Running Head: DIFFICULTY AND MISCALIBRATION

Skilled or Unskilled, but Still Unaware of It:

How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons

Katherine A. Burson

University of Chicago

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 this inaccuracy comes from unskilled performers, who also lack the metacognitive skill to evaluate their performance. Thus, the unskilled overestimate their standing. Krueger and Mueller (2002) contend that this pattern reflects only statistical regression plus an overall upward bias. We present three studies that test these explanations by examining what happens in the absence of upward bias. Moderately difficult tasks produce little overall bias and little difference in accuracy between best and worst performers. Very difficult tasks produce negative bias, and the *unskilled* are better calibrated. These patterns are consistent with regression-plus-bias, although differences in metacognitive ability may still play some role.

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. Kruger and Dunning (1999; 2002) argue that overestimates of relative performance exist primarily because people who are unskilled at a task also lack the metacognition to realize that they are unskilled. On the other hand, people who are more skilled have the true ability to justify believing that they are in the upper percentiles of performance, plus they have a more accurate perception of their standing. This means that unskilled participants are more miscalibrated than skilled participants are, as shown in Figure 1. We will 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 experienced. However, Krueger and Mueller (2002) argue that this phenomenon does not require a metacognition explanation. They propose that poor performers are no more error-prone than good performers—they just appear that way because of two effects: 1) subjective and objective measures of performance are imperfectly correlated, so estimates of relative performance regress toward the mean and 2) overall, people tend to give inflated estimates of relative ability.[1]

In other words, regardless of skill level, people do not have much knowledge about how they compare to others; their guesses contain a lot of random error. Thus, the average guess of top performers does not differ very much from the average guess of the poorest performers. In Kruger and Dunning's (1999) tasks, most participants estimated their performance as better than average, so naturally those who actually *were* above average were closer to the truth. Kruger and Dunning (2002) disagree, and present data showing that the less skilled are still less aware after controlling for the effects of regression and overall bias. Kruger and Dunning (2002) and Krueger and Mueller (2002) focus their discussion on the question of whether metacognitive skills can be shown to mediate the asymmetry. Again, they disagree. In the end, the evidence provided by Kruger and Dunning (1999; 2002) and Krueger and Mueller (2002) remains equivocal.

In the present paper, we approach the argument using a different methodology. We ask what pattern would be observed if people did not, on average, see themselves as above average. In fact, people do not always see themselves as better than average. Our experiments take advantage of research by Kruger (1999) showing that there are contexts in which people grossly understate their abilities. 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. We can use this effect to determine the extent to which it is *everyone* who is unaware of relative position or poor performers in particular.

What would one expect to see when there is a downward bias in estimating one's percentile, as is the case with hard tasks? The answer is clear if everyone is equally unaware, that is, if people at all skill levels are equally poor at estimating their relative performance and equally prone to underestimate their relative standing on hard tasks and overestimate it on easy tasks. In that case, percentile estimates will be similar across skill levels, resulting in relatively flat, parallel lines like those in Figure 2.

It is less clear what would be observed if those who are less skilled are also less accurate in judging their relative standing. Kruger and Dunning's (1999; 2002) hypothesis is not specific enough to provide precise predictions for what would happen if participants' mean percentile estimates were below average. One plausible interpretation is that less skilled participants are simply more error-prone in estimating their relative performance. This will tend to push their estimates closer to the mean because noisier estimates are naturally more regressive. The task difficulty effect can be thought of as changing the mean to which estimates regress. On an easy task, people who are just guessing tend to guess that their performance is above average. On a hard task, guessers guess that they are below average. The simulated results of such a model are shown in Figure 3. There are a number of other possible models for the unskilled-unaware

hypothesis, but in general, we should expect to see some variation from the pattern of relatively flat, parallel lines of Figure 2.

In the remainder of this paper, we describe three experiments that manipulate the difficulty of tasks and hence participants' beliefs about their relative standing. By looking at participants' estimates by their actual skill level, we are able to see which of the two patterns described above is a better description: the unskilled are unaware or everyone is equally unaware and biased by difficulty.

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 easy quiz about University of Chicago trivia and 43 students took a moderately difficult quiz about University of Chicago trivia. Care was taken to ensure that the moderate task was not below some "minimal threshold of knowledge, theory, or experience" as cautioned by Kruger and Dunning (1999, p. 1132); the moderate trivia were answered well above chance level (as, of course, were the easy 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 easy or moderate). 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 moderate trivia resulted in a lower actual score than the easy 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 moderate trivia were also rated as significantly more difficult than the easy 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 easy trivia and in the 48th percentile for the moderate 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.

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, $F(3, 81) = 2.68$, $p = .05$, $?^2 = .09$, 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.

