

A timidity error in evaluations: Evaluators judge others to be too risk averse

George Wu,^{a,*} Chip Heath,^b and Marc Knez^c

^a Graduate School of Business, Center for Decision Research, University of Chicago, 1101 E. 58th St., Chicago, IL 60637, USA

^b Graduate School of Business, Stanford University, Stanford, CA, USA

^c Sibson Consulting, 101 N. Wacker Dr., Chicago, IL 60606, USA

Abstract

Managers often lament that their employees are risk averse and do not take sufficient risks. While in some instances employees might in fact be too risk averse, we explore situations in which managers may incorrectly judge their employees to be overly risk averse or timid. In two studies, we find evidence of a *timidity error* in evaluations—evaluators judge target decision makers to be risk averse, even when the targets are actually employing a more thoughtful approach (as measured by better calibration) than their evaluators.

© 2003 Elsevier Science (USA). All rights reserved.

1. Introduction

Managers often face the difficult task of inducing employees to take the appropriate level of risk in their decisions. There is ample evidence in the psychological literature that individuals have conflicting tendencies that may lead them to be too bold in some situations and too timid in others (Kahneman & Lovallo, 1993). The academic literature on risk has responded to this managerial dilemma by suggesting tools that managers can use to overcome the overconfident forecasts that might make employees too bold (Heath, Larrick, & Klayman, 1998) or the aversion to losses that might make employees too timid (Shapira, 1994; Tushman & O'Reilly, 1997). However, in order to apply such tools effectively, managers must first accurately evaluate the level of risk that their employees are taking. If managers make incorrect evaluations, the steps they take to manage risk-taking may fail or even backfire.

While we share the view that employees will often be too bold or too timid, we question whether managers are always accurate when they evaluate the level of risk their employees are taking. We suggest that evaluators may sometimes incorrectly judge their employees to be

too risk averse or “timid.” We focus on the *timidity error* because it may help explain why evaluators often judge other decision makers to be overly timid. Popular accounts of organizational life generally lament how conservative employees are, when faced with important organizational changes or new innovations (Hamel, 2000; Kotter, 1996). Indeed, when we polled a range of managers from various companies and industries, we found that managers frequently judged their employees to be too risk averse. We asked participants ($N = 123$) in an executive MBA course on managerial decision making to rate “the level of risk your direct reports are willing to take” on a scale ranging from 1 (*take much too little risk*) to 5 (*take the appropriate amount of risk*) to 9 (*take much too much risk*). Participants had an average of 17 years of work experience and 7.2 direct reports. On average, participants indicated that their reports took too little risk ($M = 4.42$, differs from scale midpoint, $t(122) = 4.34$, $p < .0001$).

There are reasons to believe that employees may indeed be too timid—standard models of choice assume that agents are risk averse, while research on loss aversion suggests that agents often behave timidly (Kahneman & Lovallo, 1993). However, we suspect that the evaluation process itself might sometimes lead evaluators to conclude incorrectly that their employees are inappropriately risk averse. In the next section, we

* Corresponding author.

E-mail address: george.wu@gsb.uchicago.edu (G. Wu).

describe more fully when and why a timidity error may arise when one person evaluates another. Before doing so, we motivate our argument by presenting two examples that suggest that the timidity error may occur in at least some important evaluation situations.

Our first example is historical. A few months into the Civil War, Abraham Lincoln faced a formidable set of obstacles. He had not yet had time to build an effective army, his general-in-chief was still recovering from typhoid, and even his own party was divided on how vigorously the war should be pursued and how much Southerners should be punished for their rebellion. Despite these obstacles to quick action, which seem rather obvious, Lincoln was widely criticized at the time because observers felt he was moving too slowly. Even his own Attorney General expressed grave doubts: “The President is an excellent man, and, in the main wise, but he lacks *will* and *purpose*, and, I greatly fear he has *not the power to command*” (italics original, Donald, 1995, p. 328). According to the Pulitzer Prize-winning historian, David Donald, this summary “represented a widely held opinion.” A Senator from Maine remarked, “If the President had his wife’s *will* and would use it rightly, our affairs would look much better” (p. 331). Instead of recognizing the difficulty of Lincoln’s task, his evaluators accused him of lacking “will.”

Consider another example from a business context. Bhide (1992, 1994) has argued that successful entrepreneurs rarely start with the business strategy that eventually allows them to succeed. Instead they pursue a “try-it, fix-it” approach and they modify their initial strategy in response to feedback from the environment. Bhide argues that entrepreneurs may want to avoid financial support from venture capitalists when they are initially building their firm because venture capitalists may reduce their ability to modify their strategy in response to new contingencies. Venture capitalists, it seems, do not like it when entrepreneurs respond to contingencies by changing their initial strategy. Frequently, when entrepreneurs make such changes, the venture capitalists question their motivation. In some cases, venture capitalists have been known to fire the entrepreneur and hire another person “with the guts” to pursue the original strategy (Bhide, 1997).

Both Lincoln’s critics and the venture capitalists may have been mistaken in their accusations. Both sets of evaluators accused a decision maker of being too timid—lacking the “will” or “guts” to pursue a risky course of action. We argue that these judgments are probably inappropriate, and that they hint at a more general problem that evaluators face when they try to evaluate the risky actions of others. Lincoln may have been indecisive, not because he lacked “will,” but because he correctly appreciated the delicate tradeoffs he faced in pursuing a war with an unprepared army and divided political support. Entrepreneurs may decide to switch

strategies, not because they “lack guts” to pursue the original plan, but because they acquire information that their original plan was flawed. In both cases, evaluators attributed timidity to decision makers in situations in which the decision makers may have actually employed a more thoughtful and well-reasoned approach than their evaluators.

Although suggestive, these real world examples are difficult to interpret cleanly because of hindsight biases and also because evaluators and front-line decision makers may have had access to different information. In the next section, we describe why evaluators may judge the behavior of decision makers to be timid even when decision makers are behaving appropriately. We then present two experiments that demonstrate a similar timidity error: in a situation in which evaluators and decision makers have identical information and in which decision makers tend to be more thoughtful and accurate (as measured by better calibration) than their evaluators, evaluators incorrectly conclude that decision makers are overly timid and risk averse. We conclude with some remarks on managerial risk taking and evaluation.

2. Previous research

There are reasons to suspect that evaluators may judge actions more favorably than front-line decision makers. In the remainder of the paper, we refer to a front-line decision maker as a “target”, because they are the target of a social evaluation by an evaluator. In many cases, the difference between evaluators and decision makers reflects a difference between breadth and depth. Evaluators may have a broader view of a decision than their targets (e.g., Buehler, Griffin, & Ross, 1994; Shapira & Berndt, 1997). They may be more likely to aggregate outcomes across multiple decisions and to focus on the base rate of comparable cases rather than the details of the case at hand (Kahneman & Lovalló, 1993). They may also be less subject to motivational biases that would lead them to be overconfident about a particular course of action (Buehler, Griffin, & MacDonald, 1997; Kunda, 1990).

