Working Paper

Skilled or Unskilled, but Still Unaware of It: How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons

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

Richard P. Larrick
The Fuqua School of Business
Duke University

Joshua Klayman
Graduate School of Business
University of Chicago

Ross School of Business Working Paper Series
Working Paper No. 956
December 2005

Under review at the *Journal of Personality and Social Psychology*

Running Head: DIFFICULTY AND MISCALIBRATION

REVISION 10

Skilled or Unskilled, but Still Unaware of It:

How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons

Katherine A. Burson

University of Michigan

Richard P. Larrick

Duke University

Joshua Klayman

University of Chicago

Under review at *Journal of Personality and Social Psychology*.

Abstract

People are inaccurate judges of how their abilities compare to others'. Kruger and Dunning (1999; 2002) argue that most inaccuracy is attributable to unskilled performers' lack of metacognitive skill to evaluate their performance. They overestimate their standing, whereas skilled performers accurately predict theirs. Consequently, the majority of people believe they are above average. However, not all tasks show this bias. In a series of ten tasks across three studies, we show that moderately difficult tasks produce little overall bias and little difference in accuracy between best and worst performers, and that more difficult tasks produce a negative bias, making the worst performers appear more accurate in their judgments. This pattern suggests that judges at all skill levels are subject to similar degrees of inaccuracy and bias. Although differences in metacognitive ability may play a role in the accuracy of interpersonal comparisons, our results indicate that, for the most part, the skilled and the unskilled are equally unaware of how their performances compare to those of others.

Skilled or Unskilled, but Still Unaware of It:
How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons

Research on overconfidence has found that subjective and objective measures of performance are poorly correlated (see Alba & Hutchinson, 2000 for a comprehensive review). While most of this research compares confidence in one's estimates with one's actual performance, one particular vein focuses on people's accuracy in estimating their ability compared to their peers. Such judgments are important in many contexts. In many societies, success in school, jobs, entrepreneurship, sports, and many other activities are largely a function of how one's ability and performance compare to others'. Thus, the ability to estimate one's relative standing can have a major impact on one's life choices and one's satisfaction with those choices.

The most common finding in this area is a "better-than-average" effect: On average, people think that they are above average. However, this tendency is not uniform. The overestimation comes mostly from poor performers. Figure 1 summarizes results from studies by Kruger and Dunning (1999) showing this effect. Kruger and Dunning (1999; 2002) argue that this happens because people who perform poorly at a task also lack the metacognitive skill to realize that they have performed poorly. On the other hand, people who are more skilled have both the ability to perform well and the ability to accurately assess the superiority of their performance. Borrowing from the title of Kruger and Dunning's paper, we refer to this as the "unskilled-unaware hypothesis."

The unskilled-unaware hypothesis has logical and intuitive appeal. As Kruger and Dunning (1999) point out, the skills it takes to write a grammatically correct sentence are the same skills it takes to recognize a grammatically correct sentence. The most incompetent

individuals overstate their abilities in many contexts. One of this paper's authors spent several years leading horseback rides and was struck by the number of incompetent riders who actually put their lives in danger by claiming that they were highly skilled. However, Kruger and Dunning look at only one judgment context—one in which participants on average believe they are above average. In fact, research by Kruger (1999) showed that this condition is not as universal as it was thought to be. He found that on easy tasks (such as using a computer mouse), people estimate their performance as better than average, whereas on hard tasks (such as juggling), people estimate themselves as worse than average. He argues that participants anchor on their perception that they will perform well or poorly in an absolute sense and adjust insufficiently for the fact that the task may be easy or hard for everyone.

Are the unskilled responsible for most of the error in more difficult tasks as well, when the average judgment is unbiased or negatively biased? This question is important because the answer can help distinguish the unskilled-unaware hypothesis from a simpler alternative explanation for the pattern illustrated in Figure 1. The alternative hypothesis, proposed by Krueger and Mueller (2002), is that people at all skill levels are prone to similar difficulties in estimating their relative performance. Their subjective estimates of performance are imperfectly correlated with objective performance measures, so estimates of relative performance regress toward the mean.[1] Additionally, people at all skill levels make estimates of relative performance that are biased upward. In other words, regardless of skill level, people do not have much knowledge about how they compare to others, and the average estimates of poor and good performers tend to be similar and high. Good performers are more accurate, but not because of greater metacognitive skill. Rather, when most participants estimate their performance as better than average, those who actually *are* above average are necessarily closer to the truth. Kruger

and Dunning (1999; 2002) and Krueger and Mueller (2002) examine judgments of relative standing in tasks with overall positive biases. The two explanations are difficult to distinguish in that context. In a published exchange, the two sets of authors focus on the question of whether metacognitive skills can be shown to mediate the difference between good and poor performers, and they disagree. In the end, the evidence provided by Kruger and Dunning and Krueger and Mueller remains equivocal on whether population-level errors in interpersonal comparisons should be attributed mainly to the metacognitive failings of poor performers.

In the present studies, we take a different approach to investigate the cognitive processes underlying judgments of relative standing. We examine a range of judgment contexts that vary in perceived difficulty, including several for which there is no overall bias or a negative bias. This permits us to test a generalization of Krueger and Mueller's (2002) basic hypothesis. That is, people at all performance levels are equally poor at estimating their relative performance, and equally prone to underestimate their relative standing on tasks that are perceived to be hard and overestimate it on tasks that are perceived to be easy. The results expected under this hypothesis are illustrated in Figure 2. (See the appendix for the model used to simulate these results). An interesting implication is that higher-skilled performers are better judges of their relative standing only for easy tasks. For difficult tasks, the opposite is true: The most skilled are the *least* accurate. Poor performers account for most of the above-average effect in easy tasks, whereas good performers account for most of the below-average effect in difficult tasks.

If judgments of relative standing show the pattern in Figure 2 (i.e., parallel lines with modest upward slopes), then noise and bias across all performers provide a sufficient explanation. There is no need to appeal to metacognitive differences between better and worse performers. On the other hand, significant departures from this pattern suggest that such

differences may be important. Figure 3, for example, shows one plausible instantiation of the unskilled-and-unaware hypothesis. Here, worse performers are more prone to random error and to difficulty-related biases. (See the Appendix for details of the simulation.)

In the remainder of this paper, we describe three experiments that manipulate the perceived difficulty of tasks and hence participants' beliefs about their relative standing. By sampling a wider range of judgment contexts, these studies provide evidence on a fundamental question about the psychological processes underlying comparative judgments in these kinds of tasks: Does miscalibration reflect the poor insight of poor performers or the poor insight of all performers?

Study 1

Method

Participants. Ninety University of Chicago students were recruited using posted advertisements and paid two dollars for participating in this 15-minute experiment.

Design. In this between-participants design, 47 students took an easier quiz about University of Chicago trivia and 43 students took a harder quiz about University of Chicago trivia. Care was taken to ensure that the harder task was not below some "minimal threshold of knowledge, theory, or experience" as cautioned by Kruger and Dunning (1999, p. 1132); the harder trivia were answered well above chance level (as, of course, were the easier trivia).

Procedure. Participants were told that they would be taking a 20-question quiz about the University of Chicago. They were given a two-page quiz (either easier or harder). After taking the quiz, participants estimated the number of questions out of 20 that they thought they would get right, the percentile rank into which they believed they would fall in relation to their peers, and the difficulty of the task for themselves and for the average participant on a 1 (*very easy*) to 5

(*very difficult*) scale. The use of the percentile scale was explained in detail. The order of the performance estimates and the questions about difficulty was counterbalanced: For half the participants, performance estimates appeared first followed by the difficulty questions, and for half it was the reverse.

Results

Manipulation check. The order of the performance estimates and the difficulty estimates did not lead to a difference in estimates, so we collapsed across orders. As expected, the harder trivia resulted in a lower actual score than the easier trivia (M = 10.62 versus M = 14.64), $t(87) = 7.53$, $p < .001$, $d = 1.60$ and both were better than chance level of 6.67, ts > 10.86, ps < .001. The harder trivia were also rated as significantly more difficult than the easier trivia (M = 3.91 versus M = 2.96), $t(87) = 5.24$, $p < .001$, $d = 1.11$.

Percentile estimates. Next, we looked at percentile estimates at each level of difficulty. Participants estimated their performance to be in the 62nd percentile for the easier trivia and in the 48th percentile for the harder trivia, $t(88) = 3.66$, $p < .001$, $d = .77$. This replicates the results of Kruger (1999) that the more difficult the task, the lower the overall percentile estimate. As hoped, two distinct levels of difficulty were sampled in this study. In this case, our harder trivia turned out to be only moderately difficulty while the easier trivia was indeed easy. Therefore, we will call our harder condition the "moderate" condition.