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 (see Table 1). 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 unskilled-unaware effect largely disappeared. This result is consistent with the hypothesis that accuracy in estimating one's percentile is difficult regardless of skill level, in accord with Krueger and Mueller's (2002) assertion.

Study 2

The next experiment extends this logic by using tasks that were perceived to be more difficult than those used in Study 1. If unawareness is universal, then it will be the *unskilled*

participants who will appear to be more aware of their relative standing in *difficult* tasks. This is illustrated by the lowest line of Figure 2.

As in Study 1, we manipulated perceived difficulty, but this time we compared moderate and difficult conditions. We used two manipulations to create the desired range of perceived difficulty. First, we sampled several trivia domains that we expected to vary in perceived difficulty. Second, we manipulated the criterion for accuracy in the tasks from less to more exacting. 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 difficult or an easy 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 moderate 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 difficult 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 difficult version.

Two domains 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 (financial worth and hockey) 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 difficult 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 (difficult first or easier first) was a between-participant variable. The difficulty manipulation worked; the hard conditions were perceived as significantly more difficult than the easier versions (M = 6.59 versus M = 7.94), $F(1, 35) = 30.43$, $p < .001$, $?^2 = .47$. Hard and easier conditions also differed significantly in actual performance (M = 19.84% correct versus M = 68.77% correct), $F(1, 35) = 808.15$, $p < .001$, $?^2 = .96$).

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 $M_{\text{Nobel Prize}} = 8.19$), $F(2, 70) = 15.16$, $p < .001$, $?^2 = .30$. 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, $F(2, 70) = 14.25$, $p < .001$, $?^2 = .29$. Also, percentile estimates were lower in

the difficult (narrow range) versions than in the easier versions, $F(1, 35) = 22.57$, $p < .001$, $?^2 = .39$ (see Table 2). 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, $F(2, 70) = 7.18$, $p < .001$, $?^2 = .17$, the implications of which are unclear.

Asymmetry by quartiles. Next, for each of the twenty subsets of estimates (five domains x two subsets x two difficulty versions), we divided the participants into four quartiles of performance relative to the performance of other participants on the same subset of questions. The quartile that a particular participant fell in on the difficult subset of a given subdomain was almost completely uncorrelated with the quartile that a participant fell in on any other subdomains (*r*s from -.39 to .31, median = .02). 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 2).

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 (see Table 3). In both subsets of college acceptance rates, the estimated performance was near 50 across quartiles. In those subsets, the mean estimation error was of approximately the same magnitude in the lowest and highest quartiles, 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. 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-unaware pattern failed to replicate. The results of this domain fell between colleges and Nobels, with a nonsignificant trend toward higher estimation error in the top quartile.

The overall pattern in Table 3 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 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. Instead, the relative degree of miscalibration between low and high performers is driven by the task difficulty: Domains that feel harder (Nobel prizes) and criteria that feel harder (a narrow range of acceptable answers) make low performers appear better-calibrated than high performers. However, just as the apparent unskilled-unaware effect is largely a function of perceived task difficulty, so is the opposite, unskilled-aware effect.

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 easy pop music estimates to .52 on one set of hard 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 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 compared participants' ratings of how difficult they found the task and their performance scores using MANOVAs with difficulty as a between-participants variable and first vs. second word as repeated measures. Scores were lower in the difficult condition than in the easy condition, $F(1, 74) = 95.49$, $p < .001$, $?^2 = .56$, and ratings of difficulty were significantly higher, $F(1, 74) = 24.78$, $p < .001$, $?^2 = .25$ (see Table 4). There was also an interaction between difficulty and word for score, $F(1, 74) = 15.21$, $p < .001$, $?^2 = .17$, and for reported difficulty, $F(1, 74) = 4.98$, $p = .05$, $?^2 = .05$, suggesting that the word *petroglyph* was and seemed more difficult than the word *gargantuan*, and *typewriter* was and seemed slightly more difficulty 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, $F(1, 68) = 5.07$, $p = .028$, $?^2 = .07$, and an interaction between difficulty and estimated vs. actual percentile $F(1, 68) = 6.88$, $p = .011$, $?^2 = .09$. 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). There was also an interaction between quartile and estimated vs. actual percentile, $F(3, 68) = 42.77$, $p < .001$, $?^2 = .65$. As shown in Table 5, those in the bottom quartile underestimated their percentile, while those in the upper quartile overestimated theirs. There was no three-way interaction between quartile, difficulty and estimated vs. actual measures, Fs < 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 are shown in Table 5. There was no main effect of difficulty ($F < 1$), but a significant difficulty by quartile interaction, $F(1, 34) = 8.54$, $p = .006$, $?^2 = .20$. 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 6. 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, $?^2$ = .12, consistent with the pattern we observed before: Top performers were better calibrated when the task was perceived as easy (i.e., average percentile estimates were above 50), and low performers were better calibrated when the task was perceived as hard (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 6. 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 (as in the moderate trivia condition of Study 1).