However, the front-line decision maker who is the target of the evaluation typically has deeper knowledge of the decision at hand. Consider a situation where a decision maker faces a new or unusual decision without ready access to a set of comparable choices. Front-line decision makers who better understand the “structure” of the novel decision (e.g., the compound lottery that governs success and failure) might make better predictions than evaluators who rely on a broad understanding of the decision. Put simply, a front-line decision maker’s “deep” understanding may compensate in some situations for her “narrow” use of base rate information.

Thus, there may be some situations where front-line decision makers have a superior understanding of the situation because of their deeper view of the situation. One question is whether evaluators realize when they are in such situations and acknowledge the expertise of decision makers who have thought more carefully about a particular decision. The Lincoln and entrepreneur anecdotes above suggest that they may not always do so, and at least some experimental evidence also supports this supposition.

In an elegant experimental paradigm, Koehler (1994) and Koehler and Harvey (1997) asks one person, the *target*, to generate an answer to a particular question (e.g., Which picture will win the Academy Award for Best Picture?) and then state a probability that her answer will be correct. Another person, the *evaluator*, sees the target's answer and also provides a probability that her answer will be correct. On average, evaluators tend to believe that answers are more probable than do targets. For example, in one experiment, Koehler (1994) asked targets to predict Oscar winners for best film, actor, and actress, before the Oscar nominees had been announced. Interestingly, evaluators liked the target's answers better than the targets—on average, evaluators thought the targets had predicted 66% of the winners, while targets thought they had predicted only 47%. Both groups were wrong in the sense that both were overconfident, but targets were less wrong than evaluators. Targets showed better calibration (i.e., the probabilities they assigned better matched the objective probabilities that an answer was correct) and resolution (i.e., their probabilities better distinguished correct from incorrect answers).

The reason for these results, Koehler suggests, is that evaluators think less carefully about the task. Although targets typically considered a number of alternatives before selecting the one they prefer, evaluators may not—in the extreme, they may only consider the single alternative suggested by the target. When evaluators consider fewer alternatives, the option suggested by the target automatically seems more probable, a result consistent with support theory (Tversky & Koehler, 1994). In keeping with this interpretation, we refer to the original experiment as the “partial unpacked” condition (from the standpoint of the evaluator). Interestingly, the results reversed in a “fully unpacked” condition wherein targets and evaluators considered the same set of alternatives. Koehler (1994) repeated his study after the Academy released the five Oscar nominees in each category. When both groups considered this identical set of five alternatives, evaluators thought the target's answers were less probable (66%) than did targets (78%).

Psychologists have studied the attribution process for years, and one of the key results in this literature is that evaluators frequently attribute the behavior of targets to the targets' dispositions, whereas the targets typically attribute their own behavior to the demands of their

situation (Fiske & Taylor, 1991, pp. 22–38; Ross & Nisbett, 1991). However, this traditional literature does not provide enough detail to predict what evaluators will do in the kind of situation we study. Evaluators are likely to attribute the target's behaviors to disposition, but which disposition?

In our experiments, we allow evaluators to see not only targets' answers but also targets' willingness to act (i.e., bet) on their answers. What kinds of dispositional attributions will evaluators make about targets? We suggest that dispositional attributions are less straightforward in situations where the alternatives are fully unpacked than when they are partially unpacked. Consider evaluators in a situation where alternatives are fully unpacked: in Koehler's study, evaluators have considered all five alternative winners of an Academy Award, so they may find themselves making different attributions about targets depending on the quality of the targets' choices and how much targets are willing to bet on their answers. Evaluators may attribute targets' behavior to ignorance (if targets prefer a different winner than evaluators), or boldness (if targets bet on a dark horse candidate that is plausible but a long shot), or keen insight (if evaluators agree with the target's candidate and bet). Thus, in this condition it is not immediately clear which dispositional attribution evaluators will make. On the other hand, we suggest that the most straightforward dispositional attribution will occur in a situation where alternatives are partially unpacked, where the targets are more thoughtful because they have considered more alternatives than evaluators. In this case, the particular alternative chosen by the target is likely to look more plausible to evaluators because evaluators have not considered any alternatives. Thus the evaluator's major task is to explain why the target is not terribly willing to act on what the evaluator views as a very promising alternative. We suggest that evaluators in this situation may attribute targets' behavior to a disposition of timidity, assuming that the targets are risk-averse, conservative, or timid. Unfortunately, because evaluators have not thought as much about the situation as targets, this attributional process produces a timidity *error* in their evaluation: targets are not timid but simply realize that there are attractive alternatives other than the specific one they chose.

3. Study 1

In our studies, we extend Koehler's (1994) paradigm to test for a timidity error in evaluation. First, targets generated answers to a set of questions in a domain in which they had some expertise (most admired companies in a variety of industries) and indicated how willing they were to bet on their answers by means of cash equivalents (CEs). Second, evaluators saw the targets' answers and indicated how willing they were to bet on

these answers (again by means of CEs). We used CEs to make the judgments of both targets and evaluators incentive compatible—both targets and evaluators were paid based on their choices in the experiments.

In the key step of our experimental procedure, we asked the evaluators to rate the target as well as a “typical classmate” on a number of dimensions: knowledge, risk attitudes, confidence, etc. This procedure allowed us to assess whether evaluators thought that their particular target was more or less knowledgeable, risk averse, and confident than their typical classmate. This comparison also permits us to assess whether evaluations were accurate—because the targets are sampled at random from the evaluator’s classmates, evaluators would be incorrect to claim that their targets on average differ from their typical classmates in their willingness to take risks.

If the timidity error is enhanced, as we argued above, because evaluators think less hard than targets, then we should be able to control the size and direction of evaluation errors by manipulating the ease of the task faced by evaluators and targets. To do so, we included high- and low-knowledge conditions in our experiment. In the high-knowledge condition, it was relatively easy to come up with a plausible answer, whereas the opposite was true in the low-knowledge condition. Based on Koehler’s (1994) results, we expected that in the high-knowledge condition, where alternatives were obvious and readily available, targets and evaluators would unpack and consider roughly the same number of alternatives. On the other hand, in the low-knowledge condition, where participants had to work harder to generate alternatives, we expected that evaluators would not reason as carefully as targets. Thus, we predicted that the timidity error would be most noticeable in the low-knowledge condition where evaluators are less thoughtful than targets.

3.1. Participants

Participants in Study 1 were 124 MBA students at the University of Chicago: 62 served as targets and 62 served as evaluators. Participants randomly received a low- or high-knowledge questionnaire. Because of an error in photocopying questionnaires, we had unequal samples in the two conditions: 24 target-evaluator pairs in the high-knowledge condition and 38 pairs in the low-knowledge condition.

3.2. Method

Targets first predicted the most admired company in 10 industries (based on the 1997 *Fortune* survey of “America’s Most Admired Companies”). The *Fortune* survey appears in March of each year, and our study was conducted two months later. Targets wrote their guess for each industry in a blank space, and they were

told to answer each question even if they had to guess. For each of the 10 industries, participants then provided a cash equivalent for a gamble that paid \$10 if their answer for that industry was correct. Targets also provided cash equivalents for two risky prospects, one based on the flip of a coin and one based on the roll of a die. At this stage, we eliminated responses of targets who did not provide a prediction for each industry.