Asymmetry by quartiles. To examine how estimated percentiles varied with skill level, we divided the participants in each condition into four groups based on performance.[2] These groups represented four quartiles of performance relative to other participants in that condition. As shown in Figure 4, percentile estimates are fairly uniform across quartiles on both the easy and the moderate task and are lower on the more difficult task. An ANOVA on percentile estimates

with the independent variables of difficulty and quartile showed the main effect of task difficulty already discussed. There was also a marginal main effect of quartile, $F(3, 81) = 2.68$, $p = .05$, $\eta^2 = .09$, but no significant interaction. The main effect of quartile was tested with a polynomial contrast and showed a significant linear trend ($p = .022$); as quartiles increased, so did percentile estimates. We also explored this relationship by regressing percentile estimates on actual percentiles. This analysis supported the linear trend above (B = .142, SE = .068, β = .219, $t(87) = 2.097$, $p = .039$).

Paired *t* tests confirm some of Kruger and Dunning's (1999) findings. In both conditions, those in the bottom quartile overestimated their percentile, and those in the top quartile underestimated theirs. Participants in the bottom quartile on the easy trivia were actually in the 12th percentile but thought they would be in the 57th ($t(11) = 8.35$, $p < .01$, $d = 2.32$) and on the moderate trivia they were actually in the 9th percentile but thought they would be in the 48th ($t(7) = 4.25$, $p = .004$, $d = 1.40$). Participants in the top quartile on the easy trivia were actually in the 89th percentile but thought they would be in the 72nd ($t(9) = -3.09$, $p = .013$, $d = -.93$) and on the moderate trivia they were actually in the 86th percentile but thought they would be in the 58th ($t(10) = -4.43$, $p = .001$, $d = -1.28$). To compare the magnitude of errors of top and bottom performers, we coded errors as (estimated percentile – actual percentile) for the bottom quartile and (actual percentile – estimated percentile) for the top quartile, and then compared the two quartiles. (This simple transformation preserves the variance around the means, but gives the means the same sign so that they can be tested against each other.) On the easy quiz, we replicated the asymmetry observed by Kruger and Dunning; the lowest quartile was much more miscalibrated than the highest ($M = 44.34$ versus $M = 16.84$), $t(20) = 3.59$, $p = .002$, $d = 1.54$.

However, in the moderate condition, the first and fourth quartiles did not differ significantly (M = 39.23 versus M = 28.13), $t(17) = 1.03$, $p = .32$, $d = .48$.

Discussion

As can be seen in Figure 4, percentile estimates varied only slightly with actual performance. Difficulty lowered estimates for low and high performers alike. Thus, in the absence of an overall upward bias, the skilled and the unskilled are similarly accurate. There is no evidence that the results are driven by any special lack of metacognition on the part of the unskilled participants; these results are consistent with the hypothesis that estimating one's percentile is difficult regardless of skill level.

Study 2

The next experiment looks more closely at the psychological underpinnings of the observed pattern by using tasks that were perceived to be more difficult than those used in Study 1. Participants perceived Study 1's stimuli as easy and moderately difficult. If unawareness is universal, then it will be the *unskilled* participants who will appear to be more aware of their relative standing in more difficult tasks, in which the average percentile estimate is less than 50. This is illustrated by the lowest line of Figure 2.

As in Study 1, we manipulated perceived difficulty by sampling a variety of stimuli, but this time we compared what turned out to be moderate and difficult conditions. We used two manipulations to create the desired range of perceived difficulty: We selected several domains of trivia questions that we expected to vary in perceived difficulty, and we manipulated the strictness of the criterion for judging an estimate to be correct. Our prediction was that domains that were perceived to be more difficult and criteria that were more exacting would lead to significantly lower perceived percentiles.

We hypothesized that as the perception of task difficulty increased, low performers would appear to be more accurate and high performers less accurate. If the task is difficult enough to produce below-average estimates overall, low performers should be more accurate in their estimates than the high performers are (as in the lowest lines of Figure 2). We want to emphasize that this pattern should not be interpreted as showing that poor performers are actually more perceptive than high performers. Rather, in a task in which everyone is biased toward believing their performance is poor, those whose performance truly is poor will appear to be right.

Method

Participants. Forty University of Chicago students were recruited with posted advertisements and were paid nine dollars for this 45-minute experiment.

Design. Three variables were manipulated within participant: domain, question set, and difficulty. There were five domains: college acceptance rates, dates of Nobel prizes, length of time pop songs had been on the charts, financial worth of richest people, and games won by hockey teams. For each domain, there were two subsets of 10 questions each. These questions were selected randomly from the available information sources. Each 10-question subset was presented in either a harder or an easier version. The more difficult version required participants' estimates to fall within a narrower range to be considered correct (e.g., within 5 years of the correct date, vs. 30 years in the harder version).

The order of the 100 estimates was the same across participants, consisting of 10 questions from each of the five domains, followed by another 10 questions from each of the five domains. The order of difficulty was counterbalanced. Half the participants received the first five

subsets of questions in the harder version and the second five in the easier version. For the other half, the first five subsets were in the easier version and the second five in the harder version.

Two domains (financial worth and hockey) included tests that were so difficult or so easy that almost all of the participants got nearly all or nearly none right, making it hard to distinguish levels of performance. We dropped these two domains from the analyses.

Procedure. Participants were told that they would be making a series of estimates about a range of topics. They were given a booklet containing 10 subsets of estimates preceded by an unrelated example. One page was devoted to each subset of questions. For each of the 10 subsets, participants indicated their predicted percentile rank, the difficulty of the task for themselves, and the difficulty of the task for the average participant on a 1 (*very easy*) to 10 (*very difficult*) scale. Prior to each set of 10 questions, participants read an explanation of the required estimates, along with information about the mean of the sample and the range in which 90% of the sample fell. For instance, when making estimates of years of Nobel Prizes in the easier version, participants read:

> In this section, you will estimate the year in which particular people received the Nobel Prize in Literature. You should try to be accurate within 30 years of the truth. These 10 Nobel Laureates were selected randomly from the 100 Nobel Laureates in Literature. Within the 20 Laureates in this packet, the average year of the Nobel Prize is 1949 and 90% of the Laureates fall between 1921 and 1985.

In the harder version of the test, participants had to give an estimate within five years of the actual year.

Results

Manipulation check. A repeated measures MANOVA was performed with actual performance, estimated performance, and estimated difficulty as dependent measures. Domain and difficulty were within-participant variables and order (harder first or easier first) was a between-participant variable. The difficulty manipulation worked; the harder conditions were perceived as significantly more difficult ($M = 7.94$) than the easier versions ($M = 6.59$), $F(1, 35) = 30.43$, $p < .001$, $\eta^2 = .47$. Harder and easier conditions also differed significantly in actual performance ($M = 19.84\%$ correct versus $M = 68.77\%$ correct), $F(1, 35) = 808.15$, $p < .001$, $\eta^2 = .96$).

We also checked the extent to which different subsets of estimates provided independent tests of relative ability. For each of the 12 subsets of estimates (three domains x two subsets x two difficulty versions), we divided the participants into quartiles of performance relative to the performance of other participants on the same subset of estimates. We then compared participants' quartiles on one subset of estimates to their quartile on a different subset. The subsets proved to be largely independent in terms of relative performance. Correlations for the (12 x 11) / 2 = 66 pairs of subsets ranged from -.39 to .31, with a median of .02.

Percentile estimates. Overall, the mean percentile estimate was 37.04. This was significantly less than 50, $t(39) = -4.68$, $p < .001$. The repeated measures MANOVA showed that some domains (like Nobel Prize dates) seemed more difficult than others ($M_{\text{colleges}} = 6.36$, $M_{\text{pop songs}} = 7.17$, and $M_{\text{Nobel Prize}} = 8.19$), $F(2, 70) = 15.16$, $p < .001$, $\eta^2 = .30$. Furthermore, the percentile estimates tracked these perceptions of difficulty ($M_{\text{colleges}} = 45.98$, $M_{\text{pop songs}} = 39.47$, and $M_{\text{Nobel Prize}} = 26.98$); the more difficult the domain seemed to participants, the lower the percentile estimate, $F(2, 70) = 14.25$, $p < .001$, $\eta^2 = .29$. Also, percentile estimates were lower in

the difficult (narrow range) versions than in the easier versions, $F(1, 35) = 22.57$, $p < .001$, $\eta^2 = .39$ (see Table 1). In other words, average percentile estimates decreased as tasks became more difficult (through more stringent evaluation standards or domain differences). This replicates the effect reported by Kruger (1999). There was no effect of order or any significant two-way interactions. However, there was an unexpected three-way interaction between domain, difficulty, and order, $F(2, 70) = 7.18$, $p < .001$, $\eta^2 = .17$, the implications of which are unclear.

Asymmetry by quartiles. As shown in Figure 5, the overall picture is one of a fairly uniform level of percentile estimates across quartiles within each domain. For those in the top quartile, estimated percentiles were significantly lower than actual percentiles in each of the combinations of domain and difficulty. For those in the bottom quartile, estimated percentiles were significantly higher than actual percentiles in most cases (see Table 1).

