Working Paper

Judgments of Performance: The Relative, the Absolute, and the In-between

Katherine A. Burson
Stephen M. Ross School of Business
at the University of Michigan

Joshua Klayman
University of Chicago Graduate School of Business

Ross School of Business Working Paper Series
Working Paper No. 1015
August 2005

Judgments of performance: The relative, the absolute, and the in-between

Katherine A. Burson

University of Michigan Ross School of Business

Joshua Klayman

University of Chicago Graduate School of Business

Abstract

People often evaluate how their abilities or their achievements compare to those of others. Such judgments tend to show *asymmetric weighting*: They are more influenced by impressions of one's own performance than by impressions of the comparison group. We challenge interpretations of this effect as an egocentric focus. We show that asymmetry is much smaller when predicting concrete performance measures rather than general skill level and when the judge has experienced the task in question. We attribute this to a tendency to understand poorly-specified performance scales as implicitly relative. Moreover, judges' modest tendency toward asymmetrical weighting may be adaptive, because judges often know more about their own performance than about their peers'. This does not mean, though, that judges are sensitive to optimality: We find that they are insensitive to the effects that objective feedback has on the optimal weighting of estimates of one's own and others' performance.

Judgments of performance: The relative, the absolute, and the in-between

In many domains, people need to evaluate how their abilities or their achievements compare to those of other people. A key component in choosing which jobs to apply for, which slopes to ski on, or which camera to purchase is where one stands relative to other candidates, other skiers, or other amateur photographers. Indeed, research suggests that consumers often do not know their tastes in any absolute sense, but instead choose products based on their beliefs about how their tastes compare to others' (Prelec, Wernerfelt, & Zettelmeyer, 1997; Wernerfelt, 1995). For example, a wine drinker may think of herself as of average sophistication, and choose seemingly average wines on that basis.

Evidence suggests, however, that people do not accurately gauge their own relative standing in many contexts (see Alba and Hutchinson, 2000 for a comprehensive review). One systematic element of misjudgment is that perceptions of relative standing vary with the perceived difficulty of the task, even when the task is difficult or easy for people in general. It seems that anything that makes a task seem harder causes estimates of relative performance to decline. This includes manipulating the difficulty of the target domain (Kruger, 1999) and tightening the required precision of answers (Burson, Larrick & Klayman, in press). An anecdotal example comes from the Chinese national college examinations in 1999. Chinese students applied to colleges and universities between the time that they took the examination and the time they learned their results. That year, the examination was more difficult than usual, and most of the 50,000 students taking it underestimated their relative performance, leaving premier universities short of

applicants. As the *China Daily* put it "A student who is dying for entering China's prestigious Qinghua University applied for an ordinary college due to his wrong assessment of his performance" (China Daily, 1999). Indeed, with difficult tasks one often sees pessimistic rather than optimistic biases: On average, people think they are below average (Burson et al., in press; Kruger, 1999).

This *difficulty effect* reflects the tendency for judges to overweight or anchor on their own perceived degree of success, and to underweight their understanding of how well others are likely to perform (Giladi and Klar, 2002; Klar and Giladi, 1999; Klar, Medding, & Sarel, 1996; Kruger, 1999; Kruger & Burrus, 2004; Moore and Kim, 2003; Windschitl, Kruger, & Simms, 2003). Study participants show a high correlation between judgments of their own absolute performance and judgments about their relative performance (typically, judgments of the percentile into which their performance will fall within a specified peer group). The correlation is much weaker between estimates of peer performance and the participant's relative standing. See Figure 1.

Most investigators have attributed this *asymmetric weighting* to one or both of two psychological processes: egocentrism and focalism (Klar & Giladi, 1997; Kruger & Burrus, 2004; Windschitl et al., 2003). Egocentrism is implicated in that the self figures more prominently in judgment than do others. Focalism is implicated in that judges who are asked to evaluate themselves focus on the information most obviously relevant, namely information about themselves. Both egocentrism and focalism lead judges to put more weight on what they know about themselves than what they know about members of the comparison group That asymmetric weighting is the focus of this paper, and thus,

we will not need to distinguish between egocentrism and focalism; we will refer to them collectively as *egocentric focus*.

In this paper, we challenge the hypothesis of egocentric focus on both descriptive and normative grounds. We hypothesize (as do Moore and Kim, 2003) that the observed asymmetric weighting is due, at least in part, to ambiguity in whether a particular judgment is meant to be absolute or relative. In particular, when judges lack a well-learned absolute scale in a skill domain, we hypothesize that they naturally fall back on relative judgments. For example, how skilled are you at cleaning your kitchen? Lacking an appropriate absolute scale of cleaning performance, it may be natural to make a judgment based largely on how good you think you are compared to other people. We believe that some experimental judgments that have been taken to be judgments of absolute performance may have this quality to some degree. If so, then it is natural that answers to these questions correlate highly with more explicit relative judgments, such as percentiles. In contrast, a question like, "How often do you clean your kitchen?" is more unambiguously absolute, and we hypothesize that more objective performance questions like this will show lower correlations with relative judgments of kitchen-cleaning frequency. Support for this hypothesis comes from the stereotyping literature. Biernat (2003) argues that subjective language like that used in "skill" scales elicits comparison to a category. When asked to judge the intelligence of a 4-year-old, a parent will claim she is brilliant because the comparison is implicitly to other 4-year-olds. In contrast, the meaning of "common-rule" judgments (e.g., her score on a specific task) necessarily remain constant across contexts. Similarly, the tendency to use relative judgments to set the scale for absolute judgments may be greater when the question is hypothetical than

when the judge has direct experience with the performance in question. The appropriate scale of measurement may be more vague when asking, “How well do you tell jokes” than “How well did you tell those three jokes?” In Studies 1 and 2, we show that general skill questions show much more asymmetry of weights than do questions about specific behaviors, and that hypothetical estimates show more asymmetry than do estimates based on performance experience.