A separate group of participants played the role of evaluators. Each evaluator was paired with a specific target. First, evaluators received a sheet of paper on which the 10 industries were listed. The experimenter hand-copied the 10 answers originally provided by the target next to the name of each industry. Evaluators were told that the answers were generated by a student in another section of the same course. Each evaluator then provided a cash equivalent for each of her target’s 10 answers. She was then shown the target’s original form. After examining the target’s cash equivalents for each of the 10 questions, each evaluator rated her target and a “typical classmate” on seven dimensions: (i) knowledge of business; (ii) risk aversion; (iii) overconfident about his/her knowledge; (iv) willingness to take gambles; (v) interest in business; (vi) boldness; and (vii) self-doubt. Ratings were taken on a 5-point Likert scale (1 = *this person rates very low on this dimension*; 5 = *this person rates very high on this dimension*). Finally, evaluators provided cash equivalents for risky prospects based on the flip of a coin and the roll of a die.

The following week in class, 10 participants were randomly selected to play one of the lotteries generated by their answers, using a Becker, DeGroot, and Marschak (1964) procedure. This procedure is incentive compatible, so participants made less money if they overstated or understated their cash equivalents.

3.3. Materials

The companies for the high- and low-knowledge conditions were developed from the responses of a separate group of students in the same program ($N = 46$). They evaluated 23 industries on how much they knew about each industry (1 = *very little*; 7 = *a great deal*). After ordering the industries based on the average knowledge ratings, we designated the bottom 10 as “low-knowledge” industries and the top 10 as “high-knowledge” industries. The average knowledge ratings for the low- and high-knowledge industries were 2.55 and 3.33, respectively.

3.4. Results

Analysis of cash equivalents. We manipulated low- and high-knowledge industries to try to influence how readily participants could generate answers. The manipulation was successful: although accuracy was low

overall, targets correctly answered 18% of the questions in the high-knowledge condition and 12% in the low-knowledge condition ($t(60) = 2.24, p < .05$).

Consistent with our hypothesis, knowledge also affected whether evaluators or targets had a higher cash equivalent (CE). Fig. 1 plots, for each of the 20 industries, average industry knowledge ratings against average CE differences (evaluator's CE minus target's CE, or CEDIFF for short). For example, the evaluator's CE was \$1.18 higher on average than the target's CE for the railroad industry (average knowledge rating of 2.10), whereas evaluators had a \$2.04 lower CE on average than targets for the motor vehicles industry (average knowledge rating of 3.28). There is almost no overlap in CE differences across the low- and high-knowledge conditions (the dotted lines represent median splits for industry knowledge ratings and CEDIFF). This is confirmed by a Wilcoxon test ($p < .0001$). Note that although the low-knowledge condition was difficult as measured by the low accuracy rates, participants were quite willing to bet on their answers. For example, participants indicated that they were the least knowledgeable about the railroad industry, but they assigned a cash equivalent of \$2.58 to their answer in this industry, thus expressing a reasonably high willingness to bet.

The left panel of Fig. 2 plots the average CEs provided by evaluators and targets for the low- and high-knowledge condition, collapsing across the 10

corresponding industries. Evaluators stated higher CEs than targets in the low-knowledge condition ($t(36) = 2.65, p < .02$) and lower CEs in the high-knowledge condition ($t(22) = 2.01, p = .057$). Note that most of this difference was driven by the targets. While targets provided different CEs in the two knowledge conditions ($t(58) = 3.18, p < .01$), evaluators did not ($t(58) = .57, p > .20$). Note also that both evaluators and targets were, on average, overconfident. If participants were calibrated and risk neutral, the average CE for both evaluators and targets should be \$1.80 in the high-knowledge condition, and \$1.20 in the low-knowledge condition. (Participants were told they would receive \$10 if their answer was correct. Since their answers for high-knowledge industries were correct 18% of the time, the CE for a risk neutral decision maker is $\$10 \times .18 = \1.80 .) Of course, risk neutrality is a best case for calibration. As participants become more risk averse, their implied subjective probabilities would increase and hence their implied calibration would become worse.

To test the hypotheses outlined in the introduction, we need to show that our knowledge manipulation successfully made targets more thoughtful than evaluators. To do so, we assess the quality of the CEs, as measured by the internal consistency of probabilistic judgments (i.e., calibration) and the ability to discriminate or distinguish between different states of the world (i.e., discrimination or resolution) (cf. Yaniv, Yates, & Smith,

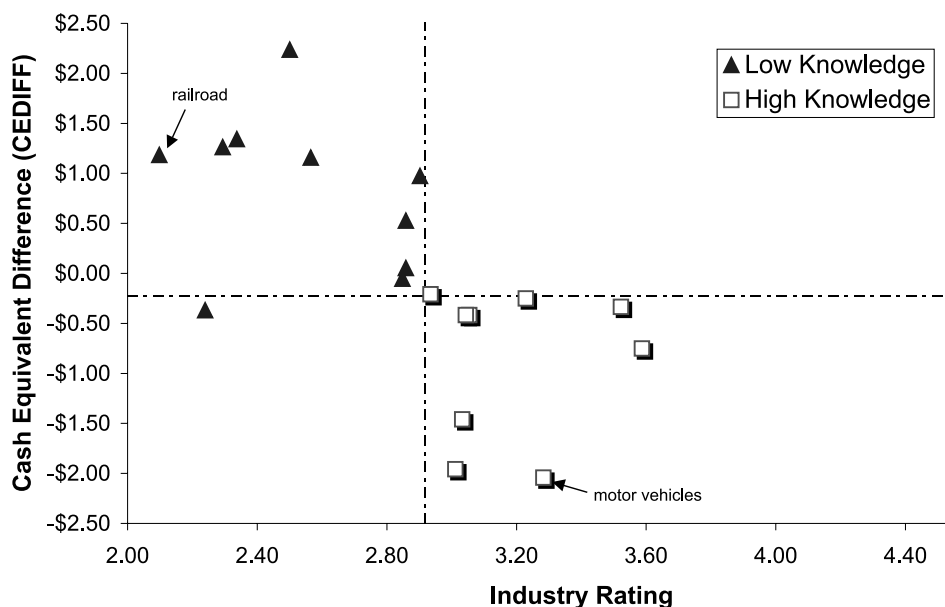


Fig. 1. Mean cash equivalent differences (evaluator minus target; CEDIFF) by industry (Study 1). Each data point represents a single industry (plotted from lowest knowledge industries to highest knowledge industries as measured by average knowledge ratings on 1–7 scale). The dotted lines show the median industry ratings (2.92) and the median average cash equivalent differences ($-\$0.23$) across the 20 industries. The 10 low-knowledge industries and the knowledge ratings given to these industries were: railroads (2.10), petroleum refining (2.24), rubber and plastics products (2.29), publishing, printing (2.34), wholesalers (2.50), chemicals (2.57), specialty retail (2.85), health care (2.86), engineering, construction (2.86), and electric and gas utilities (2.90). The 10 high-knowledge industries were: apparel (2.93), general merchandise-retailers (3.01), food (3.03), food & drug stores (3.04), hotels, casinos, and resorts (3.05), pharmaceuticals (3.23), motor vehicles and parts (3.28), commercial banking (3.52), telecommunications (3.59), and computers and office equipment (4.65).