To compare the errors of best and worst performers, we coded errors as (estimated percentile – actual percentile) for the bottom quartile and (actual percentile – estimated percentile) for the top quartile, and then compared the two quartiles. In both subsets of college acceptance rates, the estimated performance was near 50 across quartiles. In those moderately difficult subsets, the mean estimation error was of approximately the same magnitude in the lowest and highest quartiles, $t(13) = -1.04$, $p = .32$, $d = -.52$ in the easier condition and $t(20) = -.61$, $p = .55$, $d = -.25$ in the harder condition), replicating the moderate condition of Study 1. In the two subsets of Nobels, average percentile estimates across quartiles were well below 50. In this domain, we see a reversal of the asymmetry reported by Kruger and Dunning (1999): Underestimation in the highest quartile was much larger than overestimation in the lowest quartile, $t(20) = -3.11$, $p = .006$, $d = -1.30$ in the easier condition and $t(13) = -5.00$, $p = .000$, $d = -2.48$ in the harder condition). In this harder domain, it was the *skilled* participants who appeared

more unaware. Similarly, in the two subsets of pop music, average percentile estimates across quartiles were below 50 and the unskilled performers were as accurate as the highly skilled. The results of this domain fell between colleges and Nobels, with a nonsignificant trend toward higher estimation error in the top quartile, $t(19) = -.82$, $p = .43$, $d = -.35$ in the easier condition and $t(20) = -2.00$, $p = .06$, $d = -.82$ in the harder condition.

The overall pattern of difficulty ratings and miscalibration shows that the difference in relative miscalibration between high and low performers is a direct function of perceived task difficulty. In other words, who looks more accurate depends on the difficulty of the task simply because difficulty affects estimates of relative ability (but not actual relative ability). The difference in miscalibration between high and low performers correlates with perceived task difficulty ($M_{\text{easier colleges}} = 5.35$, $M_{\text{harder colleges}} = 7.15$, $M_{\text{easier pop}} = 6.77$, $M_{\text{harder pop}} = 7.62$, $M_{\text{easier Nobel Prize}} = 7.55$, $M_{\text{harder Nobel Prize}} = 8.88$) at $r(6) = -.70$, $p = .12$.

Discussion

The results of this study are consistent with Krueger and Mueller's (2002) hypothesis that skilled and unskilled people are similarly unaware of how they perform relative to others. The relative degree of miscalibration between low and high performers is driven by the task difficulty: With harder domains that feel harder (Nobel Prizes) and with more stringent criteria, low performers are better-calibrated than high performers. However, just as the apparent unskilled-unaware effect is largely a function of perceived task difficulty, so is the appearance in hard domains that the *un*skilled are more aware.

In the present study, as in Study 1, we find only a weak positive relation between objective and subjective measures of relative performance. Good and poor performers alike seem to have limited insight into how their skills and abilities compare to others, and it is task

difficulty that determines whether high or low performers appear better calibrated. Alternatively, these might be tasks for which relative performance is inherently unpredictable. If so, we might not have provided the high performers with adequate opportunities to demonstrate their superior metacognitive abilities. Kruger and Dunning (2002) make a similar point in their critique of Krueger and Mueller's (2002) studies, although they focus on task reliability rather than predictability per se. (We will elaborate on the difference between reliability and predictability in the General Discussion.)

Reliability in the 12 subdomains of the present study ranged from poor to moderate (Spearman-Brown's from -.24 on one set of easier pop music estimates to .52 on one set of harder Nobel Prize estimates). Our "unskilled-aware" effect holds even within the latter, most reliable subdomain, $t(7) = -3.71$, $p = .008$. However, one might wish to have more and stronger evidence about the relation between skill level and estimates of relative standing in more reliable, predictable tasks.

Study 3

In this study, we use a task that is more amenable to prediction of one's relative standing than were our previous tasks. In line with Kruger and Dunning's (2002) focus, the selected task is highly reliable; it also has other features that may help participants to some degree in judging their relative standing. The task we chose was a "word prospector" game. In this game, the player attempts to construct as many four, five, and six letter words as possible from the letters contained in one 10-letter word. For example, from the word "typewriter" one can construct type, writer, trite, pewter, etc. Participants receive some performance feedback, in that they can score their own word lists as they produce them. However, as in previous studies, the participants do not receive reliable, objective feedback during the task. Those with poor spelling or weaker

vocabularies might mistakenly believe that they will get credit for, say, *weery* or *twip*. The other component of relative standing is of course the performance of others. Here, too, participants may have some information to go on, but it is limited. They may have a general sense of where they stand on games and tasks involving spelling and vocabulary, but lacking specific feedback on other people's performance, they cannot know where a (self-calculated) score of say, 37, would put them in the distribution.

In this study we gave each participant two different word prospector problems of similar difficulty and asked them for estimates about their relative standing on each word individually and overall. This facilitated two approaches for comparing predicted to actual performance at different levels of ability. The first approach is the same as that used in all previous studies: Participants are separated according to their total performance on both subtasks. Because the word prospector task has good reliability, this gives us a stable measure of each participant's ability.

The second approach is to separate participants according to their performance on one subtask, and measure how accurately they estimated their relative performance on *the other* subtask. This method provides a noisier measure of ability, but it avoids the possible biasing effects of mean reversion in comparing poor and good performers. Those found in the bottom quartile or the top quartile on a given test appear there partly because of ability and partly because of bad and good luck, respectively. Even in tasks that are largely skill based, judges cannot perceive all the elements of good and bad luck that contributed to their high or low performance. Thus, their estimates of their performance will naturally be regressive, and this will be counted as error. Given reasonable reliability, the worst and best performers will still do poorly and well, respectively, on the other test, but now good and bad luck will be equally

distributed among them, on average. Thus, judging ability on one subtask and measuring estimated and actual relative performance on another subtask provides a luck-neutral (i.e., mean-zero error) way of comparing good and poor performers.[3]

Method

Participants. As in Study 2, 76 University of Chicago students were recruited with advertisements posted around campus and were paid five dollars for their participation, which required approximately 15 minutes.

Design. Task difficulty was manipulated between participants. Those in the harder condition were given two words that prior testing had shown to be relatively difficult to work with (*petroglyph* and *gargantuan*) and were given three minutes to work on each. Those in the easier condition received two easier words (*typewriter* and *overthrown*) and were given five minutes for each. The order of words was not varied: all participants received them in the order shown.

Procedure. At the beginning of the procedure, participants received one page of written instructions including an explanation of the word prospector task, an example, and the scoring rules for the task. These rules were repeated at the top of the page containing the 10-letter word, as well. Participants received points for each letter of each correct word they spelled, and lost points for non-existent, repeated, or misspelled words. For example, if a participant looking at the word "gargantuan" spelled the word "grant," five points would be counted toward the overall score. But, if the participant spelled the non-existent word "naut," four points would be subtracted from the overall score.

After reading the page of instructions, the experimenter repeated the instructions and the rules for scoring. Then, participants were allowed to turn the page and begin creating words from

the first 10-letter word. After working on the first 10-letter word for three or five minutes, participants were stopped and asked to fill out the following page where they estimated the number of points that they expected to receive, the percentile rank into which they would fall in relation to their peers, and the difficulty of the task for themselves and for the average participant, using a scale from 1 (*very easy*) to 10 (*very difficult*). As in Studies 1 and 2, the use of a percentile scale was described in detail. Participants were then given a five-minute, unrelated questionnaire. Next, they were given three or five minutes to repeat the task using a different 10-letter word. Lastly, after the experimenter stopped them, they were given another one-page questionnaire with the same questions as after the first 10-letter word, plus a request for an estimate of their percentile rank for word prospector tasks in general.

Results

Manipulation checks. First, we checked the reliability of the task by comparing the first half with the second half. The split-halves reliability was very high for both the easier and harder versions (Spearman-Brown = .74 and .78, respectively). Next, we checked the difficulty manipulation using MANOVAs with difficulty as a between-participants variable and first vs. second word as repeated measures. Scores were lower in the harder condition than in the easier condition, $F(1, 74) = 95.49$, $p < .001$, $\eta^2 = .56$, and ratings of difficulty were significantly higher, $F(1, 74) = 24.78$, $p < .001$, $\eta^2 = .25$ (see Table 2). There was also an interaction between difficulty and word for score, $F(1, 74) = 15.21$, $p < .001$, $\eta^2 = .17$, and for reported difficulty, $F(1, 74) = 4.98$, $p = .05$, $\eta^2 = .05$, suggesting that the word *petroglyph* was and seemed more difficult than the word *gargantuan*, and *typewriter* was and seemed slightly more difficult than *overthrown*.

Percentile estimates. Next, we looked at percentile estimates using a MANOVA with difficulty level and performance quartile as between-participants variables. Participants were grouped into performance quartiles according to their overall performance across both 10-letter words. The dependent measures were the estimate of overall percentile participants made after having completed both words and their actual overall performance percentile.

There was no significant overall difference between estimated and actual percentiles, $F < 1$, but there was a significant main effect of difficulty, $F(1, 68) = 5.07$, $p = .028$, $\eta^2 = .07$, and an interaction between difficulty and estimated vs. actual percentile $F(1, 68) = 6.88$, $p = .011$, $\eta^2 = .09$. These results reflect the difficulty effect observed in the previous studies: Percentile estimates averaged 54.39 in the easier condition and 43.50 in the harder condition. (Average *actual* percentile was by definition the same in the two conditions).