Note that the overall magnitude of errors is lower when measured on a different task (compare Tables 5 and 6). 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 easy tasks the unskilled seem unaware of their relative standing, in hard tasks the skilled seem unaware.

General Discussion

The results from all three studies show a consistent picture. 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) also 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.

We looked at tasks that varied in how difficult they were to perform and in how difficult it was to estimate one's relative standing. Across the board, our results are consistent with this 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. Our studies replicate, eliminate, or reverse the unskilled-unaware effect as predicted by a combination of regressive estimates and overall bias. Where there is a positive bias, the best performers are also the most accurate in estimating their standing, but where there is a negative bias, the worst performers are the most accurate.

Detecting Skilled-Unskilled Differences in Percentile Estimation

We do not mean to imply that there is no relation between task ability and metacognitive ability—only that the technique of plotting perceived percentile against actual percentile does not

demonstrate such a link. However, evidence for such a relationship might be found elsewhere. For example, Kruger and Dunning (1999) present a regression analysis showing that deficits in metacognitive skill predict absolute miscalibration for participants in the bottom quartile (but see also the subsequent debate over the mediating role of metacognition: Kruger & Dunning, 2002; Krueger & Mueller, 2002). Here we present three other techniques for overcoming the inherent limitations of plotting subjective estimates against objective measures of performance percentile.

Using independent samples. Perhaps the best technique for representing the true degree of miscalibration is the one introduced in Study 3, which is to use a measure from one subtest to create the groupings of low and high performers, and then to plot the perceived versus actual percentile differences from a second, independent sample of performance from the same domain. Both the sorting of performers by one subtest and the measurement of estimation error on the other are of course prone to error. However, this technique ensures that the errors are independent and unbiased. Otherwise, a given instance of bad luck simultaneously pushes the participant toward the bottom quartile and produces a performance that is worse than it seemed. As Study 3 demonstrated, this reduces the degree to which the extreme quartiles appear biased. Across a number of tasks, it could also provide evidence that one quartile is more systematically biased than the other is. No such evidence, however, was found in Study 3. It should be noted, though, that this technique does not remove the biasing effect of task difficulty. Thus, it would also be important to adjust for task bias (as in the following example) or to use a set of tasks in which perceived percentiles are centered on 50.

Adjusting estimations for overall bias. To remove the biasing effect of task difficulty, one can measure the accuracy of percentile estimates after subtracting out the overall task-bias (as measured in the overall sample). That is, for a task in which people believe they are in the 60th

percentile on average, 10 points would be subtracted from each person's percentile estimate; if they believe overall that they are in the 40th percentile, then 10 points would be added. One could then look at the mean magnitude of error at different levels of performance after adjusting for overall bias. Note that this technique does not remove the selection bias due to sorting and measuring error on the same sample (discussed in the previous paragraph), which tends to exaggerate the apparent miscalibration in both the first and fourth quartile (roughly equally). It does remove the biasing effect of task difficulty, which is the critical variable that tends to make one quartile appear more able than the other.

As an example of this test, we performed an ANOVA on the estimation error measure adjusted for overall percentile bias, using all domains of all three studies. We treated domain (n = 12) and quartile (lowest or highest) as between-participant variables. This particular implementation is extremely liberal, since counting multiple estimates from one participant as independent observations overstates the appropriate degrees of freedom. It showed a small, non-significant trend for fourth quartile participants (M = 27.07) to be more accurate than first quartile participants (M = 30.98), $F(1, 210) = 1.86$, $p = .175$, $?^2 = .01$. There was a marginal effect of domain on adjusted estimation error, $F(11, 210) = 1.76$, $p = .064$, $?^2 = .08$, and no interaction between quartile and domain. Even with this lenient test of the effect of quartile on error, there was little indication that fourth quartile participants were more accurate than were first quartile participants.

Measuring correlations between estimated and actual percentiles. Another approach is to test whether the estimates of more skilled participants are better correlated with actual performance. To illustrate this test, 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. Looking across all 12 tasks, we do see some indication that top-half performers were better at estimating their relative standing. Top-half and bottom-half performers did not differ significantly on any single task, but taking all tasks together, a difference does appear. We transformed these twelve 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. This analysis is the best support we found in our three studies for Kruger and Dunning's (1999) unskilled-unaware hypothesis.