We also question whether, as most investigators imply, it is optimal to put equal weight on estimates of one’s own performance and estimates of the average performance of the comparison group. Because relative judgments are necessarily based on judgments about the difference between one’s own performance and the average, logic seems to dictate weighting those equally in estimating relative performance. However, we hypothesize that there is a parallel here to the phenomenon once known as *false consensus*. People’s judgments about others' beliefs and attitudes correlate with their own beliefs and attitudes. Thus, people are systematically biased toward thinking that others feel as they do—a seemingly egocentric bias. However, later research showed that such a bias may not be an error from the point of view of the judge (Davis, Hoch, & Ragsdale, 1986; Dawes, 1990; Dawes & Mulford, 1996; Hoch, 1987; Hoch, 1988). People are not very accurate in judging how their attitudes differ from those of other people. Although they may also misjudge their own attitudes, those are nevertheless a valid cue to what other people think, because there is a fair measure of actual consensus among one’s peers. Thus, one’s own opinions may be the best cue to others’. Indeed, Davis, Hoch, and Ragsdale (1986) and Hoch (1987) showed that judges could be more accurate if they placed even *more* weight on their own opinion.

The parallel in relative judgments is that judges may have better insight into their own ability at a given task than they have in the average ability of their peers, and thus weighing one's own performance more heavily may improve accuracy in estimating relative standing. Absolute performance is certainly a predictor of relative performance, albeit an imperfect one. The worse one does on a test, the more likely one is to have performed below average. Put another way, if a task is difficult for you, that may indicate that it is difficult for everyone, or that it is more difficult for you than for most people. If it is hard to judge the former, than it is reasonable to assume that the latter is likely to be true, at least probabilistically.

However, even if we can demonstrate that asymmetrical weighting is useful in judgments of relative standing, that does not prove that people are aware of the fact, or that they weight self and other appropriately. In Study 3, we show that asymmetric weighting may indeed be appropriate in many situations, but that judgments of relative ability are not sensitive to changes in the validity of information. That is, people apply very similar weights to the available information regardless of whether the source is their own intuitive estimate or objective information. Thus, it cannot be said that people follow an optimal weighting strategy.

Study 1

In this study, we manipulated two variables that we hypothesized would affect the extent to which participants treated a judgment of ability as absolute or relative. One variable was whether judgments concerned a general rating of *skill* in a domain or an estimate of a particular performance measure, *score*. Our hypothesis is that general skill

questions are more prone to be treated implicitly as relative judgments, whereas concrete score questions more clearly elicit absolute performance estimates. This should manifest itself in a lower correlation between estimates of one's own performance and estimates of relative performance for score questions, leading to more symmetrical weighting of self and other in estimates of relative performance.

The second manipulated variable was whether the judgments were made on the basis of only a description of the task or following experience performing the task. We hypothesize that estimates made without direct experience would tend to be treated as relative judgments, because the appropriate scale to use for absolute judgments would be unclear. Following experience at performing a task, we expect participants to have a better idea of an appropriate absolute scale of performance, and thus to give more clearly absolute estimates.

Methods

Participants. 40 University of Chicago students were recruited with advertisements posted around campus and were paid $9 for their participation, which required approximately 45 minutes.

Materials. We used three game-like tasks. One was a "Word Prospector" game like one previously used by Burson et al. (in press). In this game, the player attempts to construct as many four, five, and six letter words as possible from the letters contained in the word *gorgonzola*. Participants receive points for each letter of each correct word they spell, and lose points for nonexistent or misspelled words. For example, if a participant spelled the word *along*, five points would be counted toward the overall score. But, if the

participant spelled the nonexistent word *gool*, four points would be subtracted from the overall score.

Another task was a Remote Associates Test, used by Kruger (1999). In this task, participants are asked to find the word that is associated with three other words. For instance, given the set of words “athletes—web—rabbit”, the correct answer would be *foot*. There were ten sets of words.

For the third task, we selected a non-verbal, memorization game commonly called “Concentration.” This game requires participants to turn over pairs from a 6 by 6 matrix of cards containing a variety of pictures and symbols (in this case, represented on a computer monitor). Players must remember the location of the cards in order to consecutively turn over all matching pairs of cards as quickly as they can. Participants played four rounds of this game.

Design. Task type was a within-participants variable. The order of tasks—Word Prospector, Remote Associates, Concentration—was not varied. We manipulated type of absolute dependent measure (skill or score) between participants. Participants were randomly assigned to a condition in which they were asked about their *score* or about their *skill* (e.g., for the Remote Associates Test, “On how many of the 10 sets do you think you would have correctly guessed the fourth word?” versus “How would you rate your ability to correctly guess the fourth word?”). We also manipulated between participants whether the questions were *hypothetical* or *experiential*. In the hypothetical condition, participants were given the game instructions, but did not play the games. In the experiential condition, estimates were made after participants played each game, but without receiving any additional feedback from the experimenter.

Procedure. Participants read several pages of instructions from the computer screen including an explanation of the first task (Word Prospector), an example, and the scoring rules for the task. In the *hypothetical* condition, participants next made their performance estimates based on how they thought they would do at the game. Those in the *experiential* condition played the game before making estimates. For them, a screen containing the 10-letter word "gorgonzola" and indicating the time limit (3 minutes) followed the instructions. They typed their list of words into the computer until the time limit was reached, at which point the computer ended the game.