A main effect of quartile is inevitable, given that quartile was determined by the same performance that determined actual percentiles. However, follow-up tests showed that there was also a positive linear trend of estimated percentiles across quartiles; participants in higher quartiles of performance gave higher estimates of performance than participants in lower quartiles, $p = .011$ (see Figure 6). We also explored this relationship by regressing percentile estimates on actual percentiles. This analysis supports the linear trend above (B = .224, SE = .072, $\beta = .343$, $t(73) = 3.118$, $p = .003$). There was also an interaction between quartile and estimated vs. actual percentile, $F(3, 68) = 42.77$, $p < .001$, $\eta^2 = .65$. Those in the top quartile underestimated their percentile ($M_{\text{easier estimate}} = 67.33$ versus $M_{\text{easier actual}} = 87.00$ and $M_{\text{harder estimate}} = 54.20$ versus $M_{\text{harder actual}} = 87.00$), while those in the bottom quartile overestimated theirs ($M_{\text{easier estimate}} = 52.22$ versus $M_{\text{easier actual}} = 12.00$ and $M_{\text{harder estimate}} = 35.00$ versus $M_{\text{harder actual}} = 11.90$). There was no three-way interaction between quartile, difficulty and estimated vs. actual

measures, *F*s < 1. That is, there is no evidence to contradict the hypothesis that the estimate lines for easier and harder tasks are parallel.

We also performed a MANOVA using participants' estimates of their performance on each of the word prospector words they saw, split by quartile of performance on their overall performance on the two words. Difficulty and quartile were between-participants variables. Actual performance percentile and estimated performance percentile were measured on each word separately, so first vs. second word and actual vs. estimated performance were within-participants variables. There were no significant effects involving first vs. second word, and the pattern of results was the same as in the previous analysis.

Asymmetry by quartiles. To test the unskilled-unaware hypothesis, we recoded errors as (estimated overall percentile – actual overall percentile) for the bottom quartile and (actual overall percentile – estimated overall percentile) for the top quartile. We then performed an ANOVA on these transformed difference scores, with difficulty and quartile as a between-participants variables. Only participants in the top and bottom quartiles of overall performance were included. Means showed no main effect of performance level: $M_{\text{lowest}} = 31.21$ versus $M_{\text{highest}} = 26.58$. There was no main effect of difficulty ($F < 1$), but a significant difficulty by quartile interaction, $F(1, 34) = 8.54$, $p = .006$, $\eta^2 = .20$. In the easier condition, miscalibration was significantly greater in the first ($M_{\text{lowest}} = 40.22$) than in the fourth quartile ($M_{\text{highest}} = 19.67$), $F(1, 34) = 7.492$, $p = .01$. However, in the harder condition, miscalibration was not greater in the fourth ($M_{\text{highest}} = 32.80$) than in the first quartile ($M_{\text{lowest}} = 23.10$), $F(1, 34) = 1.854$, $p = .182$. The degree of miscalibration from the easier to the harder condition within the first quartile was significantly different ($F(1, 34) = 5.472$, $p = .025$) and marginally significant in the fourth quartile ($F(1, 34) = 3.219$, $p = .082$).

The means show that in the easier condition, those in the bottom quartile made larger estimation errors, whereas in the harder condition, those in the upper quartile made larger errors. This is consistent with findings from our previous studies. The same pattern of results was found using the average of the participant's miscalibration errors on each of the two words they saw.

As an alternative measure of estimation errors, we divided participants according to their quartile of performance on one word, and measured the difference between their estimated performance on the other word and their actual performance on the other word. We again calculated error as (estimated percentile – actual percentile) for those in the bottom quartile and (actual percentile – estimated percentile) for the top quartile. The results yielded different patterns depending on which word was conditioned on, but in a predictable way. Results for the second word conditioned on the first are shown in the top half of Table 3. We performed an ANOVA on the transformed differences with difficulty and quartile as between-participants variables. The only significant effect was a difficulty by quartile interaction, $F(3, 67) = 5.08$, $p < .03$, $\eta^2 = .12$, consistent with the pattern we observed before: Top performers were better calibrated when the task was perceived as easier (i.e., average percentile estimates were above 50), and low performers were better calibrated when the task was perceived as harder (i.e., average percentile estimates were below 50). Then, we did the reverse, dividing participants according to their quartile of performance on the second word, and measuring the difference between estimated and actual performance on the first word. Results are shown in bottom half of Table 3. This time, there was no significant interaction with quartile, but this is not surprising because the perceived percentile for both words (typewriter and petroglyph) averaged to 50 across the bottom and top quartiles (suggesting that it was moderately difficult just as in the moderate trivia condition of Study 1).

Note that the overall magnitude of errors is lower when measured on a different task. This reflects the removal of the bias induced by regression to the mean (because, in this analysis, actual percentiles are unbiased and closer to 50). Of course, the total amount of error across all participants on a given task is a constant. However, removing the effects of regression toward the mean makes those at the extremes of performance look much less extreme in their errors of self-perception.

Discussion

It is clear that the word prospector task allows participants to estimate how well they have done compared to others to a moderate degree. However, that ability does not seem to be the province of skilled performers. Rather, the results of the present study support the conclusion we reached on the basis of more difficult-to-estimate tasks in Studies 1 and 2. That is, the skilled and the unskilled are similarly unaware of their relative standing; who makes the larger error is mostly a function of the overall bias in judgments across people. Overall bias varies according to task difficulty, also without any apparent difference in bias between low and high performers. Thus, in easier tasks the unskilled seem unaware of their relative standing, in harder tasks the skilled seem unaware.

General Discussion

The results from all three studies show a consistent picture of the psychology behind relative miscalibration. People have a difficult time judging how their performance compares to the average performance of their peers. Accordingly, estimates of relative standing are rather regressive: The best performers do not guess how well they have done; the poorest performers do not guess how badly they have done. At the same time, as Kruger (1999) found, there is a systematic effect of task difficulty. People give lower estimates of their relative standing when

they find the task more difficult. The well-known above-average effect turns out to be only half the picture. On difficult tasks, the average person thinks he or she is performing below average.

Our studies replicate, eliminate, or reverse the association between task performance and judgment accuracy reported by Kruger and Dunning (1999) as a function of task difficulty. On easy tasks, where there is a positive bias, the best performers are also the most accurate in estimating their standing, but on difficult tasks, where there is a negative bias, the worst performers are the most accurate. This pattern is consistent with a combination of noisy estimates and overall bias, with no need to invoke differences in metacognitive abilities. In this regard, our findings support Krueger and Mueller's (2002) reinterpretation of Kruger and Dunning's (1999) findings. An association between task-related skills and metacognitive insight may indeed exist, and later we offer some suggestions for ways to test for it. However, our analyses indicate that the primary drivers of errors in judging relative standing are general inaccuracy and overall biases tied to task difficulty. Thus, it is important to know more about those sources of error in order to better understand and ameliorate them.

Sources of Inaccuracy

The results of our three studies indicate that there is often a weak positive relation between objective and subjective measures of relative performance. This suggests that people have limited insight into their skills and abilities. We believe that it is important for future research to examine the sources of insight and the sources of error that produce this weak relationship, and the conditions that facilitate or hinder judgment. Two variables we think are worthy of further study in this regard are randomness and feedback.

Randomness. Using a broad definition, randomness can be thought of as any source of variability that is unpredictable for a judge. Thus, people attempting to predict or estimate their

performance relative to others must deal with different kinds of randomness that have different effects on the accuracy of their judgment. We discussed one source of error in connection with Study 3. That is that performance on any given test is subject to random variation, holding ability constant. This kind of randomness stems from luck as to which particular test items are included, transient effects on performance such as distraction or fatigue, etc. Some of the effects of this randomness are irrelevant to judgmental accuracy. For example, as we pointed out in Study 3, classifying participants into quartiles on the basis of performance on a single test will inevitably produce regression toward the mean with regard to any other performance-related measures. Accordingly, we recommend using one sample of performance to segregate low and high performers, and an independent sample of performance to measure perceived and actual relative performance. As Study 3 demonstrated, this reduces the degree to which the extreme quartiles appear biased, although there is still substantial error in relative performance judgments across performance levels.

Other effects of randomness do impact accuracy in important ways, by limiting the predictability of one's relative performance. The less predictable one's performance, the less performers can be expected to guess what their standing will be. In terms of a graph like Figure 1, the less predictability, the flatter the lines relating predicted to actual performance.[4] For example, suppose there is great variability in difficulty from item to item on a test, or that performance is greatly affected by momentary distraction. Then one's performance may vary considerably from one test to another, and will be less predictable from a general notion of one's underlying skill level. It would be useful to learn more about the task and person characteristics that affect predictability, and how different aspects of predictability affect judgments of relative standing.