Our procedures were designed with different goals in mind, so the data reported in the previous two analyses are only suggestive evidence about the existence of a performance-metacognition relationship. However, they serve to illustrate the kinds of approaches that researchers might want to pursue in order to discover whether such a relationship exists once the effects of regression and difficulty-related bias have been systematically controlled.

Error in Estimating Percentile Ratings

The results of these studies indicate that there is often a weak positive relation between objective and subjective measures of relative performance, producing substantial regression to the mean effects. This suggests that people have limited insight into their skills and abilities. We believe that is it important for future research to examine the various sources of error that produce this weak relationship and the conditions that exacerbate or ameliorate them. With that

in mind, we briefly review some of the sources of error that we think could be distinguished and studied systematically.

Kruger and Dunning (1999; 2002) emphasize the role of task reliability in producing regression to the mean, but we want to emphasize that it is unpredictability that is the critical determinant of these regression effects. Reliability may often be associated with predictability, but it is neither necessary nor sufficient. Consider one extreme case in which both reliability and predictability are very low, because random, undetectable luck drives performance. Imagine the task of tossing coins into a box, out of sight. 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. Without being able to see the results, those with the most heads and those with the fewest will of course give very similar estimates (50 percent heads), and both will appear quite inaccurate.

But unreliability does not imply poor performance at estimating percentiles. Imagine the throw-the-coins-in-the-box task, only this time the tosser can see the coins. The tosser will now be very accurate at guessing how his or her proportion of heads compares to the rest of the population, but the task is just as unreliable as before. 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. Conversely, a reliable test may nevertheless be unpredictable. Suppose, for example, that the blindfolded coin-tossers were each given differently-biased coins, and were tested repeatedly with the same coin. Relative standing would now be much more reliable, but no more predictable. Similarly, one may have no idea of one's

relative performance on, say, emergency driving maneuvers, no matter how reliably they can be tested.

Thus, unreliability does not necessarily imply unpredictability, nor vice versa. The two may be often be associated, because a task with a large element of luck—i.e., unobservable random variation in performance—will have both. However, this is only one potential source of unpredictability. Other sources are:

Feedback on one's own performance. It may be very difficult to know the extent to which one has succeeded or failed relative to a given absolute criterion. This is presumably the critical source of error in the metacognition explanation for expertise differences: The more able have a better ability to judge their absolute performance

Feedback on other's performance. People who have very clear knowledge of how well they performed may nevertheless have very little information about how well others are likely to have done.

Translation mistakes. Finally, people may make errors in translating their sense of how their performance compares to others into a percentile (by misgauging the population dispersion, for example).

What is clear from this taxonomy is that situations will vary in the degree to which each source is influential. It suggests that as more sources of error are minimized, people should become more accurate in their estimates of their own percentile. These considerations point to a significant limitation in the tasks used by Kruger and Dunning (1999), Krueger and Mueller (2002), and us—all of these tasks, by design, provided participants with little information about how they were doing in either an absolute or a relative sense. For example, participants were not told whether their quiz answers were correct or not. Though many tasks in life do have this

quality, there are also many that do provide degrees of performance feedback. For instance, it is transparent how well one is doing at bowling in an absolute sense—pins are counted and totaled from frame to frame. Our tasks and Kruger and Dunning's also offered little information about how well others were likely to do. This, too, is probably a common enough situation, but there are notable exceptions. Those who participate in organized sports, for example, often have considerable experience observing the performances of other participants. Students, similarly, tend to learn where they have scored in a distribution.

The word-prospector task we used in Study 3 went part way toward reducing some sources of uncertainty, and it did not show differences between high and low performers's accuracy in estimating their standings. However, all of our tasks, including that one, follow the lead of Kruger and Dunning (1999) in providing only limited, ambiguous feedback. A full account of how higher and lower performers differ in their ability to recognize their own relative performance will need to consider what information is available in different environments and how different people make use of the various sources of information. The weak relationship between actual and perceived percentile in Studies 1 through 3 is likely to be a robust effect for similar tasks, but should be generalized with care.

Note, though, that estimates of relative standing are not only noisy, but also prone to systematic bias: People feel they are worse than average on tasks that are hard for everyone, and above average on tasks that are easy (confirming the finding of Kruger, 1999). This is an interesting phenomenon in itself, that merits further investigation. Proposed explanations include anchoring and insufficient adjustment (Kruger, 1999), confirmatory hypothesis testing (Menon, Block, & Ramanathan, 2002), and support theory (Brenner, Koehler, & Rottenstreich, in press). We hope to explore this issue, too, in future research.