Following the task description and, in the experiential condition, the game itself, participants were asked to answer three questions about their performance on the game. The order of the three questions was counterbalanced. They estimated their own performance or ability, the average performance or ability of "other University of Chicago students participating in this experiment," and the percentile rank into which their performance or ability would fall in relation to other participants. The use of the percentiles was described as follows (in this example, for Word Prospector):

> Compared to other University of Chicago students participating in this experiment, how successful were you at finding 4, 5, or 6 letter words in the word gorgonzola? Please estimate what percentile you are in compared to other participants in this experiment. Writing 10% means that you did better than only 10% of your peers, writing 90% means that you did better than 90% of your peers, and writing 50% means you did better than half of the students participating in this study. Write any percentile between 0 and 99 in the space below.

The nature of the two absolute-estimate questions varied with condition. In the *score* condition, participants estimated the number of points that they expected to receive and the number of points that they expected the average person to receive. In the *skill* condition, participants estimated their ability and the average person's ability using the

method used previously by Kruger (1999)—a 10-point scale ranging from *very unskilled* to *very skilled*. Both groups estimated their relative ability using percentiles.

Participants then repeated the procedure for the next task, Remote Associates, which had a six–minute time limit. They made performance estimates for that task, and then continued on to the Concentration game and its estimates. All participants also provided demographic information. Participants in the hypothetical condition stayed on to complete an additional experiment, which we will describe later as Study 2.

Results

For each of the three tasks in each of the experimental conditions, we calculated the path weights between estimates of one's own absolute and own relative performance and between estimates of peer-average performance and own relative performance.[1] We also determined the correlation between estimates of own and peer-average performance. Results are shown in Figure 2.

For estimates of skill using a 10-point rating scale, the resulting paths clearly show greater weight on one's own performance than on average performance when estimating relative performance. This replicates the findings reported by Kruger (1999) using similar questions, as well as those reported by Klar and Giladi (1999) and Kruger and Burrus (2004). However, we see a very different picture for estimates of scores. Here, participants estimates of the percentile of their performance show nearly equal weighting of own and average performance. These results using estimates of performance rather than general skill ratings are consistent with findings by Moore and Kim (2003, Study 3), who found less asymmetry of weights in a task requiring participants to estimate their number of correct answers on a trivia quiz. The variable of hypothetical

versus experiential basis for estimates shows little effect. However, it may be that our hypothetical condition was not really hypothetical enough. Participants received detailed task descriptions, including examples, which may have allowed them to simulate task experience. We examine this possibility later, in Study 2.

Discussion

The results of Study 1 support the hypothesis that Moore and Kim (2003) and we propose: Seeming overweighting of one's own performance in judgments of relative skill may be due in large part to the fact that the rating one gives one's own skill is already relative in nature. That is, the interpretation of a scale from *very unskilled* to *very skilled* is heavily influenced by a sense of the distribution in the reference group. What does it mean to have a skill of 5 in computer programming? In operating a computer mouse? Participants do not say that their peer-group average is exactly 5 in all cases—they are willing to recognize that their peers are more skilled at some tasks than at others. Yet even that may reflect comparison to some larger population that is used to calibrate the scale (see Giladi & Klar, 2002 for more on this argument). The upshot is, we propose, that judgments of one's own skill already reflect a large measure of relative judgment, and thus it is no surprise that they correlate very highly with percentile estimates, and that variation in the estimated average score adds little.

In contrast, score questions have a scale that is more clearly absolute—how many points scored, how many questions answered correctly. Here, we find little asymmetry. Participants' judgments of the percentile of their performance are predicted almost equally by their estimates of their own scores and their estimates of average peer scores.

Study 2

Study 1 showed little impact of hypothetical versus experiential basis for estimates. However, in that study, even the hypothetical condition provided more concrete information about the task than has been provided in some other studies. For example, Kruger (1999) provided participants with task experience in some cases, but also asked participants to make estimates about their skills based only a brief description of the skill (e.g., “programming a computer”). So, in this study we similarly provided only brief definitions, and, as in Study 1, compared estimates of skill ratings versus estimates of performance measures. As before, we predicted that hypothetical estimates would show greater asymmetry because lack of experience would contribute to uncertainty about the appropriate scale for absolute judgments, which would in turn lead to a tendency to scale even one’s absolute estimates in relative terms. Naturally, since the questions here are based on even less task information than in Study 1, we expected any such effects to be stronger here than they were in the previous study.

Methods

Participants. The participants were the 20 students from the hypothetical condition of Study 1. Because these participants did not play the games in that study, they completed their estimates more quickly than did those in the experiential condition. So, following the three tasks described in Study 1, these 20 participants were provided with an additional set of 12 tasks to evaluate, which provide the data for this study. They were paid $9 at the end for the combination of the two studies, which required approximately 45 minutes in total.

Materials. We used 12 everyday tasks, similar to those used by Kruger (1999). These are listed on the left side of Table 1. As in Study 1, we manipulated the type of performance estimates asked for (skill or score), with participants randomly assigned to one of those two conditions. For example, in the score condition, participants received the question "How far can you throw a baseball (in feet)?". In the skill condition, "How would you rate your ability to throw a baseball far?". The questions used to elicit score estimates are shown on the right side of Table 1. Skill questions used the same 10-point scale as in Study 1. Judgments of relative ability used the percentile scale with which participants were also familiar from Study 1.

Design. Question type (skill or score) was a between-participants variable. Task was manipulated within participants, with all participants receiving the 12 tasks in the same order.

Procedure. Participants received no description of the 12 tasks other than what was provided in the question (see Table 1). They were asked to answer three questions about their performance on each task, regarding their own predicted performance, the average performance of their peers, and the percentile into which they thought their performance would fall. The order of these three estimates was counterbalanced across participants.