Feedback. People's ability to estimate their relative performance is largely determined by the kind of feedback they receive, and how they use it. Relevant feedback may be accumulated over a long time to produce a general sense of one's own ability in a domain compared to others. Other feedback is more immediate, indicating how well a particular performance is going. Relative performance judgments will be affected by feedback about one's own performance, as well as feedback about the average performance of others and about the dispersion of performances.

Most of the tasks used by Kruger and Dunning (1999), Krueger and Mueller (2002), and us provided participants with little specific information about how they were doing in an absolute sense, or about how their peers performed. For example, participants were not told whether their quiz answers were correct or not, and were not told what the average score was. Though many tasks in life have this quality, there are also many that do provide considerable performance feedback, such as sports and academics. For instance, a baseball player quickly knows the outcome of each turn at bat and has access to the performance of other players and other teams. Students get direct feedback about their relative standing every time they are graded on a curve.

Ultimately, general theories about accuracy in judging relative performance need to take into account differences in specific feedback conditions. The original "unskilled-unaware" Kruger and Dunning (1999) hypothesis pertained to environments offering impoverished feedback on both absolute and relative performance for both self and others. General claims about accuracy will hinge on discovering more about how, and how well, people use different kinds of feedback about performance.

Sources of Systematic Misestimation

Estimates of relative standing are not only noisy because of random error and poor feedback, they are also prone to systematic bias: People feel they are worse than average on tasks on which everyone performs poorly, and above average on tasks on which everyone performs well (confirming the finding of Kruger, 1999). Kruger interprets this result as implying that judges anchor on their own absolute performance and adjust insufficiently for the knowledge they have of other people. However, little is known as yet about the processes that underlie this phenomenon.

A number of different cognitive processes are known to contribute to people's impressions of their own confidence and ability. Many of these processes can contribute to relative judgments as well. In general, people's feelings of knowing are far from perfect predictors of their actual knowledge (Koriat, 1993; 1995), and people may not fully appreciate that the factors that make a task easy or difficult for them also have a very similar effect on others (Moore & Kim, 2002). Thus, processes that affect perception of absolute performance level may also affect perception of relative ability. For example, people use subjective feelings of cognitive effort as a cue to performance (Schwarz, Strack, Bless, Klumpp, Rittenauer-Schatka, & Simons, 1991). Tasks for which it is easy to produce a response (e.g., a multiple choice recognition test) will lead to upwardly-biased estimates of both absolute and relative performance compared to tasks for which producing a response is difficult (e.g., uncued recall). In another vein, tasks in which alternative choices are clearly specified may be less prone to upward bias. Judgments about the adequacy of a single, focal hypothesis with unspecified alternatives are more prone to overconfidence and confirmation biases (Brenner, 2003; Klayman, 1995; Soll & Klayman, in press).

In general, explanations like these imply over-reliance on one's own experience and discounting of relevant information about others. However, it is also possible that the association between perceived difficulty and estimated relative standing is not a "bias" at all. Instead, it may represent the fact that, in the natural ecology, absolute performance and relative performance are often correlated. If you have poor information about how others perform, it might in fact be the best strategy to guess that you are worse than average when you do poorly and better than average when you do well (Klayman & Burson, 2002).

It is likely that factors other than difficulty also contribute to systematic bias in a given task. For example, people may be unclear about differences among different subpopulations to which they are being compared. College students at top schools, for instance, often experience a shock in moving from an environment in which they were nearly all in the upper percentiles of school performance to one in which they are, on average, only average. Indeed, the above-average effect found in many studies may stem in part from college students' inability to fully adjust for this effect. This systematic bias will help create a seemingly "unskilled-unaware" pattern in many studies involving talented undergraduates. Failure to adjust for the reference group could also produce the opposite effect. If historically poor performers on a task are systematically grouped together and asked to assess their relative performance within that untalented group, it could produce the "skilled-unaware" pattern we observed in our studies. Accordingly, in our studies, we manipulated task difficulty by varying characteristics of the task, and not by selecting more or less-talented subpopulations.

Motivation may also play an important role. For example, self-enhancement undoubtedly contributes to over-estimation, and constraints on self-enhancement may produce under-estimation when performance feedback is unambiguous (Dunning, Meyerowitz, & Holzberg,

1989; Kunda, 1990; Larrick, 1993) or temporally near (Gilovich, Kerr, & Medvec, 1993; Shepperd, Ouellette, & Fernandez, 1996). So, even if errors of judgment are "only" due to general inaccuracy plus systematic bias, there are many factors to explore to understand who misestimates relative performance and when.

Reexamining the Unskilled-Unaware Hypothesis

Kruger and Dunning (1999, 2002) argue that those who are less skilled at a task are also less able to judge their relative skill. This argument is based on two hypotheses: (1) There is a performance-metacognition association, such that those who perform worse at a task are less able to assess their own performances and those of others; (2) Because of this, most of the error in judging relative performance is produced by poor performers. Kruger and Dunning provide evidence for the performance-metacognition association (1999, Studies 3 and 4). Although our procedures were designed with different goals in mind, we can test whether a performance-metacognition relationship exists in our studies by measuring the correlations between estimated and actual percentiles. We separated participants into two groups, above and below median performance. We then looked at the correlation between estimated and actual performance percentile within each group. Naturally, these correlations will be smaller than for the population as a whole, because we are looking only within each half of the range of performance. However, the comparison between better and worse performers can be informative. Top-half and bottom-half performers did not differ significantly on any single task, but taking all 12 tasks together, we do see some indication that top-half performers were better at estimating their relative standing. We transformed these 24 correlations using Fisher's *r*-to-*z* and ran a paired samples *t* test on the *z*s, comparing top-half and bottom-half performers' correlations across the 12 tasks. The average correlation between estimated and actual percentile was .24 for top-half performers and .03 for

bottom-half performers, $t(11) = 2.13$, $p = .06$, suggesting that the top-half performers had better insight into their relative standing,. These results suggest that in our studies, as in Kruger and Dunning's (1999), those who perform better at tasks may also be more accurate in evaluating their performance. Yet, in our studies, poor performers are not disproportionately responsible for error in judgments of relative performance. This illustrates our basic argument: We do not say that there is no relation between cognitive skill and metacognitive skill, but rather, that such a relationship is not a primary determinant of who makes what errors in judging relative performance. Better performers might be somewhat more sensitive to differences in their achievements, but there is still a significant degree of noise and bias in translating that sensitivity into judgments of relative standing. Thus, when feedback is ambiguous and the overall task bias is negative, the judgments of better performers deviate more from the truth. Ultimately, who deviates more from the truth is a more a function of task-induced bias than of metacognitive advantage.

Conclusions

It is a well-established and entertaining factoid that, on average, people think they are above average (e.g., Svenson, 1981). However, recent research tells a more interesting story about who is wrong and when. Kruger and Dunning (1999, 2002) suggest that there is a relationship between task performance, metacognition, and judgmental accuracy. They conclude that most of the error in judging relative performance comes from poor performers' tendency to overestimate their abilities, which is in turn due to their poorer metacognitive skills. Some secondary analyses of our data do provide some evidence that task skill is correlated with calibration, and Kruger and Dunning (1999, Study 4) present a regression analysis showing that deficits in metacognitive skill predict absolute miscalibration for participants in the bottom

quartile. However, evidence for the mediating role of metacognition remains controversial (Kruger & Dunning, 2002; Krueger & Mueller, 2002), and it remains unclear how important this may be in explaining inaccuracies in relative judgment. Evidence from three new studies indicate that the answer to "Who makes errors?" is, more or less, "Everyone." On the kinds of tasks that have been used to date, the skilled and the unskilled are similarly limited in judging how their performance compares to others'. The answer to "When?" is, in part, a matter of perceived task difficulty. When the task seems hard, top performers underestimate their standing. When the task seems easy, those near the bottom overestimate theirs.

Judgments of relative ability play an important role in decisions about engaging in competitive activities, purchasing goods and services, and undertaking challenging tasks (Burson, 2003; Simonsohn, 2003). Overestimates of relative ability can lead to frustration, loss, and even physical harm. (Think, for example, of those middling horseback riders or skiers who attempt advanced trails.) On the other hand, there are also significant domains in life where relative ability may be *under*estimated, so that people fail to participate when they would have succeeded (Moore, 2003). The research presented in this paper provides a foundation for further exploration of how and how well people know where they are on the curve, and how we can help people to do that better.

References

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

Brenner, L. A. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes, 90*, 87-110.

Burson, K. A. (2003, October). *The effect of interpersonal and interproduct comparison on product choice.* Paper presented at the annual conference of the Association for Consumer Research, Toronto, ON, Canada.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082-1090.

Gilovich, T., Kerr, M., & Medvec, V. H. (1993). Effect of temporal perspective on subjective confidence. *Journal of Personality and Social Psychology, 64*, 552-560.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*, 1161-1166

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5*, 55-71.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*, 226-246.

Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Psychology of learning and motivation: Vol. 32. Decision making from a cognitive perspective* (pp.365–418). New York: Academic Press.