Conclusions

Knowing how one's ability and performance compare to others' is an important kind of judgment, which is a component of many choices people make. Accordingly, it would be useful to understand more about the who, when, and why of errors in estimating relative standing. The primary goal of this paper was to determine the extent to which it is *everyone* who is unaware of relative position or just poor performers in particular. We have shown that, to a considerable extent, unawareness is universal 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'. Across the board, estimates are noisy, and they are prone to an overall bias depending on perceived task difficulty. Measures of actual ability are also inherently noisy. This produces methodological obstacles in the search for relations among skill, metacognition, and judgments of relative standing: Simple comparisons between estimates of and actual relative standing are not sufficient to provide a clear picture of who is a more accurate (at least on the kinds of tasks that have been used to date). Although our secondary analyses suggest that there may be some relationship between degree of miscalibration and skill level, noisy estimates plus overall task-biases are a substantial part of the story. We believe the research presented in this paper provides a foundation for future exploration of how and how well we know where we are on the curve.

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., Koehler, D. J., & Rottenstreich, Y. (2002, in press). Remarks on support theory: Recent advances and future directions. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment.*

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., 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.

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., & 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.

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.

Menon, G., Block, L. G., & Ramanathan, S. (2002). We're at as much risk as we are led to believe: Effects of message cues on judgments of health risk. *Journal of Consumer Research, 28*, 533-549.

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.

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.

Author Note

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

Correspondence concerning this article should be addressed to Katherine A. Burson, GSB, University of Chicago, 1101 E. 58th St., Chicago, IL, 60637.

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).

Table 1

Tests of Differences Between Means of Expected and Actual Percentile in Study 1 by Difficulty of Quiz and Quartile of Performance

	Quartile			
Domain and measure	Lowest		Highest	
Easy trivia				
Expected percentile	56.58	(13.53)	72.00	(20.98)
Actual percentile	12.25	(6.85)	88.84	(5.49)
Difference	$t(11) = 8.35$**		$t(9) = -3.09$**	
Moderate trivia				
Expected percentile	48.25	(23.37)	58.09	(17.68)
Actual percentile	9.03	(4.44)	86.22	(7.02)
Difference	$t(7) = 4.25$**		$t(10) = -4.43$**	

Note. Numbers shown in parentheses are standard deviations.

**$p < .01$

Table 2

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

		Quartile			
Domain and measure	Overall	Lowest		Highest	
Easy 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		$t(8) = 3.62$**		$t(5) = -3.63$*	
Hard 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		$t(8) = 3.61$**		$t(12) = -6.04$**	
Easy 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		$t(12) = 2.23$*		$t(7) = -2.90$*	
Hard 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		$t(9) = 2.19$		$t(11) = -4.93$**	

Easy 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		$t(8) = 1.07$		$t(12) = -5.59$**	
Hard 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		$t(8) = 2.59$*		$t(5) = -6.63$**	

Note. Numbers shown in parentheses are standard deviations.

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

Table 3

Tests of Miscalibration by Quartile of Performance on Each Quiz

		Quartile		
Domain	Difficulty rating	Lowest	Highest	Difference
Easy college acceptance rates	5.35	31.00	46.33	$t(13) = -1.04$
Hard college acceptance rates	7.15	38.33	46.04	$t(20) = -.61$
Easy pop songs on charts	6.77	15.00	23.75	$t(19) = -.82$
Hard pop songs on charts	7.62	15.50	36.17	$t(20) = -2.00$
Easy year of Nobel Prize	7.55	8.52	45.06	$t(20) = -3.11$**
Hard year of Nobel Prize	8.88	12.11	56.67	$t(13) = -4.97$**

**$p < .01$.

Table 4

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	
Easy word prospector	60.97	(22.41)	5.79	(1.77)
Typewriter	57.31	(22.25)	5.92	(1.70)
Overthrown	64.64	(22.27)	5.67	(1.85)
Hard word prospector	21.44	(17.45)	7.44	(1.52)
Petroglyph	25.73	(19.50)	7.20	(1.67)
Gargantuan	17.15	(14.36)	7.68	(1.33)
Mean	45.32	(28.96)	6.46	(1.85)

Table 5

Expected Overall Percentile, Actual Overall Percentile, and Miscalibration for the Word Prospector Problems of Study 3, By Quartile of Overall Performance

	Quartile			
	Lowest		Highest	
Easy word prospector				
Expected percentile	52.22	(9.72)	67.33	(21.97)
Actual percentile	12.00	(7.56)	87.00	(7.60)
Miscalibration	40.22	(7.95)	19.67	(16.78)
Hard word prospector				
Expected percentile	35.00	(13.84)	54.20	(19.94)
Actual percentile	11.90	(7.64)	87.00	(7.52)
Miscalibration	23.10	(19.12)	32.80	(16.93)

Note. Expected percentiles overall and actual percentile across the two words seen by each participant. Miscalibration was calculated as (estimated percentile – actual percentile) for the bottom quartile and (actual percentile – estimated percentile) for the top quartile. Numbers in parentheses are standard deviations.