Results

We calculated path weights across task using the same statistical methods as in Study 1 (see Footnote 1). Results are shown in Figure 3.

As shown in Figure 3, we find that *both* skill and score ratings show asymmetrical weighting of own and average performance in these tasks. Taken together with the results

of Study 1, this suggests that ratings of one's own performance are highly correlated with judgments of relative performance percentile if either the estimates are made using a general skill scale or they are made on a purely hypothetical basis. Put differently, participants give nearly equal weight to their own and peer-average performance when they have the combination of a concrete performance metric *and* some exposure to the task requirements. This supports the hypothesis that uncertainty about the appropriate scale for absolute judgments leads judges to treat them more like relative judgments.

Study 3

This study examines two further questions about relative judgments: Is equal weighting of own and others' performance optimal for predicting relative standing, and are judges sensitive to conditions that affect what the optimal weighting is? As we know from the not-necessarily-false-consensus literature, optimal weights in practice may not be the same as those that would theoretically be optimal in an error-free environment. For one thing, estimates about others are likely to be less accurate and reliable than estimates about oneself. Incorporating more poor information about others into one's percentile estimates could actually make them *more* errorful (paralleling Hoch, 1987). We present an illustrative example of a situation in which this is the case. In this study, we used the Word Prospector game, we asked for estimates based on score, and we allowed participants to play the game before responding. Based on the results of our previous two studies, we did not expect to see a great deal of asymmetry in weighting for these tasks. However, we also anticipated that the optimal weighting scheme would in fact put more weight on estimates of one's own performance. Thus, although the prototypical finding is

that judges overweight themselves and should not, we anticipated the opposite pattern in this case: Participants will *not* place much more weight on their own performance, but *should*.

Even if people do demonstrate asymmetric weighting when asymmetric weighting is appropriate, that does not necessarily indicate that they are sensitive to the conditions that make it appropriate. In particular, the weights put on own and others' performance should depend on the accuracy and reliability of each of those sources. To test judges' sensitivity to these variables, we varied the accuracy of each source of information, and looked at how participants' use of the sources of information varied. We did not expect that judges would respond appropriately to differences in information quality. We suspect that any tendency to put more weight on one's own perceived performance stems from a general pattern of experience, rather than from an understanding of the controlling processes. Across life experiences, better absolute performance is in fact correlated with better relative performance. That is tautological, given that one's own performance is part of the function that determines one's relative performance. But the relationship is enhanced in people's experience by the tendency for tests and other tasks to be deliberately adjusted so that the average performance is neither near the floor nor the ceiling. School examinations are a prime example: A low percentage of correct answers on a test is strongly associated with doing poorly relative to others.

Methods

Participants. 95 University of Chicago students were recruited with advertisements posted around campus. They were paid $9 for their participation, which

required approximately 45 minutes. We accepted only students who had not participated in the previous two studies.

Materials. In this study, we gave participants ten different Word Prospector problems of varying difficulty and asked them for estimates about their relative standing on each word individually. We varied the amount of feedback participants received about their own and the average participant's scores.

Design. Task difficulty was manipulated within participants in two ways. Some words were relatively difficult to work with (e.g., *petroglyph* and *gargantuan*) and while some were easy to work with (e.g., *typewriter* and *overthrown*). Secondly, different time limits were provided to work on the words (1, 2, 3, or 4 minutes). All participants received all ten problems in the same order, with the same time limits.

Feedback level varied between participants. In the *self-feedback* condition, participants were told their actual score for each word prior to making their estimates for that word. In the *median-feedback* condition, participants were told the median participant's absolute score on the word after each trial. Those in the *no-feedback* condition received neither kind of information, and those in the *full-feedback* condition received both.[2]

Actual and optimal weights were determined individually for each participant by running regressions for each individual across their ten trials.

Procedure. At the beginning of the procedure, participants read three pages of instructions presented on the computer screen including an explanation of the Word Prospector task, an example, the scoring rules for the task, and the appropriate description of feedback to expect. After reading the instructions, a screen containing the first 10-

letter word and its time limit was presented. After working on the first 10-letter word for the allotted time, the program stopped the game, calculated the participant's score, and provided the feedback appropriate to the participant's condition. Participants then answered questions about the number of points that they expected to receive on that word, the number of points that they expected the average participant to receive, and the percentile rank into which they would fall in relation to other participants. The order of questions was counterbalanced across participants, and they were not asked to make estimates about any quantities for which they received feedback. All participants also estimated the difficulty of the task for themselves and for the average participant, using a scale from 1 (*very easy*) to 10 (*very difficult*).

Participants were then allowed to continue on to the next of the ten words. After completing all ten trials, they were asked to estimate their percentile for Word Prospector tasks in general and the difficulty of the entire task for themselves and for the average participant, and to provide demographic information.

Results

Difficulty effect. Results replicate the difficulty effect reported by Kruger (1999) and Burson et al. (in press). Across tasks, average perceived difficulty and estimates of absolute performance correlated with average estimated percentile, $r(23) = -.42$ and .47, respectively, p's $< .05$.

Weighting and accuracy. We will first consider the no-feedback condition, which most closely resembles the information provided in earlier studies. A paired-sample t test showed that participants in this group did put more weight on estimates of their own score than the peer-average score ($t(23) = 8.307$, $p < .001$); 75% of participants did so.