Klayman, J., & Burson, K. A. (2002, November). *Looking for Lake Wobegon: Why sometimes we're all below average.* Paper presented at the annual conference of the Society for Judgment and Decision Making, Kansas City, MO.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79*, 216-247.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609-639.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*, 311-333.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*, 180-188.

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*, 221-232

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121-1134.

Kruger, J., & Dunning, D. (2002). Unskilled and unaware—but why? A reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology, 82*, 189-192.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498.

Larrick, R. P. (1993). Motivational factors in decision theories: The role of self-protection. *Psychological Bulletin, 113*, 440-450.

Moore, D. A. (2003). *Egocentric biases and the failure of strategic prediction*. Unpublished manuscript. Carnegie Mellon University.

Moore, D. A., & Kim, T. G. (2002). *Myopic social prediction and the solo comparison paradox*. Unpublished manuscript. Carnegie Mellon University.

Schwarz, N., Strack, F., Bless, H., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*, 195-202.

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Journal of Personality and Social Psychology, 70*, 844-855.

Simonsohn, U. (2003, November). *Weather to go to college.* Paper presented at the annual conference of the Society for Judgment and Decision Making, Vancouver, BC, Canada.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior & Human Decision Processes, 65*, 117-137.

Soll, J. B., & Klayman, J. (in press). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica, 47*, 143-148.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10*, 243-268.

Appendix

We used simulations to verify the pattern of results that we would obtain if tasks differed in overall bias (e.g., as a function of task difficulty), but the ability to judge one's relative position did not vary with one's relative ability. The results presented in Figure 2 were produced using these simulations.

We produced a Monte Carlo simulation of 1000 participants having a range of different abilities at a task. The basic assumptions of our model are that the participant's performance score and his or her predicted performance are both imperfect estimates of underlying ability.

Observed Performance

We assume that participant *j*'s observed score, S_j, is determined by the *j*'s level of ability, A_j, plus some random error, e_j. The random error represents all the elements that make any single test of performance less than 100% reliable and valid in representing underlying ability.

(1) $S_j = A_j + e_j$

We used a standard normal distribution for A_j, representing the participants' abilities relative to others in the population. The error term, e_j, is drawn randomly from a normal distribution with mean of zero. The variance of the error distribution can be manipulated to represent the quality of the test; higher error variance represents lower reliability. Like ability (A_j), the observed score (S_j), is a relative measure, having a mean of zero. However, because of the addition of error, it has a higher variance than *Aj*.

Estimated Performance

Each participant estimates his or her performance based on his or her ability plus some error, and possibly some overall bias.

(2) $\hat{S}_j = A_j + z_j + b_t$

The error, z_j, is drawn from a mean-zero, normal distribution whose variance represents the noisiness of participants' estimates of their relative ability. The overall bias, b_t, is a function of the task. We do not distinguish here between misestimation of one's own absolute performance and misestimation of others'. Both, together cause inaccuracy in the participant's estimate of where he or she stands in the distribution of performance, and are thus included in ($z_j + b_t$).

Presentation

Figure 2 shows the results of a simulation based on Equation 2, using the format of previous reports. Results are averaged across participants in each of the four quartiles of observed performance score. The *x*-axis shows the mean percentile of scores in the distribution of all 1000 scores, by quartile of score . The *y*-axis shows the percentile of the mean predicted score, according to where each predicted score would have fallen in the actual observed distribution of 1000 scores.

The particular example shown in Figure 2 represents the following situation. The performance test has high validity, with a correlation of approximately .80 between ability and observed performance score. Participants find prediction to be difficult, but not impossible, with a correlation of about .35 between predicted and actual score. The three lines represent the results with no added bias ($b_t = 0$) and with biases of ±25 percentiles relative to the no-bias condition. Different parameter values produce lines with different slopes, but they are still nearly parallel except in extreme cases, where floor and ceiling effects produce some curvature.

Models of Unskilled–Unaware

The main goal of our simulations was to test the general pattern of results that would be expected in the absence of any relation between performance and ability to judge one's performance. The simulations confirm that the basic findings of Kruger and Dunning (1999) and

our subsequent findings using tasks of varying difficulty are consistent with that hypothesis. However, we were also curious to see whether Kruger and Dunning's unskilled-unaware hypothesis would show a distinctly different pattern.

There are numerous possible specific interpretations of the general hypothesis that those who are least able are also least able to predict their own relative performance. One simple and plausible instantiation of this hypothesis is:

(3) $\hat{S}_j = A_j + (z_j + b_t)f(A_j)$

where *f* is a function such that the amount of error and bias decrease as ability increases. We modeled this using a simple linear function such that there is no bias or error for those in the top percentile of ability, an average amount for those of median ability, and double the average error and bias for those in the bottom percentile of ability. We believe this model captures the tenor of Kruger and Dunning's (1999; 2002) unskilled-unaware hypothesis and Kruger's (1999) interpretation of task difficulty effects. It is also consistent with evidence that anchoring effects are stronger for judgments that are more ambiguous (Jacowitz & Kahneman, 1995). Thus, if poorer performers find comparative judgments to be more difficult to make, they may be more prone to anchoring on perceived absolute difficulty.

Figure 3 is based on Equation 3, using the same standards as before for the validity of the performance measure and participants' overall ability to predict relative performance. The results show a distinct pattern, which seems to be different from the one we observe in the present studies. However, we consider these findings to be only exploratory. With different parameter values, the differences between the results of Equation 2 and Equation 3 can be less clearly distinct, and other models of unskilled-and-unaware are also plausible.

Author Note

Katherine A. Burson, University of Michigan; Richard P. Larrick, Duke University; Joshua Klayman, University of Chicago.

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

Footnotes

[1] A similar explanation has been offered for miscalibration in confidence judgments in which those who are most confident in their answers are also those who are most overconfident (e.g. Juslin, 1993; Juslin, 1994; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Soll, 1996; Wallsten, Budescu, Erev, & Diederich, 1997).

[2] There are many strategies for assigning percentiles in the presence of ties. We tried several methods with no meaningful differences in results.

[3] For a more detailed explanation of how this method removes the biasing effects of regression to the mean, see Klayman et al. (1999).

[4] Kruger and Dunning (1999; 2002) emphasize the role of task reliability in producing these regression effects, but it is more precisely predictability that matters. Reliability is often associated with predictability, but it is neither necessary nor sufficient. Imagine the task of tossing coins into a box while blindfolded. After flipping 10 coins, the tosser is asked to estimate how the proportion of heads-up coins in the box will compare to the average coin-tosser's. Those with the most heads and those with the fewest should of course give very similar, arbitrary guesses, and both will appear quite inaccurate. That task is both unreliable and unpredictable. However, if we repeat the task with the blindfold removed, the tossers can easily count the number of heads. The task is still unreliable, but relative standing is now very predictable. Similarly, a first-year college student may face final exams in five required courses. As tests of academic performance, reliability may be poor—the student's position in Physics may be poorly correlated with his or her position in English, etc. Nevertheless, by the end of the semester, the student may have a good idea about where he or she is likely to fall on each of the tests. Now put the blindfold back on, but give different, biased coins to each tosser. Multiple rounds with the

same coins might now be quite reliable with regard to relative standing, but will be unpredictable for the tossers. Similarly, one may have no idea of one's relative performance on, say, emergency driving maneuvers, no matter how reliably they can be tested.

Table 1

Expected, Actual, and Difference Between Percentiles for Each Trivia Quiz in Study 2, by Quartile of Performance on That Quiz

	Quartile				
Domain and measure	Overall	Lowest		Highest	
Easier college acceptance rates					
Expected percentile	51.23	41.89	(26.21)	44.83	(30.28)
Actual percentile		10.89	(5.61)	91.17	(4.65)
Difference		31.00	$t(8) = 3.62$**	46.33	$t(5) = 3.63$*
Harder college acceptance rates					
Expected percentile	41.08	49.22	(29.28)	37.15	(25.77)
Actual percentile		10.89	(5.02)	83.19	(8.44)
Difference		38.33	$t(8) = 3.61$**	46.04	$t(12) = 6.04$**
Easier pop songs on charts					
Expected percentile	40.83	30.85	(23.59)	60.13	(27.95)
Actual percentile		15.85	(8.45)	83.88	(8.10)
Difference		15.00	$t(12) = 2.23$*	23.75	$t(7) = 2.90$*
Harder pop songs on charts					
Expected percentile	35.73	28.00	(24.07)	46.67	(24.21)
Actual percentile		12.50	(7.15)	82.83	(7.93)
Difference		15.50	$t(9) = 2.19$	36.17	$t(11) = 4.93$**

Easier year of Nobel Prize					
Expected percentile	32.05	19.78	(20.71)	35.62	(27.63)
Actual percentile		11.26	(6.59)	80.67	(6.31)
Difference		8.52	$t(8) = 1.07$	45.06	$t(12) = 5.59$**
Harder year of Nobel Prize					
Expected percentile	21.53	24.00	(15.17)	35.33	(18.62)
Actual percentile		12.00	(3.75)	92.00	(3.87)
Difference		12.00	$t(8) = 2.59$*	56.67	$t(5) = 6.63$**

Note. Numbers shown in parentheses are standard deviations. Difference = (expected percentile – actual percentile) for the lowest quartile and (actual percentile – expected percentile) for the highest quartile. *t*-test is a paired *t*-test on actual versus expected percentile.