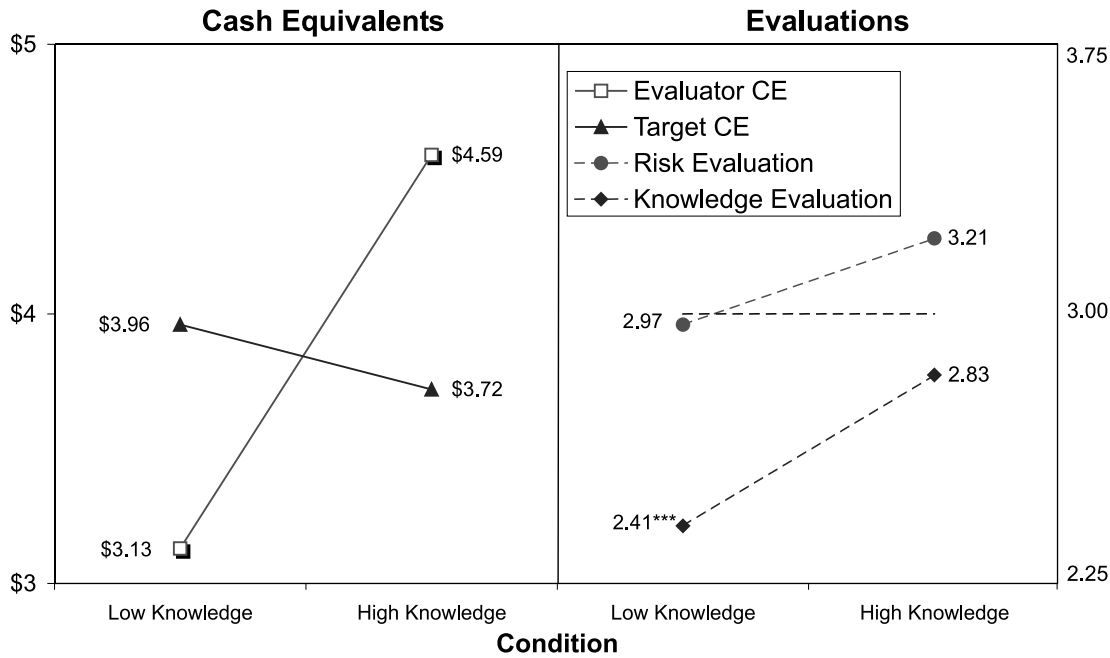


Fig. 2. The left panel shows mean cash equivalents by knowledge condition (Study 1). The right panel shows mean evaluation (risk and knowledge measures [RISK and KNOW]) of target by choice condition (Study 1). Means significantly different from 3, the middle of the scale, are denoted: (* $p < .05$); (** $p < .01$); (***) $p < .001$).

1991). We assume that our evaluators and targets are risk neutral, an assumption that describes most participants extraordinarily well; 81% of participants gave \$5 as a cash equivalent for the risky gamble that offered a 50% chance at \$10. Thus, we take their subjective probability for a particular prospect (e.g., “The most admired company in the pharmaceutical industry is Merck”) to be their CE for that prospect divided by 10.

To assess the quality of the CEs, we consider both calibration and resolution. We decompose the Brier score (Brier, 1950) into the calibration index (CI), where a score of 0 indicates perfect calibration and a higher score indicates poorer calibration, and into Murphy’s discrimination index (DI; Murphy, 1973), where higher scores indicate better resolution. We find that targets (mean CI = .14) were better calibrated than evaluators (mean CI = .19) in the low-knowledge condition ($t(36) = 2.16, p < .04$), as with Koehler’s findings, but the evaluators showed better calibration in the high-knowledge condition ($t(22) = 2.15, p = .04$). The resolution of targets and evaluators was indistinguishable in the two conditions.

Analysis of evaluations. In both conditions, evaluators rated their targets and a typical classmate on each of the seven dimensions listed above. To determine the structure of these seven dimensions, we submitted the target evaluations to a factor analysis. The first factor comprised the five items that deal with risk attitude (risk aversion, overconfidence, willingness to gamble, boldness, self-doubt; $\alpha = .83$ with risk-aversion and self-doubt reverse scored). The second factor consisted of

the items that deal with knowledge (knowledge of business, interest in business; $\alpha = .68$). We constructed a risk and knowledge measure by averaging across the appropriate items.

In our analyses, we consider the risk and knowledge measures for the target. We will label these variables as risk (RISK) and knowledge (KNOW), respectively. The right panel of Fig. 2 indicates that, in both conditions, the responses on the risk measure are consistent with our predictions on risk aversion. In the high-knowledge condition, where evaluators and targets were most likely to consider the same set of alternative hypotheses, evaluators concluded that targets were not risk averse, as measured by the difference from the middle of the scale ($t(23) = 0.96, n.s.$). In contrast, an evaluator in the low-knowledge condition may have been more likely to think primarily about the specific option proposed by the target. Thus, they explained the target’s lower willingness to bet by judging them to be risk averse ($t(37) = 4.70, p < .0001$). Moreover, evaluators judged targets in the low-knowledge condition to be more risk averse than targets in the high-knowledge condition ($t(60) = 1.95, p = .056$). For knowledge, evaluators’ judgments did not differ significantly from the middle of the scale for either condition (High: $t(23) = 1.51, p < .15$; Low: $t(37) = 0.21, n.s.$).

In particular, we hypothesized that an evaluator would question her target’s risk taking when confronted with a difference between her CEs and her target’s. If so, then CEDIFF should mediate the effect of experimental condition. The regressions in Table 1 illustrate this

mediation. Consistent with our previous analyses, Regression 1 shows that CEDIFF has a negative effective on RISK ($t(60) = 6.31, p < .0001$). Consistent with the hypothesized mediation, Regression 2 shows that although experimental condition has a negative effect on RISK (Regression 2; $t(60) = -1.95, p = .056$), the coefficient on condition drops to zero when CEDIFF is added to the regression (Regression 3; $t(59) = 0.00, n.s.$). Since, in addition, Regression 5 shows that experimental condition leads to a systematic difference in CEDIFF, the regressions suggest that the effect of the knowledge condition on evaluation is mediated by its effect on CEDIFF.

In these regressions, we used CEDIFF as an independent variable because we hypothesized that the evaluator evaluates the targets on the risk dimension primarily by exploring the difference between the target's willingness to act and their own. Of course, the timidity error could be driven more by the evaluator's willingness to act (Evaluator's CE) or the target's (Target's CE). Regression 4 measures that relative effect of the two CEs separately. Both coefficients have the correct sign (CEDIFF increases with an increase in the evaluator's CE or a decrease in the target's CE) and are significantly different after adjusting for the sign ($t(58) = 6.21, p < .0001$). Thus, the effect of CEDIFF on risk evaluation seems to be more sensitive to decreases in the target's CE than increases in the evaluator's CE, but both serve to produce an effect on risk evaluation.¹