However, estimates of one's own score and the average score were highly correlated ($r(24) = .95$). We believe this strong correlation reflects individual differences in impressions about the likely results of the Word Prospector scoring system. Regardless, the colinearity of the two variables makes it difficult to assess their separate effects. Therefore, we conducted our subsequent analyses using estimates of own score and estimates of the *difference between own and average peer score (O – P)*. Those two variables correlate more modestly ($r(24) = .57$). With these measures, the equivalent of equal weighting of own and peer performance is for $O - P$ to correlate highly with estimated percentile, and estimates of one's own performance to have no additional contribution to predictions of percentile.

In one-sample *t* tests, the average beta weight for $O - P$ was much greater than zero ($M_{O-P} = .671$, $t(23) = 5.119$, $p = .001$), whereas the average beta weight for own score was not ($M_{own} = .137$, $t(23) = .903$, $p = .376$); participants in this condition were not placing very much additional weight on their own absolute performance.

What *should* they have been doing, in order to maximize the accuracy of their predictions about the percentile into which their performance would fall? To answer this question, we conducted the same analyses using actual percentile of performance as the dependent measure to be predicted from the participants' estimates of their own score and $O - P$. Averaging across participants, the beta weight for participants' estimates of their own score was marginally significantly greater than zero ($M_{own} = .248$, $t(23) = 1.72$, $p = .099$) whereas the beta weight for the difference between own and average score was not ($M_{O-P} = .04$, $t(23) = .262$, $p = .796$). These results suggest that, ideally, participants

should not be relying on the difference between themselves and others to predict their percentile standing, but merely focusing on their own performance.

Next, we looked at all four information conditions to determine how feedback affected participants' weighting on absolute information. We conducted an ANOVA with the dependent variables being the weights given by each participant to own score and *O* – *P*, as in the previous analysis of the no-information condition (which is included here as one condition). Thus, own weight versus *O* – *P* weight was a repeated measure, and condition was a between-participants variable. Results are shown by the darker bars of Figure 4. There are two main effects that are more or less logically inevitable. A marginal condition main effect, $F(3, 91) = 2.34$, $p = .079$, indicates that weights in general increase as more information is provided (i.e., as the dependent variable becomes more predictable). A weighting main effect $F(1, 91) = 23.98$, $p < .001$ shows that more weight is put on *O* – *P* than on own score. More interesting is that there was no condition by weighting interaction ($F(3, 91) = .084$, $p = .968$). In other words, participants did not significantly adjust their weights as information became more diagnostic.

Next, we looked at how feedback affects the optimal weighting for predicting percentile from estimates of *O* – *P* and own score. We expected that when objective information was provided on a given variable, the variables in question would optimally be weighted more heavily. We repeated the previous ANOVA, but using the weights for each participant that predicted their actual percentile rather than their estimated percentile. Results are shown in the lighter bars of Figure 4. There were source and condition main effects similar to the previous analysis. This time, though, there was a

condition by weighting interaction ($F(3, 91) = 5.70$, $p = .001$), indicating that that optimal weighting of $O - P$ and own score did vary significantly as information was provided.

Discussion

The results of this study demonstrate two complementary considerations in interpreting the weights that judges of relative performance give to their own performance and that of their average peer. First, as in our previous studies, estimates of concrete performance standards, informed by task experience, show only modest asymmetry of weights. This is shown in the present study by the fact that the estimated difference between own and average performance largely accounts for predicted percentile of performance relative to peers. The additional contribution of own performance is small. The second consideration is, what would the optimal weighting be? We find that, when basing one's predictions just on subjective estimates, it is actually not such a good idea to lean so heavily on the estimated self-other difference. So, in sum, people are not always prone to an egocentric focus, and they might sometimes be better off if they were.

Looking across information conditions, we see that judges are not very sensitive to the effect that additional information should have on how they estimate their relative performance. In low-information conditions, they tend to put too much weight on *all* of their estimates. In other words, people fail to appreciate how inaccurate their estimates are, or they fail to appreciate that such inaccuracy implies that they should regress their impressions heavily toward the mean. When judges are provided with objective feedback concerning their own absolute performance, the median performance of their peers, or both, they show little change in judgmental policy, whereas of course they should.

Roughly speaking, people seem to make judgments as though their subjective estimates were as good as objective information.

General Discussion

The three studies reported here shed light on the process decision makers use to evaluate themselves relative to others. We show in the first study that judgments of relative standing are not as egocentric as some previous studies suggest. Studies such as Klar and Giladi (1999), Kruger (1999), and Kruger and Burrus (2004) find a very strong association between perceptions of one's own abilities and perceptions of one's abilities relative to others, with much less influence of perceptions of the average abilities of one's peers. However, our results suggest that this stems, at least in part, from an ambiguity in judges' interpretation of the question. When asked to rate their skill on an arbitrary scale such as 1 to 10, relative judgment is used implicitly to set the scale values. It may not always be the case that 5 is taken to be the precise average value for one's peers—sometimes a different reference group may be salient (such as those who regularly engage in the activity described). But still, low numbers mean "worse than most" and high numbers "better than most," even if the experimenter had something more absolute in mind. This interpretation is supported by our findings using well-defined scales such as points scored or time to completion. Units on these scales can readily be defined in an unambiguously absolute sense, without reference to a comparison group, and judges show more nearly equal weighting of estimates for own and average peer performance when predicting relative performance. If a relative judgment is used to set the absolute scale values, then it is inevitable that where one puts oneself on the absolute scale will

correlate highly with where one puts oneself on the relative scale, without implicating an egocentric focus.

The second study finds that, even when asking about scores, there is substantial asymmetry of weights when participants made judgments without experiencing the task or receiving detailed descriptions and examples. Hypothetical measures are, we conclude, also difficult to scale, and thus judgments of one's score are derived, in part, from one's guess about how one stands relative to others, rather than vice versa. In other words, a judge may have a clearer impression about where he stands relative to his peers in baseball throwing than he has about how far a typical baseball can be thrown. Again, this uncertainty about scaling will lead to a strong correlation between estimates of one's absolute and relative performance without egocentric focus.