Table 6

Magnitude of Average 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	
Easy word prospector: Overthrown[a]	28.59	(30.95)	6.38	(22.09)
Hard word prospector: Gargantuan[a]	11.98	(26.19)	26.10	(20.93)
Easy word prospector: Typewriter[b]	9.57	(43.01)	8.10	(27.63)
Hard word prospector: Petroglyph[b]	14.33	(26.35)	15.80	(23.06)

Note. Numbers in parentheses are standard deviations.

[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.

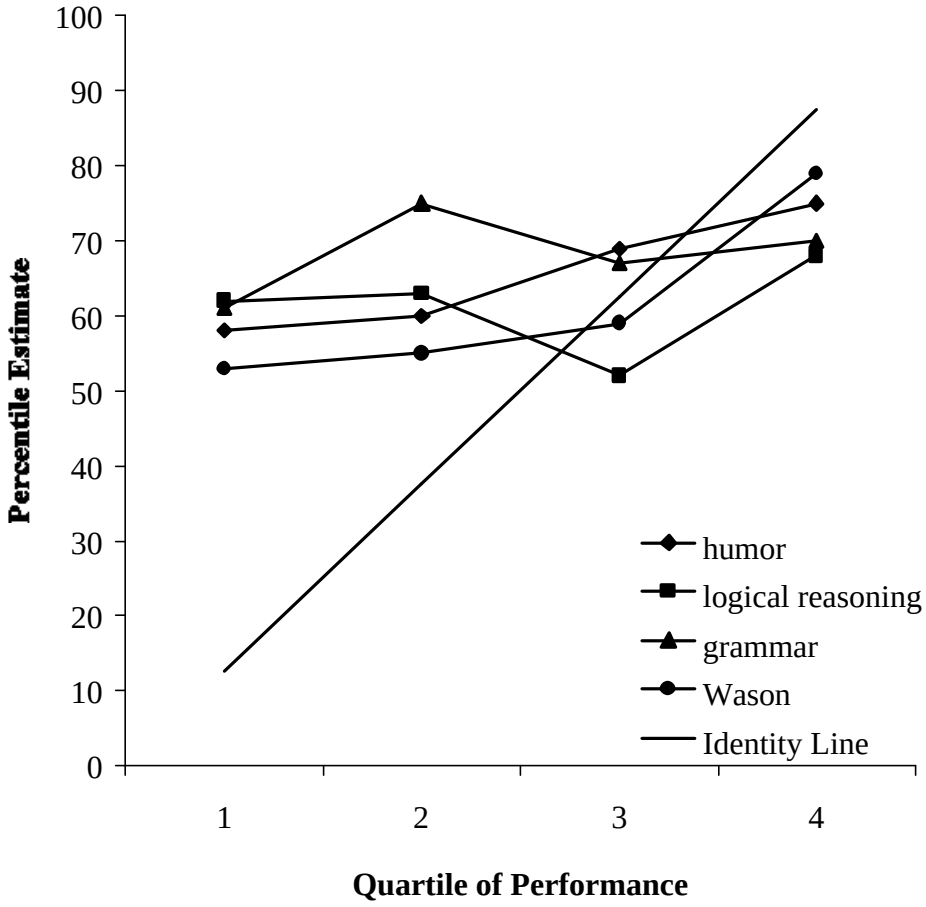
Figure 3. Hypothetical estimates of performance by actual quartile of performance on tasks of varying difficulty, assuming that less skilled participants are simply more error-prone in estimating their relative performance. Less skilled participants' estimates will regress more, and the mean to which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

Figure 4. Participants' estimates of performa nce percentile by quartile of actual performance on easy and moderately difficult tests of University of Chicago trivia in Study 1.

