behavior could also occur across situations that share the set of characteristics we have identified. For instance,

<table>
<thead>
<tr>
<th>Source word</th>
<th>Likelihood responses</th>
<th>CE responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self (difference self-prediction) (%)</td>
<td>Self (difference self-agent) (%)</td>
</tr>
<tr>
<td>Petroglyph</td>
<td>4 –3 –2</td>
<td>$140$ –7</td>
</tr>
<tr>
<td>Troglodyte</td>
<td>17 –9 –11</td>
<td>$192$ –43</td>
</tr>
<tr>
<td>Fearlessly</td>
<td>75 8 9</td>
<td>$558$ –25</td>
</tr>
<tr>
<td>Typewriter</td>
<td>84 3 1</td>
<td>$645$ –70</td>
</tr>
</tbody>
</table>
as we have mentioned, any real-world situation may be viewed as tapping some mix of rational and emotional cues (or lack thereof) that combine to engender a particular degree of likelihood forecast bias and weighting forecast bias. Overall accuracy of predictions may thus require the presence of specific combinations of rationality and emotion that happen to produce offsetting biases. Overall inaccuracy would arise given other combinations of rationality and emotion.

The balance struck between reliance on stereotypes and reliance on projection could also be an important determinant of behavior in the types of situations we have studied. People frequently predict others’ behavior by invoking stereotypes of the groups to which these others belong. For instance, Weber and Hsee (1998) found that both Americans and Chinese predicted that Americans would be less risk averse than Chinese. In actuality, in Weber and Hsee’s study, the opposite was true: Americans were more risk averse than Chinese. Weber and Hsee positied that national stereotypes, say, of Americans as free and adventurous, led to these mispredictions. Though the use of stereotypes is common, it is also typical for people to make predictions of others by projecting their own evaluations onto these others. For instance, Davis, Hoch, and Ragsdale (1986) proposed an anchoring and adjustment process whereby people anchor on their own evaluations and then arrive at a prediction of others by attempting (but somewhat failing) to adjust for ways in which others may differ from them. Such anchoring-on-the-self models (see, e.g., Hoch, 1987) are frequently invoked to explain the well-known false consensus effect (Ross, Greene, & House, 1977). It is not clear whether participants in our experiments relied on stereotypes, projection, some combination of the two, or alternative processes. But different situations may clearly engender differential reliance on these distinct approaches. Refining our understanding of the relationship between stereotypes vs. projection, on the one hand, and forecasts of likelihood and weighting, on the other, represents an important direction for future work.

Mappings from beliefs to preferences

It is common to assume that one can infer an individual’s beliefs from his behavior. For instance, if someone places a higher CE on one word puzzle than another, it seems reasonable to conclude that he is doing so because the first puzzle seems easier to solve. By the same token, it is also commonly understood that there are conditions under which inferences from behavior to belief must be limited. A person who bets on the Oakland Raiders to win the Super Bowl may do so because he or she deems the Raiders likely to win but may also do so because he or she likes the Raiders and enjoys betting on them.

Our work highlights the value of understanding such limiting conditions, whether the relevant context involves predictions of others or implicates only evaluations formed for oneself. “Straightforward mappings” often oversimplify the situation and may miss important implications of processes that include two substeps. For instance, in an influential paper, Camerer and Lovallo (1999) suggested that over-entry into competitive industries can be attributed, in part, to overly optimistic assessments of one’s likelihood of success. However, as we have seen, it is exactly when people are excessively positive about comparative likelihood that they tend to be most sober in weighting. Thus, in many settings, optimism may not correlate much with observed behavior.

The interplay of likelihood judgment and weighting may also be important when one considers inferences in the other direction, from belief to behavior. Weinstein (1980) reported that many people show “self-positivity,” viewing themselves as less likely than others to experience unfortunate life events (e.g., contracting a sexually transmitted disease). Following Weinstein, many researchers have argued that if people believe they are personally less at risk in this way, they will tend to ignore health messages and fail to take preventative measures (e.g., Menon, Block, & Ramathan, 2002; Perloff & Fetzer, 1986; Raghuvir & Menon, 1998). By noting the relevance of weighting, our work may qualify such conclusions. Simply believing that one is less likely than others to contract a sexually transmitted disease does not imply that one will take relatively few precautions.

Similar issues may be raised in reaction to discussions of better-than-average and worse-than-average effects such as expressed by Windschitl et al. (2003). These authors wrote: “It is reasonable to assume that the misguided optimism or pessimism . . . can mediate a variety of consequential decisions and behaviors. Examples . . . include: whether or not a person decides to engage in a competition, the amount of effort and resources a person invests in an outcome, the strategy used to achieve the outcome, and a person’s anxiety about and actual performance in a competition” (see also Camerer & Lovallo, 1999; Kruger & Burkus, 2004; Moore & Small, 2007). However, supposing that misunderstandings of likelihood will necessarily translate into the domain of preferences is exactly the sort of straightforward mapping that may not hold in a dual-process model of likelihood and weighting.

Conclusion

Our work was in part inspired by a prominent opposing biases analysis put forth by Kahneman and Lovallo (1993). These authors examined managers’ tendencies toward both overconfidence and risk aversion, noting that “bold forecasts and timid attitudes to risk . . . have opposite effects.” They cautioned that unfortunate managerial consequences could sometimes ensue if one error was reduced while the other was left unaffected, and they discussed the importance of pairing prescriptions in one domain with prescriptions from the other. They urged decision analysts to be highly sensitive to the interplay of the two domains.

We offer a similar thesis concerning principal–agent interactions and related settings, in which behavior depends on the evaluation of work and delegation prospects and on the prediction of others’ evaluations thereof. Lack of awareness of the interplay between forecasts of likelihood judgment and forecasts of weighting may lead to trouble. For instance, it is entirely possible that a manager relying on his or her own natural tendencies would accurately anticipate employees’ preferences between salaries and contingent pay. If this same manager then made very highly successful efforts to improve his or her understanding of employees’ likelihood judgments but maintained only an intuitive or default understanding of weighting, the manager would subsequently fail to accurately anticipate employees’ preferences. Decision analysis must be highly sensitive to the implicit interplay of belief prediction and weighting prediction.

Acknowledgment

We are grateful to William Bottom, Craig Fox, Joshua Klayman, Justin Kruger, Richard Larrick, Alex Markle, and Jack Soll for helpful comments and Rich Gonzalez and George Wu for statistical advice.

References