In our final study, we examined what weighting scheme would allow participants to best predict their relative standing from their estimates of their own performance and of the peer average, or from objective information provided on one or both variables. Using the example of Word Prospector games with task experience, we demonstrate that it is not necessarily optimal to weight estimates of own and others' performance equally. Generally, one has a better idea of one's own performance than of the average performance of peers. Thus, it can be optimal to put more weight on the former than on the latter. But judges do not seem to be sensitive to the fact that this balance depends on the sources of information. They use virtually the same weighting scheme whether their information comes from subjective estimates or from objective feedback, as though ignoring the uncertainty in their subjective estimates or the implications of that uncertainty for appropriate weighting.

The idea of egocentric focus in judgments of relative performance or skill is intended to explain a very robust and interesting phenomenon, namely the tendency for judgments of relative ability to correlate with judgments of one's own performance. With easy tasks, people on average think they are above average; with difficult tasks, people think they are below average. Nothing in our results contradicts this empirical finding. In Study 3, as well as in Burson et al. (in press), we replicate this difficulty effect even using concrete performance measures and task experience. Even in such tasks, there is still some tendency for judges to lean more on their impression of their own competence than they do on their impression of the average peer's performance, and that residual asymmetry is enough to maintain the difficulty effect.

When response scales have clear external referents, the amount of extra weight given to one's own ability is not extreme. In fact, it turns out that a moderate degree of imbalance often increases accuracy rather than hurting it, a finding that is paralleled in the "false consensus" literature. Presumably, people have learned from life experience that one's absolute performance is indeed a cue to one's relative performance, and have learned to weight the former heavily in predicting the latter. But the optimal weighting of cues of course depends on one's sources of information, and people do not seem to recognize when or why weighting one's own performance heavily is good policy. That means that they are likely to make systematic errors depending on what the kinds of experience and feedback they receive in a domain. The effects of experience and feedback on relative judgments are deserving of further research. Meanwhile, the present studies demonstrate some of the important descriptive and normative issues surrounding this important domain of judgment.

References

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research, 27*(September), 123-156.

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist 58*(12), 1019-1027.

China Daily (1999). Students underestimate scores. *People's Daily Online*, July 27, 1999.

Davis, J. L., Hoch, S. J., & Ragsdale, E. K. E. (1986). An anchoring and adjustment model of spousal predictions. *Journal of Consumer Research, 13*, 25-37.

Dawes, R. M. (1990). The potential non-falsity of the false consensus effect. In R. M. Hogarth, (Ed.), *Insights in decision making: A tribute to Hillel T. Einhorn*. Chicago: University of Chicago Press, 97-110.

Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behavior and Human Decision Processes, 65*(3), 201-211.

Giladi, E. E., & Klar, Y. (2002).When standards are wide of the mark: Nonselective superiority and inferiority biases in comparative judgments of objects and concepts. *Journal of Experimental Psychology: General, 131*(4), 538-551.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*(2), 221-234.

Hoch, S. J. (1988). Who do we know: Predicting the interests and opinions of the American consumer. *Journal of Consumer Research, 15*(December), 315-324.

Klar, Y., & Giladi, E. 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*(5), 885-901.

Klar, Y., & Giladi, E. E. (1999). Are most people happier than their peers, or are they just happy? *Personality and Social Psychology Bulletin, 25*(5), 585-594.

Klar, Y., Medding, A., & Sarel, D. (1996). Nonunique invulnerability: Singular versus distributional probabilities and unrealistic optimism in comparative risk judgments. *Organizational Behavior and Human Decision Processes, 67*(2), 229-245.

Kruger, J. (1999). Lake Wobegon be gone! The "Below-Average Effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*(2), 221-232.

Kruger, J., & Burrus, J. (2004) Egocentrism and focalism in unrealistic optimism (and pessimism). *Journal of Experimental Social Psychology, 40,* 332-340.

Moore, D. A., & Kim, T. G. (2003). Myopic social prediction and the solo comparison paradox. *Journal of Personality and Social Psychology, 85*, 1121-1135.

Prelec, D., Wernerfelt, B., & Zettelmeyer, F. (1997). The role of inference in context effects: Inferring what you want from what is available. *Journal of Consumer Research, 24*(June), 118-125.

Wernerfelt, B. (1995). A rational reconstruction of the compromise effect: Using market data to infer utilities. *Journal of Consumer Research, 21*(March), 627-633.

Windschitl, P. D., Kruger, J., & Simms, E. N. (2003). The influence of egocentrism and focalism on people's confidence in competitions: When what affects us equally affects me more. *Journal of Personality & Social Psychology, 85*, 389-408.

Author Note

We thank Maciej Szefler for programming assistance on Studies 1 and 3.

Correspondence concerning this article should be addressed to Katherine A. Burson, University of Michigan Business School, 701 Tappan St., Ann Arbor, MI, 48109, kburson@umich.edu.

Footnotes

[1] We used a method to determine weights that is equivalent to the LISREL method used by Kruger (1999). We regressed estimated percentile on estimates of own absolute performance and average peer performance; the path weights are the beta weights from that regression. The path weight between own absolute performance and average peer performance is the simple bivariate correlation. In order to be able to combine data from all three tasks, estimates for own and peer-average scores were standardized according to the distribution of these estimates from a given task within a given question-type by experience cell across participants.

[2] In order to provide realistic feedback on median performance, the self-feedback and no-feedback conditions were run first, and the median performance from those conditions was reported to participants in the median-feedback and full-feedback conditions.