¹ One reviewer was concerned that our method forced evaluators to explicitly deal with the difference between the target's willingness to act and their own willingness to act (and perhaps exacerbated this effect by using the very precise metric of the CE). To explore this concern, we ran an additional study that showed that the effect did not require evaluators to determine their own willingness to act until after they had evaluated the targets. In this study, we used bets instead of cash equivalents. For each of 8 questions, both targets and evaluators indicated whether they preferred receiving \$3 for sure, or betting on the target's answer; if the target's answer were correct, they would receive \$10 otherwise they would receive \$0. We also manipulated the order that the evaluator made judgments and choices. The control condition was much like Study 2. In our experimental condition, evaluators saw the target's answers and rated them for "reasonableness;" then they saw the target's willingness to bet on their answers and evaluated the targets on risk-taking, etc. Only after they evaluated the targets did evaluators indicate their own willingness to bet on the target's answer. The results indicated that it was not necessary for our experimental procedure to force evaluators to confront the difference between their own willingness to act and that of the targets. In this study, we defined BETDIFF as the difference between how many of the 8 questions the evaluator was willing to bet on, and how many the target was willing to bet on. A regression showed a significant relationship between BETDIFF and RISK ($\beta = -.341, t = 5.33, p < .0001$), but this relationship did not differ significantly between the control condition ($\beta = -.334$) and the experimental condition where evaluators indicated their own willingness to act only after they evaluated the targets ($\beta = -.349; t = .11, n.s.$). Thus, we argue that evaluators *implicitly* compare their willingness to act against the target's even when the experimental procedure does not force them to do so *explicitly*.

Finally, although CEDIFF mediates the effect of condition on judgments about risk, it has a much weaker effect on knowledge (KNOW): evaluations of knowledge decrease as the difference between the evaluator's CE and the target's CE increases. Regression 6 indicates that CEDIFF does have a significant impact on how targets are evaluated on the knowledge dimension ($t(60) = 2.03, p < .05$), but the effect is much weaker than the effect of CEDIFF on RISK.

3.5. Discussion

Study 1 demonstrates that evaluators may make timidity errors in evaluation. In the low-knowledge condition, evaluators judge their targets to be risk averse relative to their classmates. These evaluations are incorrect—targets were randomly assigned to the two conditions and were sampled from the evaluators' typical classmates.² The timidity error is particularly interesting because it occurred in the very condition in which targets showed better calibration than their evaluators.

While evaluators accused the targets of timidity, they did not judge them to be less knowledgeable than the typical classmate. We did not initially have a strong prediction about knowledge attributions, so we may have hindered our ability to detect effects on this dimension by only including two items (thus, the resulting scale has only marginal reliability). However, we note that if there is a theoretical, rather than an empirical, interpretation of the differences between knowledge and risk, that the differences may be produced because it harder for people to imagine that someone else has different knowledge (Camerer, Loewenstein, & Weber, 1989) than different motivations. In turn, people may be less likely to attribute any differences to knowledge, and more likely to attribute them to risk.

One interesting result of this study is that evaluators tended to attribute unfavorable qualities to the target. Such negative attributions are rare. Indeed, Klar and Giladi (1997) have shown that people rarely derogate particular individuals below the average of a group. Perhaps the negative evaluations are enhanced because the evaluators saw a portfolio of decisions by the target. If evaluators see a target make many decisions, they may overweight the situations in which they disagree with the target willingness to bet relative to those in which they agree. For this reason and to test the robustness of our

² As expected, the low- and high-knowledge condition participants did not differ in their underlying risk aversion as measured by non-significant differences in their cash equivalents for the two risky gambles ($t(45) = .22, n.s.$). In addition, targets did not rate their own risk taking to be significantly different from a "typical classmate" in either condition.

Table 1
Linear regression (Study 1)

Dependent variable	RISK				CEDIFF	KNOW			
	1	2	3	4	5	6	7	8	9
Regression ($N = 62$)									
Constant	2.61*** (.08)	2.83*** (.17)	2.61*** (.14)	2.05*** (.29)	-.87* (.41)	3.07*** (.09)	3.21*** (.15)	3.14*** (.15)	2.69*** (.28)
Experimental condition (1 = Low knowledge)		-.42^ (.22)	-.00 (.19)	-.07 (.18)	1.70*** (.52)		-.23 (.19)	-.10 (.21)	-.05 (.21)
CEDIFF	-.25*** (.04)		-.25*** (.04)			-.09* (.04)		-.08 (.05)	
Evaluator's CE				-.17*** (.05)					-.02 (.06)
Target's CE				.30*** (.05)					.12* (.05)
Adjusted R^2	.389	.044	.379	.379	.135	.049	.008	.019	.060

Standard errors are shown in parentheses. Coefficients significantly different from 0 are denoted.

* ($p < .05$).

** ($p < .01$).

*** ($p < .001$).

^ ($p < .10$).

effect, in the next study, we document the timidity error in a situation in which evaluators only see one decision by each target.

4. Study 2

Study 1 documents a timidity error, but it represents a somewhat novel evaluation situation because the evaluators had the opportunity to see how a target behaves on a repeated set of similar decisions. In the low-knowledge condition where we illustrate the timidity error, evaluators may have felt that they are getting overwhelming evidence that targets were timid because they saw a number of repeated examples where targets' behavior differed from their own. In Study 2, we wanted to give evaluators somewhat more impoverished information about targets to see whether they would be just as prone to accusing the targets of timidity. Instead of seeing a portfolio of 10 decisions by a single target, evaluators saw only *one decision* made by six different targets.

Study 2 also explores whether evaluations impact-related decisions. At the end of the experiment, evaluators were told to imagine playing this game again. They were then asked to select one target to answer a question and one target to set a cash equivalent.

4.1. Participants

Participants in Study 2 were University of Chicago MBA students. Forty three students served as targets, and 51 students served as evaluators.

4.2. Method

The method was almost identical to Study 1. However, evaluators saw 6 different targets, three of whom answered questions about low-knowledge industries and three of whom answered questions about high-knowledge industries.

4.3. Instructions to targets

The basic instructions were similar to Study 1. Targets were asked to predict the 1998 most admired company according to *Fortune* in six different industries; this study took place two months after this data appeared. For each of the 6 industries, they then provided a cash equivalent for a gamble that paid \$10 if their answer for that industry was correct. They also provided cash equivalents for a coin flip and die roll. Finally, targets also indicated how much they knew about each industry on a 1–5 scale (1 = *very little*; 5 = *a great deal*). The 6 industries ranged from 1.80 (rubber and plastics) to 3.06 (computers) for average knowledge ratings.