*$p < .05$. **$p < .01$.

Table 2

Performance Scores and Ratings of Difficulty for Oneself on Each Word Prospector Problem in Study 3 with Standard Deviations in Parentheses.

Domain and word	Score		Difficulty rating	
Easier word prospector				
Typewriter	57.31	(22.25)	5.92	(1.70)
Overthrown	64.64	(22.27)	5.67	(1.85)
Overall easier	60.97	(22.41)	5.79	(1.77)
Harder word prospector				
Petroglyph	25.73	(19.50)	7.20	(1.67)
Gargantuan	17.15	(14.36)	7.68	(1.33)
Overall harder	21.44	(17.45)	7.44	(1.52)
Overall Mean (all words)	45.32	(28.96)	6.46	(1.85)

Table 3

Mean Difference Between Estimated and Actual Percentile for each Word Prospector Problem in Study 3, by Quartile of Performance on the Other Problem

	Quartile on other word			
Domain by word	Lowest		Highest	
Easier word prospector: Overthrown[a]	28.59	(30.95)	6.38	(22.09)
Harder word prospector: Gargantuan[a]	11.98	(26.19)	26.10	(20.93)
Easier word prospector: Typewriter[b]	9.57	(43.01)	8.10	(27.63)
Harder word prospector: Petroglyph[b]	14.33	(26.35)	15.80	(23.06)

Note. Numbers in parentheses are standard deviations. Difference = (expected percentile – actual percentile) for the lowest quartile and (actual percentile – expected percentile) for the highest quartile.

[a]Quartile determined by the first word presented, estimated and actual performance percentile on the second.

[b]Quartile determined by the second word presented, estimated and actual performance percentile on the first.

Figure Captions

Figure 1. Participants' estimates of the percentiles of their performances relative to their peers, by quartile of actual performance in four experiments from Kruger and Dunning (1999). This pattern of results suggests that unskilled participants are more miscalibrated than skilled participants are.

Figure 2. Hypothetical estimates of percentile of performance by actual quartile of performance on tasks of varying difficulty, assuming everyone is equally unaware of their ability and equally prone to the overall biasing effects of task difficulty.

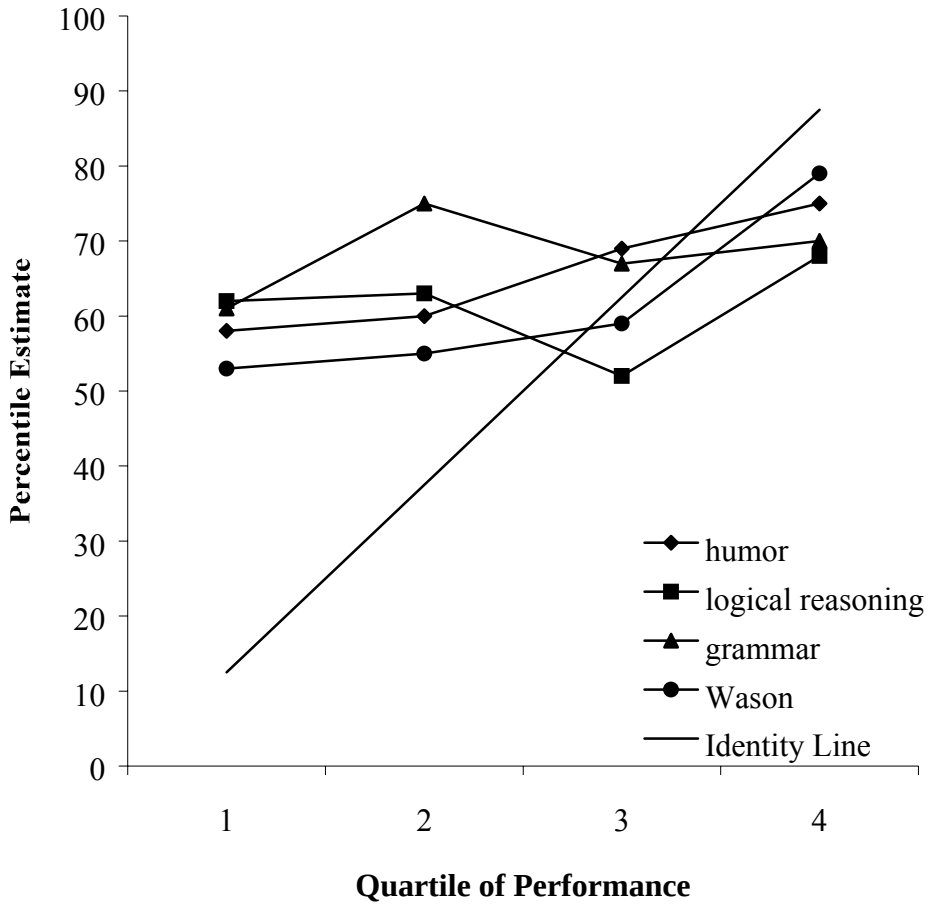
Figure 3. Hypothetical estimates of performance by actual quartile of performance on tasks of varying difficulty, assuming that less skilled participants are simply more error-prone in estimating their relative performance. Less skilled participants' estimates will regress more, and the mean to which they regress will be a function of task difficulty.

Figure 4. Participants' estimates of performance percentile by quartile of actual performance on easier and harder tests of University of Chicago trivia in Study 1.

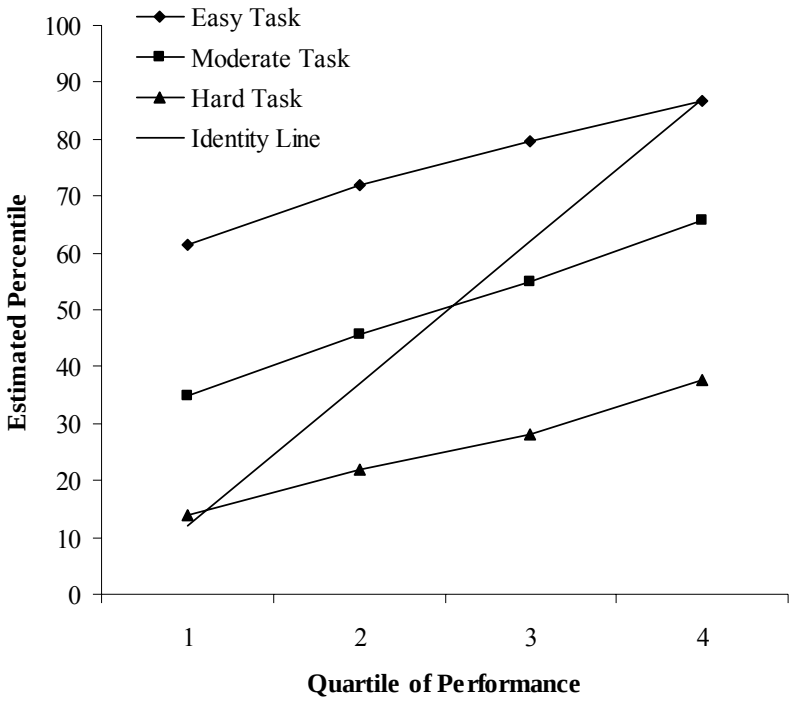
Figure 5. Participants' estimates of performance percentile by quartile of actual performance on six sets of estimates of varying difficulty in Study 2.

Figure 6. Participants' overall estimates of performance percentile by quartile of overall actual performance on an easier and harder word prospector task in Study 3.