Table 1

Tasks used in Study 2, with the question used to elicit estimates of own performance in the score condition.

Throwing a Baseball	How far can you throw a baseball?
Baking Bread	How many times have you baked bread?
Recreational Reading	How many non-school books do you read in a year?
Telling Jokes	Out of 10 jokes you tell, how many will people laugh at?
Calculus Problems	How many questions on a 20 question intro calc. test can you answer right?
Hammering Nails	What percent of nails that you hammer get bent?
Capitals	How many of the 50 state capitals do you know?
Car	What percent of car problems can you diagnose?
Orderly	What percentage of your stuff is where it belongs?
Save	For every $100 that you should save, how much do you actually save?
Sick	How many days were you sick during the last school year?
Late	What percent of your social engagements are you more than 15 minutes late to?

Figure Captions

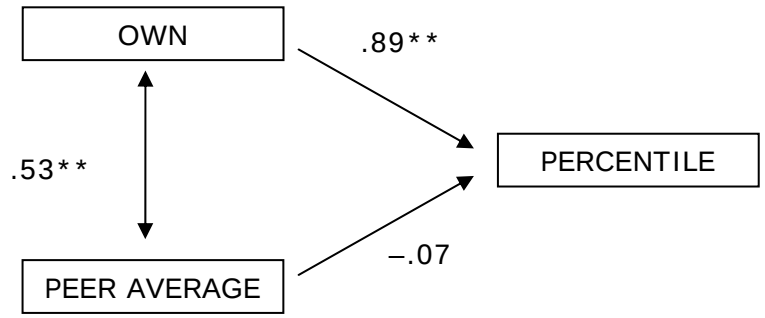
Figure 1. Path weights predicting estimates of performance percentile from estimates of one's own performance and estimates of average peer performance, and correlations between own and peer estimates, from Kruger (1999). Two asterisks indicates significant difference from 0, $p < .01$

Figure 2. Path weights predicting estimates of performance percentile from estimates of one's own performance and estimates of average peer performance, and correlations between own and peer estimates, by question type and task experience in Study 1. A + superscript indicates different from 0, $p < .10$; one asterisk, $p < .05$; two asterisks, $p < .01$.

Figure 3. Path weights predicting estimates of performance percentile from estimates of one's own performance and estimates of average peer performance, and correlations between own and peer estimates, by question type in Study 2. Asterisks represent values different from 0, $p < .05$.

Figure 4. Darker bars shown the beta weights predicting estimated performance percentile from estimates of (own score – average peer score) and own score. Lighter bars show the weights that would optimally predict actual (rather than estimated) percentile.