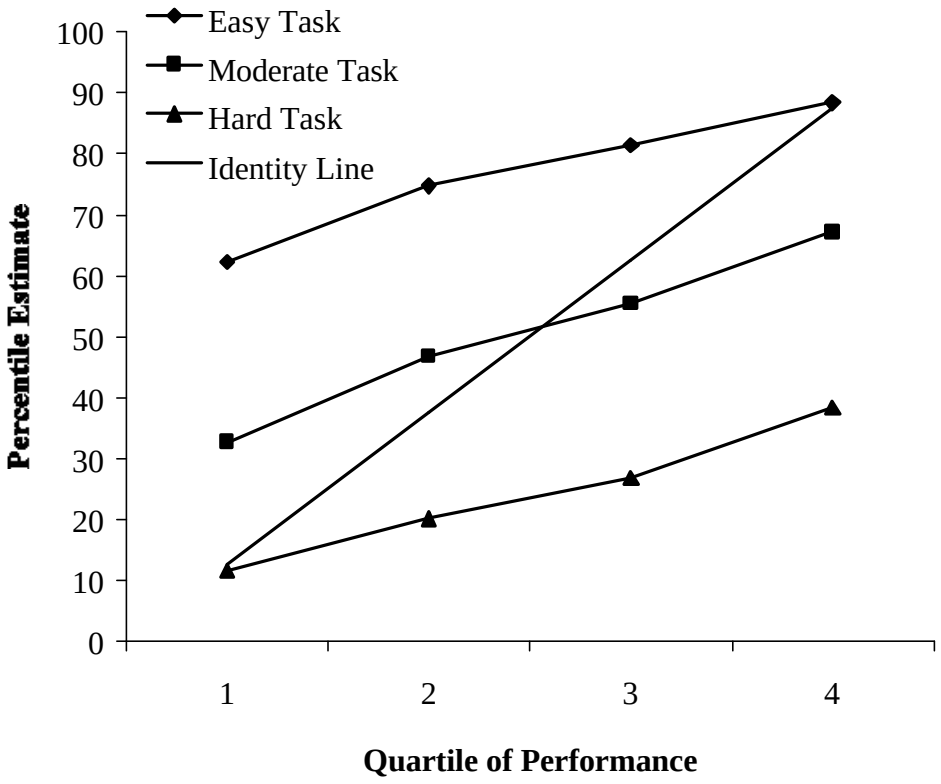
Figure 5. Participants' estimates of performance percentile by quartile of actual performance on six sets of estimates of varying difficulty in Study 2.