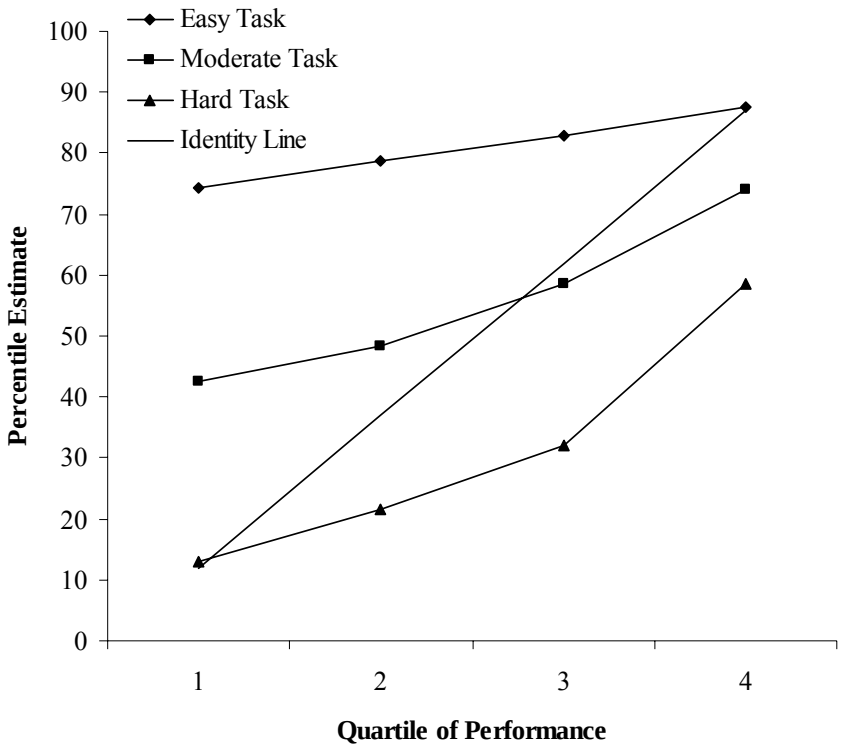
1


100
90
80
70
60
50
40
30
20
10
0
Percentile Estimate
1
2
3
4
Quartile of Performance
humor
logical reasoning
grammar
Wason
Identity Line

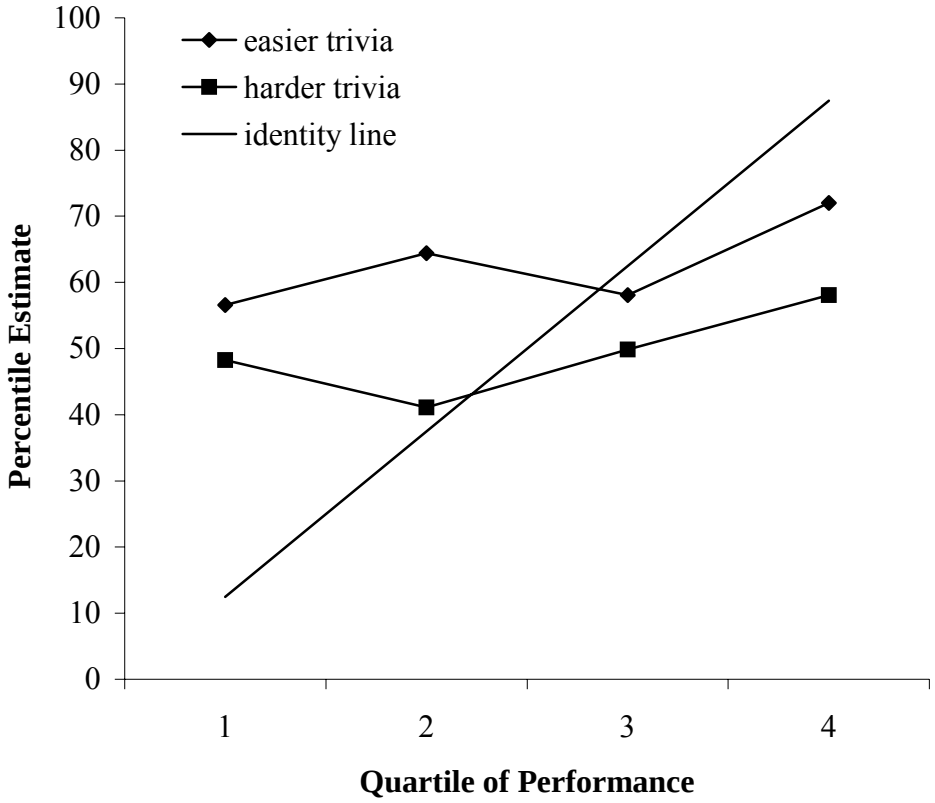
2


Easy Task
Moderate Task
Hard Task
Identity Line
100
90
80
70
60
50
40
30
20
10
0
Estimated Percentile
1
2
3
4
Quartile of Performance

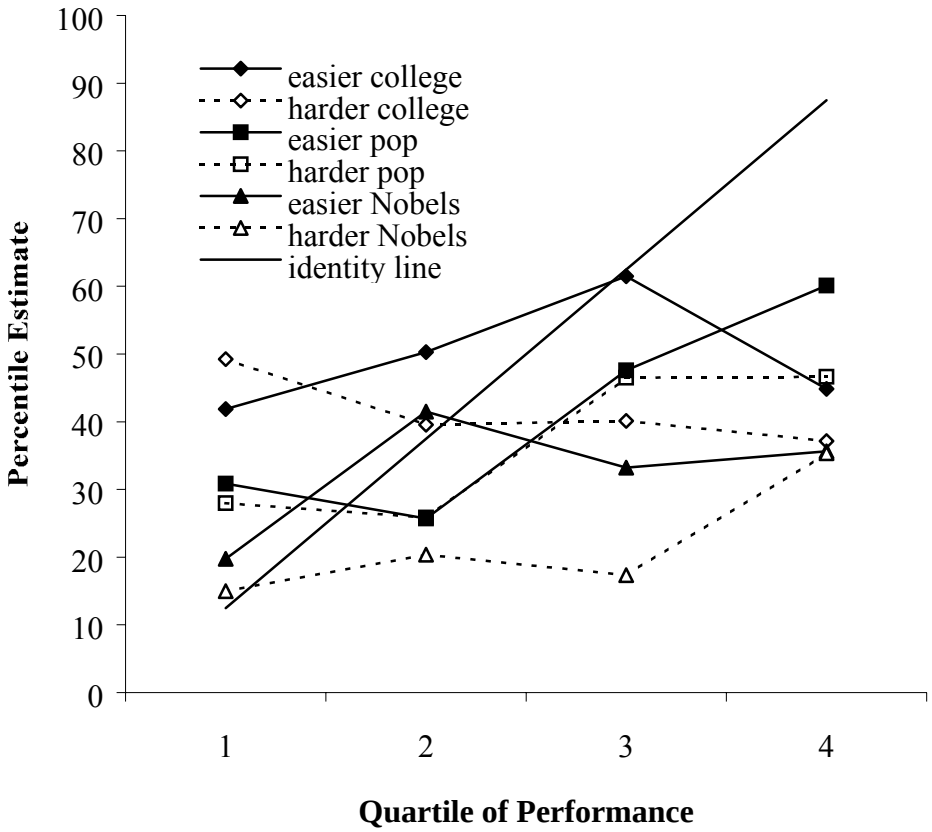
3


Easy Task
Moderate Task
Hard Task
Identity Line
100
90
80
70
60
50
40
30
20
10
0
Percentile Estimate
1
2
3
4
Quartile of Performance

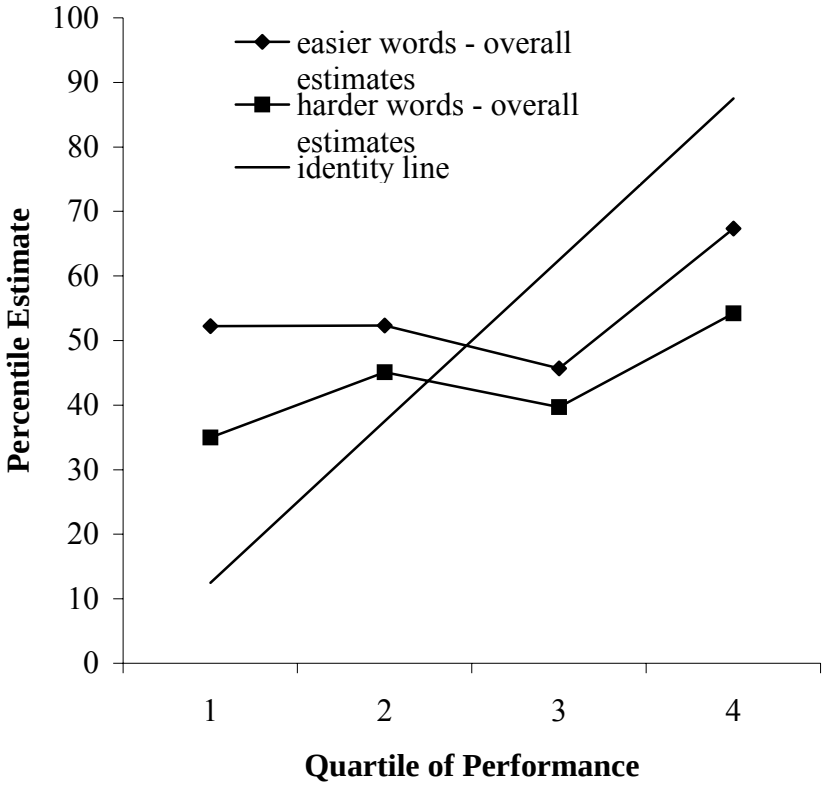
4


100
90
80
70
60
50
40
30
20
10
0
easier trivia
harder trivia
identity line
Percentile Estimate
1
2
3
4
Quartile of Performance

5


100
90
80
70
60
50
40
30
20
10
0
Percentile Estimate
easier college
harder college
easier pop
harder pop
easier Nobels
harder Nobels
identity line
1
2
3
4
Quartile of Performance

6


100
90
80
70
60
50
40
30
20
10
0
Percentile Estimate
easier words - overall estimates
harder words - overall estimates
identity line
1
2
3
4
Quartile of Performance