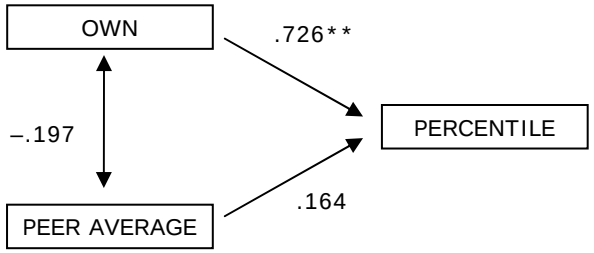
1



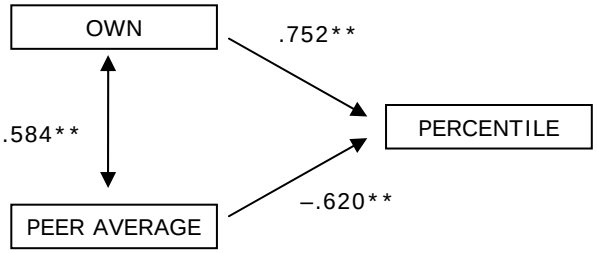
2

Hypothetical

Skill

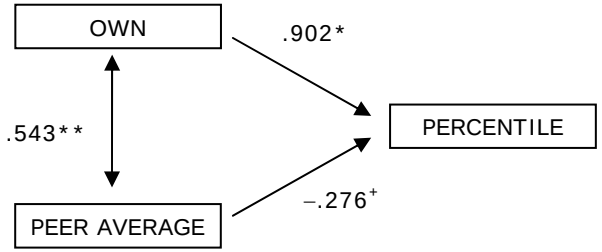


Score

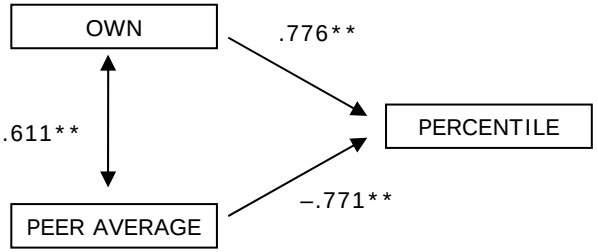


Experiential

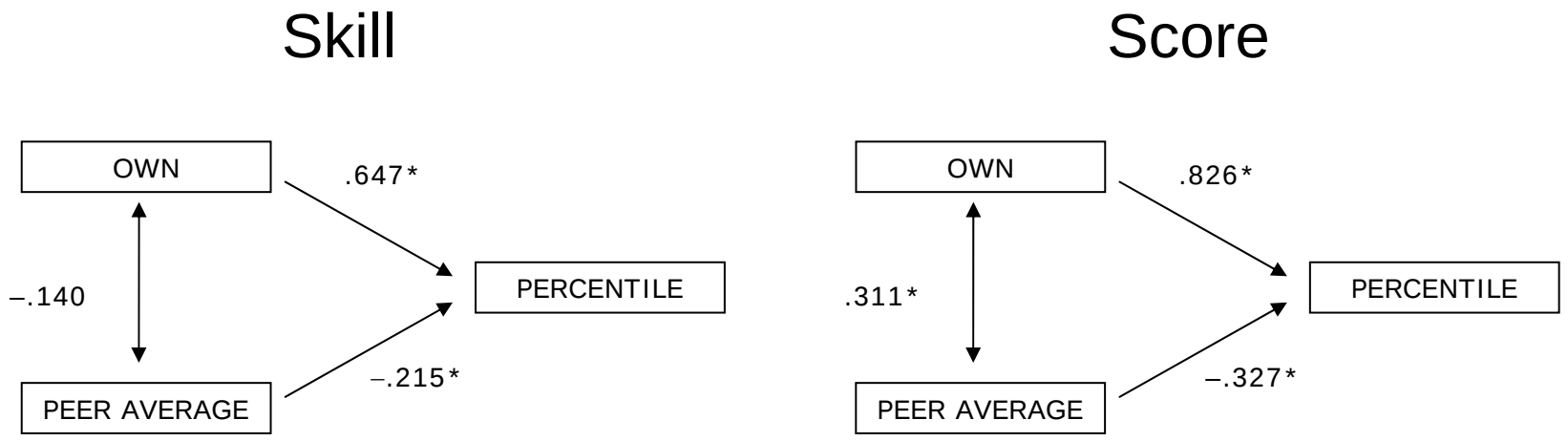
Skill



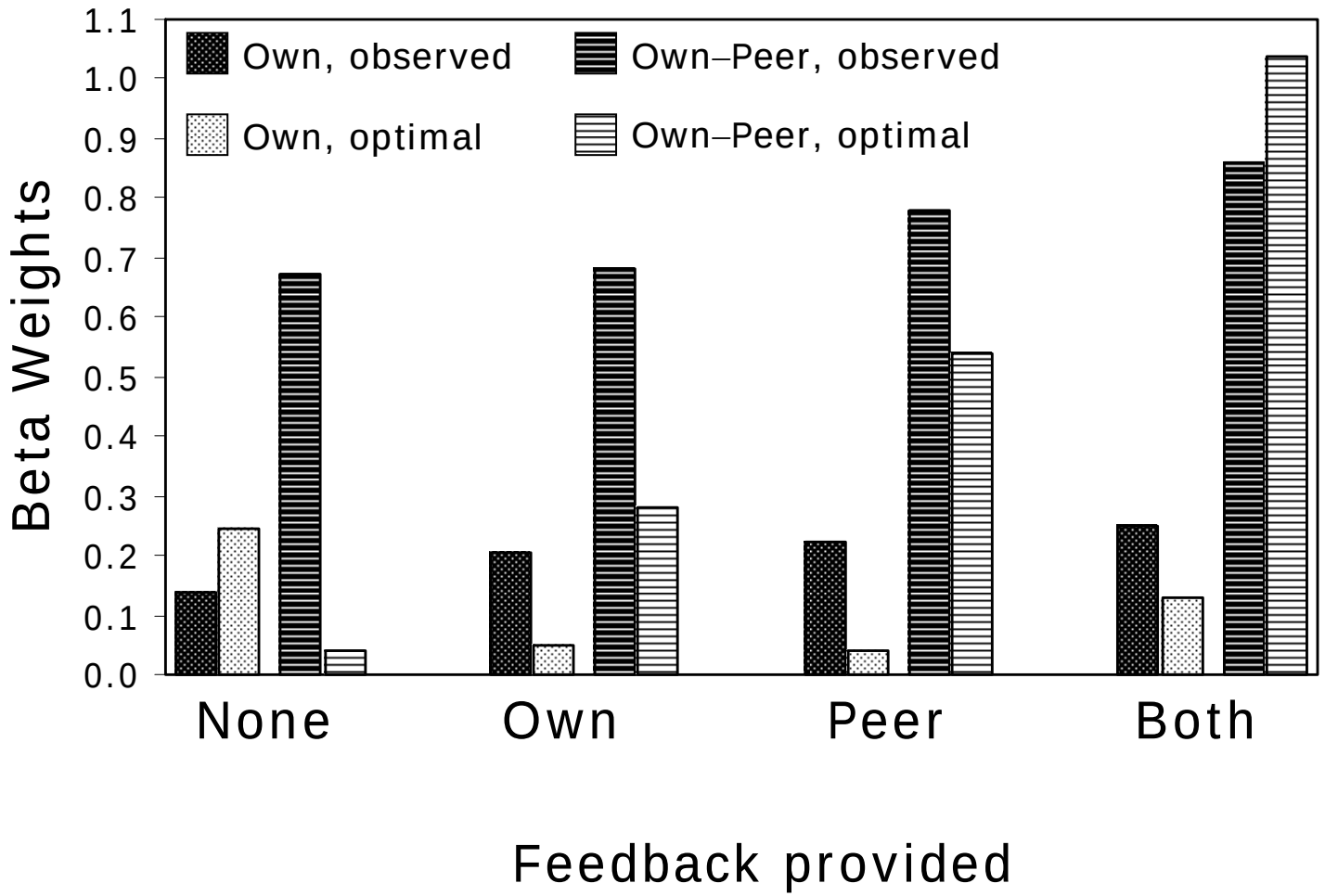
Score



3


Skill
OWN
.647*
–.140
PERCENTILE
–.215*
PEER AVERAGE
Score
OWN
.826*
.311*
PERCENTILE
–.327*
PEER AVERAGE

4


Own, observed
Own–Peer, observed
Own, optimal
Own–Peer, optimal
Beta Weights
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
None
Own
Peer
Both
Feedback provided