4.4. Instructions to evaluators

With a few exceptions, instructions to evaluators were similar to the instructions given to evaluators in Study 1. Evaluators first saw the answers generated by six different targets, each of whom generated an answer for one industry. The six targets were randomly selected for each evaluator. Evaluators then provided cash equivalents for each of the answers given by the targets. Evaluators were then presented with the cash equivalents provided by the targets. We changed the evaluation

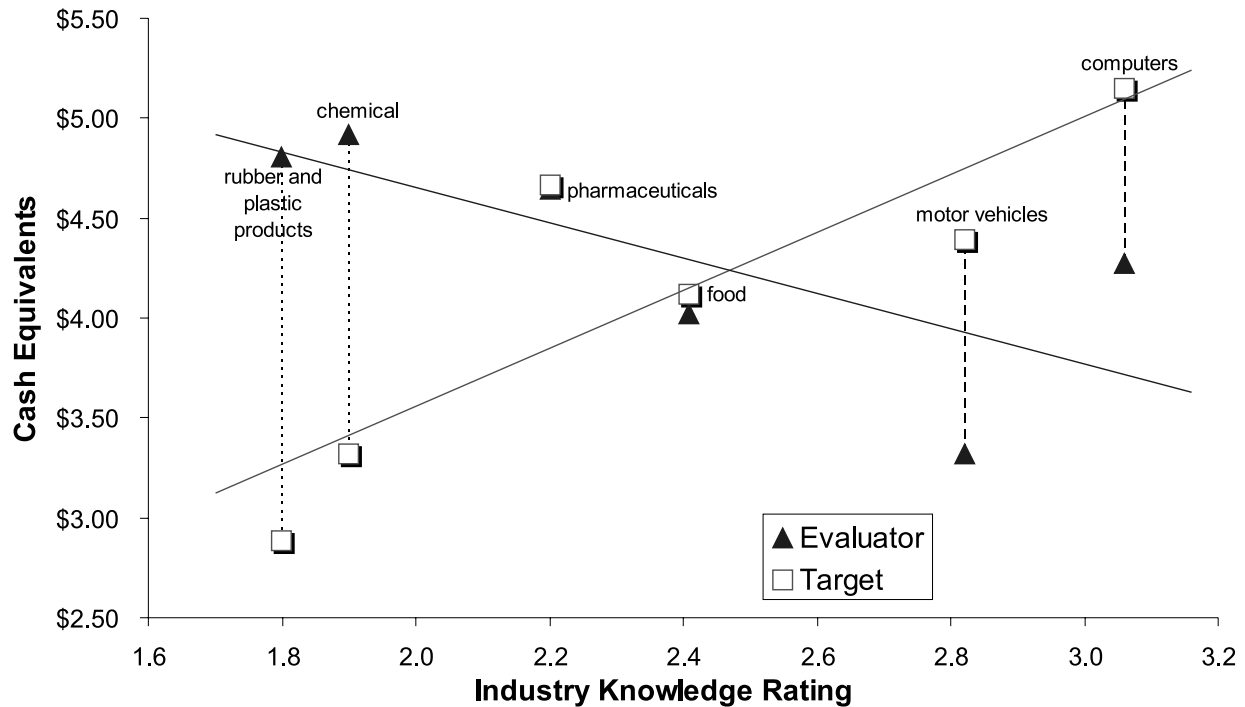


Fig. 3. Mean cash equivalents for evaluators and targets by industry (Study 2). Each data point represents a single industry (plotted from lowest knowledge industries to highest knowledge industries as measured by average knowledge ratings on 1–5 scale). The solid lines are regression lines for cash equivalents vs. knowledge ratings for evaluators and targets.

scale from Study 1: evaluators rated each of the six targets *directly* against a typical classmate using a scale anchored at -3 (“student rates *very low* on this dimension relative to a typical classmate”) to 3 (“student rates *very high* on this dimension relative to a typical classmate”). The five dimensions were: (i) knowledge of business; (ii) willingness to take gambles; (iii) interest in general business; (iv) boldness; (v) risk taking. (We chose these dimensions by choosing the individual items that loaded most heavily onto our aggregate risk and knowledge scales in Study 1.) Targets also provided a knowledge rating for each rating on the same 1–5 scale described above.

Finally, evaluators were given a selection task. They were told to imagine playing this game for another round, and then asked to select which of the six targets they would prefer to answer the question and which target they would prefer to set the CE that would determine their payoff.

4.5. Results

Analysis of cash equivalents. Fig. 3 shows that the knowledge manipulation is related to the ordering of CEs as in the previous study. For the two industries where evaluators were least knowledgeable (rubber and plastics, and chemical), evaluators had a higher average CE than the target. On the other hand, for the two industries for which the evaluators considered themselves

most knowledgeable (motor vehicles and computers), the target CE was higher than the evaluator CE. In the left panel of Fig. 4, we divide the industries into low- and high-knowledge by taking a median split along evaluator industry knowledge. The target’s CE exceeds the evaluator’s CE for the low-knowledge industries ($t(304) = 3.44, p < .001$), but the opposite is true for high-knowledge industries ($t(304) = 2.03, p < .05$).

In terms of accuracy, targets correctly answered 11% of the questions in the high-knowledge condition and 38% in the low-knowledge condition. This reversal between accuracy and industry knowledge reflects an idiosyncrasy of the *Fortune* rankings. In two of our three high-knowledge industries, the most admired American company was an American subsidiary of a foreign company: Toyota USA for motor vehicles and Nestle USA for food. Not surprisingly, few targets spontaneously generated these answers, 12% and 2%, respectively.³

³ These accuracy rates underestimate the amount of knowledge participants were able to bring to the task. We also computed an alternative measure of quality that considered how many “reasonable” answers the target provided. *Fortune* lists up to ten most admired companies in each industry. Thus, as an alternative measure of quality we computed the “ballpark rate”—the percentage of questions in which participants correctly identified one of the top ten companies. In five of the six industries, the ballpark rate was above 75% (the exception being the food industry at 19%), and overall, across all six industries, the ballpark rate was 70%. Thus, although the task was difficult, participants generally provided reasonable answers.

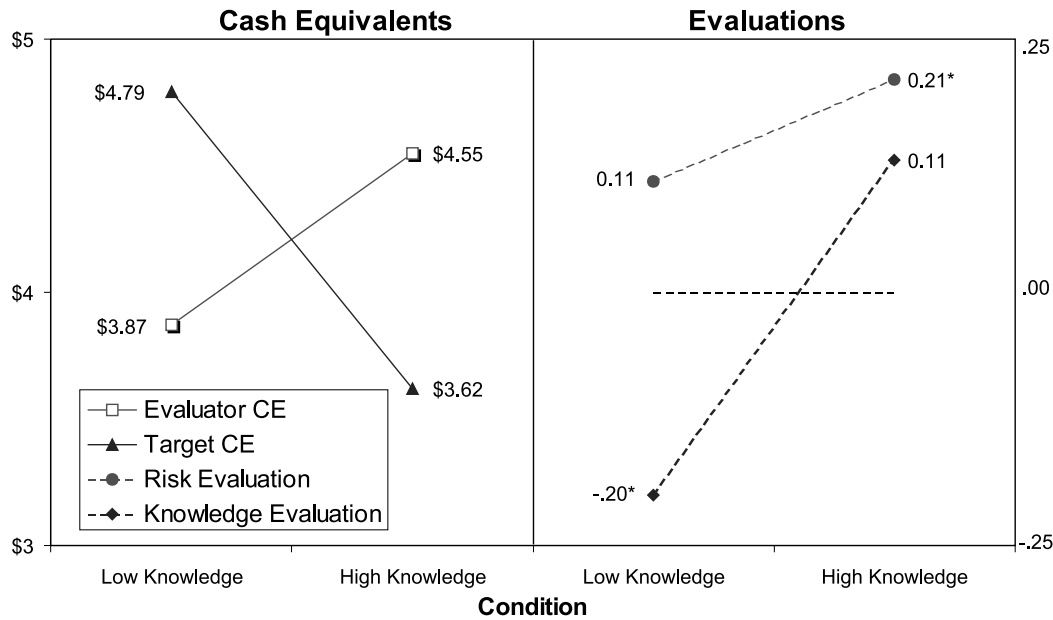


Fig. 4. The left panel shows mean cash equivalents by knowledge condition (Study 2). The right panel shows mean evaluation (risk and knowledge measures [RISK and KNOW]) of target by choice condition (Study 2). Means significantly different from 0 are denoted: (* $p < .05$); (** $p < .01$); (***) $p < .001$).

Finally, even though accuracy rates were lower for the high-knowledge industries than the low industries, as in Study 1, targets were still better calibrated than evaluators as measured by the calibration index overall ($t(99) = 2.62, p = .01$) and for both the low- and high-industries (Low: $t(99) = 1.98, p = .05$; High: $t(99) = 2.25, p = .03$). In terms of resolution as measured by the discrimination index, targets and evaluators discriminated states of nature equally well overall ($t(99) = 0.21, ns$). However, consistent with our hypothesis that targets are more thoughtful in the low-knowledge condition, targets showed better discrimination in the low-knowledge condition, but the opposite was true in the high-knowledge condition (Low: $t(99) = 2.30, p = .02$; High: $t(99) = -2.63, p = .01$).

Analysis of evaluations. We next consider the evaluations. We term the risk and knowledge evaluation measures RISK and KNOW, respectively. Consistent with timidity error, the right panel of Fig. 4 shows that targets were seen as significantly more risk averse than the typical classmate when they made decisions regarding low-knowledge industries ($t(152) = 1.98, p < .05$). In the high-knowledge condition, targets were judged to be slightly but not significantly more risk seeking than the typical classmate ($t(152) = 1.08, ns$).

Fig. 4 also shows how evaluators rated targets in terms of knowledge. Evaluators judged targets to be more knowledgeable than the typical classmate. The difference was significant in the high-knowledge condition ($t(152) = 2.19, p < .05$), but not in the low-knowledge condition ($t(152) = 1.29, n.s.$).

The linear regressions in Table 2 illustrate the relationship more clearly and suggest the same type of mediated relationship as we found in Study 1. We ran fixed-effects regressions, controlling for evaluator differences, and found that the risk evaluation was higher in the high knowledge condition than the low-knowledge condition (Regression 10), but that the effect of knowledge on risk evaluation disappeared when CE-DIFF was added to the regression (Regression 11). Furthermore, the knowledge condition led to a systematic difference in CEDIFF (see Regression 15) as in Study 1, thus suggesting that the effect of experimental condition on risk evaluation seems to be mediated by the effect of condition on CEDIFF. Finally, note that Regressions 13 and 14 show that the knowledge evaluation was not affected by either the knowledge condition or CEDIFF.

Recall that in Study 1 we found that the target's CE had about twice the effect on risk evaluation as evaluator's CE. In Study 2, there is an even more pronounced asymmetry (Regression 12). Both coefficients have the correct sign, but the coefficients for the target (.35) and evaluator (-0.06) are significantly different when adjusted for sign ($t(252) = 8.82, p < .0001$). This difference suggests that the evaluator is much more sensitive to a change in the target's CE than to a change in her own CE as they were in Study 1 (Regression 4). However, in Study 1, evaluators rated a single target based on ten answers and ten cash equivalents. In Study 2, evaluators rated six different targets. Thus, in Study 2 the contrast across targets is particularly salient to the evaluator, and

Table 2
Fixed-effects (controlling for evaluator effects) linear regressions for Study 2 (standard errors are in parentheses)

Dependent variable Regression ($N = 305$)	RISK			KNOW		CEDIFF
	10	11	12	13	14	15
Experimental condition (1 = Low knowledge)	-.33* (.16)	.01 (.14)	.16 (.12)	-.10 (.12)	-.04 (.13)	1.79*** (.42)
CEDIFF		-.19*** (.02)			-.03 (.02)	
Evaluator's CE			-.06** (.02)			
Target's CE			.35*** (.03)			
Adjusted R^2	.161	.376	.521	.049	.061	.099

Coefficients significantly different from 0 are denoted.

* ($p < .05$).

** ($p < .01$).

*** ($p < .001$).

Table 3
Fixed-effects (controlling for evaluator effects) logistic regressions for Study 2, selection task

Dependent variable: Regression ($N = 300$)	Select for risk 16	Select for knowledge 17
Experimental condition (1 = High knowledge)	.36	.06
RISK	.40**	.10
KNOW	.52**	.82***

Shown are β -coefficients. Statistically significant coefficients (Wald Statistic) denoted:

* ($p < .05$).

** ($p < .01$).

*** ($p < .001$).

thus it is not surprising that the asymmetry is more pronounced here.

Analysis of selection. Finally, we consider the selection task. Recall that evaluators chose one of the six targets to answer questions (“select for knowledge”), and one to set CEs (“select for risk”). Table 3 shows the results of fixed-effects logistic regressions to predict which target was selected to provide an answer and which to set a level of risk. The two selection decisions were highly related to the evaluations. The probability that a target is “selected for risk” is highly related to RISK and KNOW (setting the correct cash equivalent requires having the appropriate risk attitude as well as having industry knowledge), and the probability that the target is “selected for knowledge” is highly related only to KNOW.

4.6. Discussion

Study 2 shows that the timidity error we documented in Study 1 is robust to the quantity of information that evaluators have about targets. Admittedly, while these effects are significant, they are small (about .2 scale points on a 5-point scale). However, in context, we think even these relatively small effects are interesting—evaluators were willing to critique the target's level of risk-taking, even when they only saw one decision by each

target. Furthermore, evaluators' judgments of knowledge and risk strongly influenced which targets they chose for related tasks.

As in Study 1, evaluators judged targets to be more timid in the low-knowledge condition, despite the fact that targets in this condition were more thoughtful than evaluators as measured by both calibration and resolution. It is worthwhile to note that the opposite pattern in the high-knowledge condition—that evaluators showed better discrimination—is not inconsistent with our explanation. Consistent with our explanation and with Koehler's earlier findings, targets and evaluators in the high-knowledge condition consider roughly the same alternatives. In this case, there is no clear advantage for the targets in terms of how carefully they consider the alternatives, and the results were mixed—targets were better calibrated, while evaluators showed better discrimination. Most importantly, the timidity error again occurred in the condition where we expected it based on our theoretical argument.

5. General discussion

The two studies in this paper document a timidity error in evaluation. Evaluators who think less carefully about alternatives are prone to judge their targets to be

too timid. The timidity error occurred whether evaluators saw a large (Study 1) or small sample (Study 2) of the targets' behavior. The selection task in Study 2 also shows that evaluators' judgments of their targets' risk taking strongly influenced whether evaluators chose that target for another round of this decision task. Thus, these evaluations are likely to be consistent with how evaluators treat their targets on related tasks in the future.

There are two reasons to consider the timidity error to be an error. First, evaluators judged targets to be more timid on average than their typical classmates, even though they were drawn at random from the pool of classmates. Second, in the cases where we find a timidity error, targets seem to be more thoughtful about their decisions than evaluators; in the low-knowledge conditions, targets show better calibration (Studies 1 and 2) and resolution (Study 2) than their evaluators.

The evaluation error we document is particularly interesting because it occurs even when evaluators and targets share identical information. In the low-knowledge condition, all that distinguishes targets from evaluators is that the targets have thought more carefully about their decision, as measured by better calibration. In our studies, evaluators mistakenly judge others to be conservative even when the others have identical information; thus, we should perhaps be even more suspicious when evaluators judge others to be too timid in the many situations where targets have superior information. When we second-guess the President or the CEO or the quarterback on Monday morning, we typically have less training and knowledge about politics or business or football than our targets. In this situation we should probably be especially wary of an evaluator who accuses a target of timidity.

The same dynamics we have studied in this paper may sometime cause us to evaluate ourselves as overly timid. Research on regret (Gilovich & Medvec, 1995) has shown that people frequently regret not taking an action at an earlier point in their own life, in part because they have an impoverished picture of their prior self (e.g., compared with current university students, graduates underestimated how much their grade point average might have suffered from adding an extra elective). As in our studies, evaluators who evaluate a target (in this case the prior self) are likely to think about the world in an insufficiently complex way, and may find it incomprehensible that their target was too timid to pursue the obvious course of action.

Limitations of this study and areas for future research. Thus, in our studies and in prior research there is evidence of a timidity error on the part of evaluators. We feel that documenting this kind of error by evaluators is important because when targets and evaluators disagree about the correct action, our initial tendency based on our theoretical models might be to assume that the

targets are actually at fault—the standard rational model assumes that employees are risk averse and the standard psychological model assumes that employees are loss averse. We regard the data in the current paper as providing an important “existence proof” that evaluators may make timidity errors even when targets are behaving in a reasonable fashion (and indeed when targets are behaving, in some ways, more rationally than evaluators). At the same time, we have not done a thorough Brunswickian-style sampling of the environment see whether the kinds of situations we have studied in this paper are typical or atypical of the evaluation contexts in organizations. There may indeed be situations where people engage in the opposite error and accuse others of being too bold, and it would be useful to know if and when this occurs. We think that evaluation errors are interesting and have important implications for organizational life, but the current study is limited because it has only studied one particular context, and much more could and should be done to understand evaluation errors.

The psychological process underlying the timidity error also warrants further research. In explaining the error, we have used Koehler's (1994) explanation of the evaluator-generator discrepancy, and have highlighted the psychological role of differential unpacking by evaluators and targets. However, there may be other psychological processes that also exacerbate or moderate the timidity error, such as the accessibility of alternative attributions.

It would also be interesting to know how evaluation errors might be overcome. We provide one final illustration of the timidity error that suggests a possible means of repairing it. During the Manhattan Project in World War II, Du Pont was asked to manufacture enough plutonium for the bomb. In the Manhattan Project, the physicists were in charge, and this set up a situation very much like the ones in our paper where a smart, but potentially less thoughtful evaluator is in a position to evaluate a more thoughtful front-line decision-maker. Although the physics of manufacturing plutonium from uranium was well-understood at a theoretical level, there were substantial engineering problems that had to be overcome to manufacture plutonium in practice. Not surprisingly, the physicists accused the engineers of timidity: “Some physicists charged that Du Pont was too preoccupied with safety and margins of error rather than with speed in doing the job. A workable arrangement was not achieved until [the lead Du Pont engineer] succeeded in getting the physicists to review with Du Pont's designers all blueprints for the Hanford works and sign off on them. Once this process began, some of the [physicists] began to appreciate the magnitude of the engineering work involved...” (Hounshell & Smith, 1988, p. 341). Thanks to the clever reaction of the lead Du Pont engineer at

least some of the evaluators were encouraged to think more carefully about the complexities of the situation by completing the detailed review process.

As in our studies, this historical incident suggests that the engineers' conservative approach may have been justified. According to the historians, "Within hours of starting up the first full-scale reactor, the wisdom of Du Pont's safety and margin-of-error cautiousness was dramatically demonstrated." The calculations of the theoretical physicists had not taken into account the difficulties of real-world materials, and their initial design did not contain sufficient material to set off a self-sustaining reaction. Luckily, the Du Pont engineers had not acquiesced when the physicists accused them of being overly cautious and had designed the reactor with extra space to insert additional uranium slugs. If the engineers had abandoned their conservative margin-of-error approach, the reactor "would have proven to be an expensive, sophisticated heap of junk" (Hounshell & Smith, 1988, p. 341). The example illustrates that sometimes inside information may lead front-line decision makers to take an informed but cautious approach. Unfortunately, our research suggests that in such situations, evaluators may be all too willing to accuse their targets of being overly timid.

Acknowledgments

We thank Josh Klayman, Derek Koehler, two anonymous referees, and participants at various conferences for useful comments on this paper.

References

- Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9, 226–232.
- Bhide, A. (1992). Bootstrap finance: The art of start-ups. *Harvard Business Review*, 70, 109–117.
- Bhide, A. (1994). How entrepreneurs craft strategies that work. *Harvard Business Review*, 72, 150–161.
- Bhide, A. (1997). Personal communication, May 1997.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.
- Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, 67, 366–381.
- Buehler, R., Griffin, D., & MacDonald, H. (1997). The role of motivated reasoning in optimistic time predictions. *Personality and Social Psychology Bulletin*, 23, 238–247.
- Camerer, C. F., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, 97, 1232–1254.
- Donald, D. H. (1995). *Lincoln*. New York: Simon & Schuster.
- Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (second ed.). New York: McGraw-Hill.
- Gilovich, T., & Medvec, V. H. (1995). The experience of regret: What, when, and why. *Psychological Review*, 102, 379–395.
- Hamel, G. (2000). Reinvent your company. *Fortune*, 141(June 12), 99–118.
- Heath, C., Larrick, R. P., & Klayman, J. (1998). Cognitive repairs: How organizational practices can compensate for individual shortcomings. *Research in Organization Behavior*, 20, 1–37.
- Hounshell, D. A., & Smith, J. K. (1988). *Science and corporate strategy: Du Pont R & D 1902-1980*. Cambridge, UK: Cambridge University Press.
- Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk-taking. *Management Science*, 39, 17–31.
- Klar, Y., & Giladi, E. (1997). No one in my group can be below the group's average: A robust positivity bias in favor of anonymous peers. *Journal of Personality and Social Psychology*, 73, 885–901.
- Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 461–469.
- Koehler, D. J., & Harvey, N. (1997). Confidence judgments by actors and observers. *Journal of Behavioral Decision Making*, 10, 221–242.
- Kotter, J. P. (1996). *Leading Change*. Boston: Harvard Business School Press.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480–498.
- Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12, 595–600.
- Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.
- Shapira, Z. (1994). *Risk Taking: A Managerial Perspective*. New York: Russell Sage Foundation.
- Shapira, Z., & Berndt, D. J. (1997). Managing grand-scale construction projects—A risk-taking perspective. *Research in Organizational Behavior*, 19, 303–360.
- Tushman, M., & O'Reilly, C. A. (1997). *Winning through innovation: A practical guide to leading organizational change and renewal*. Boston, MA: Harvard Business School Press.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A non-extensional representation of subjective probability. *Psychological Review*, 101, 547–567.
- Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of Discrimination Skill in Probabilistic Judgment. *Psychological Bulletin*, 110, 611–617.

Received 28 September 1999