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

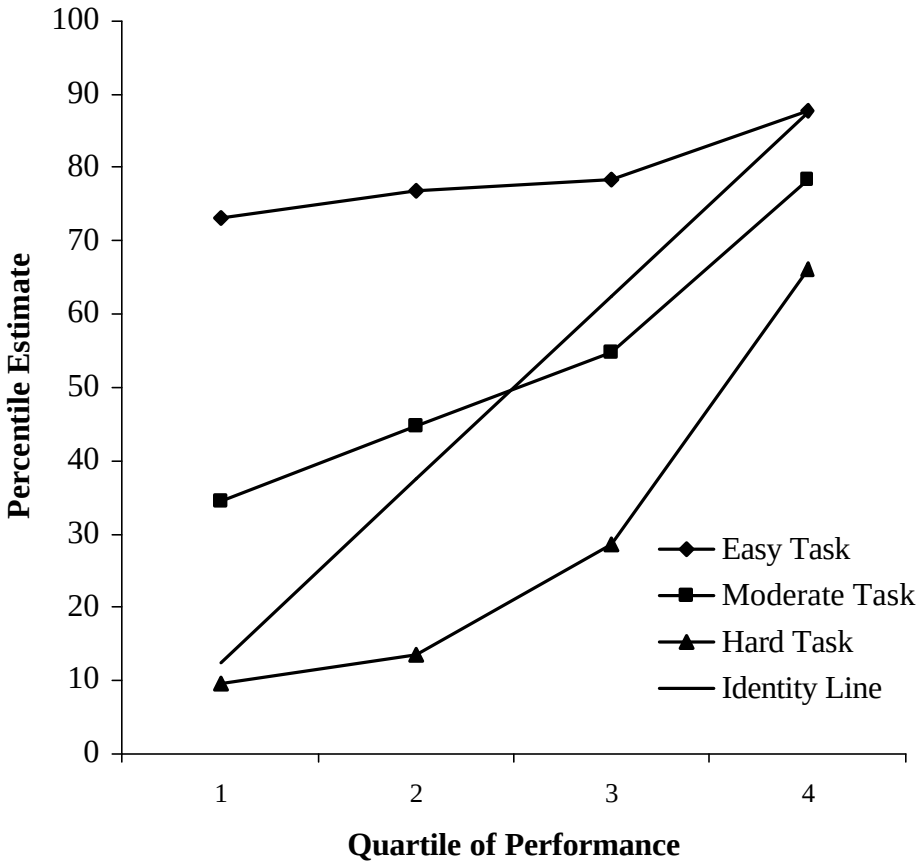
1


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

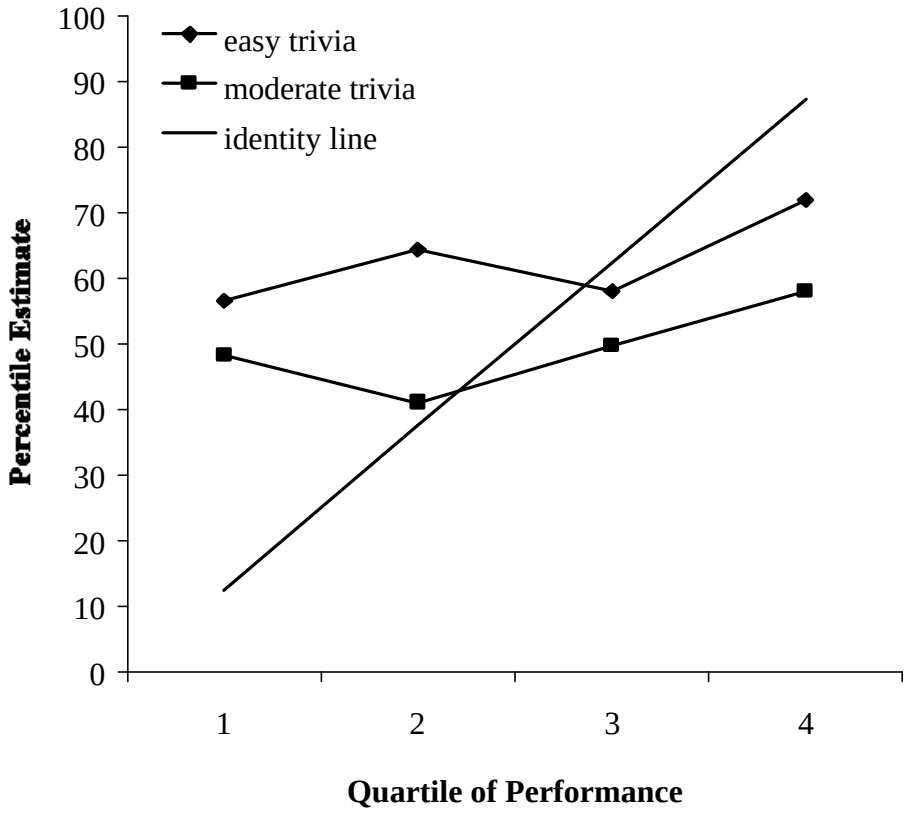
2


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

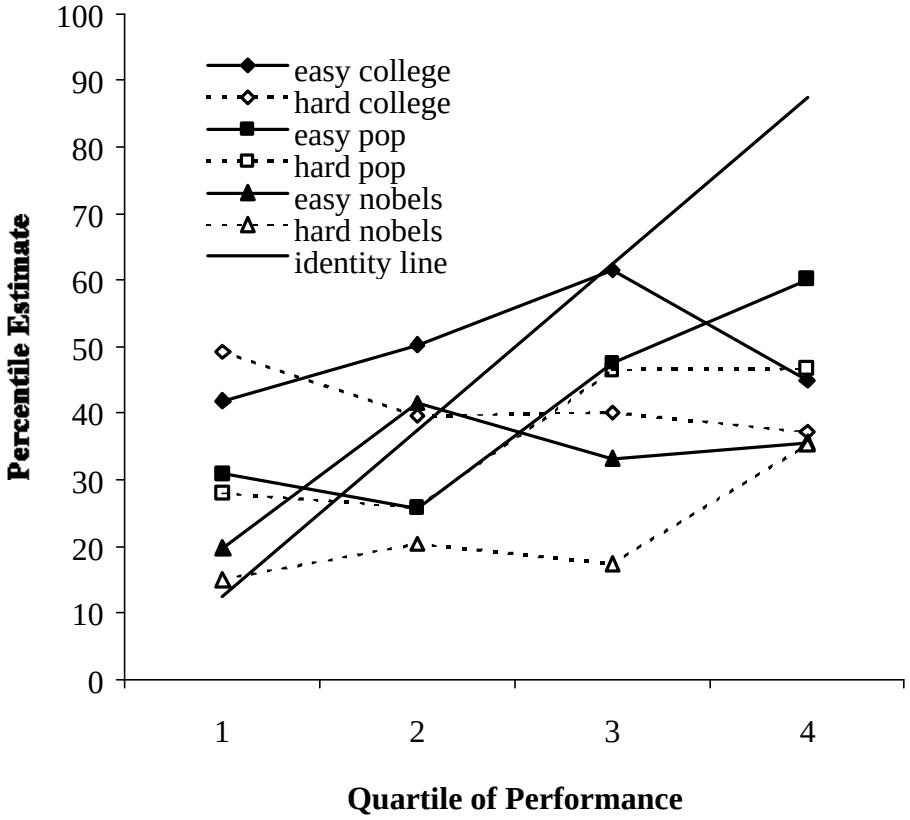
3


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

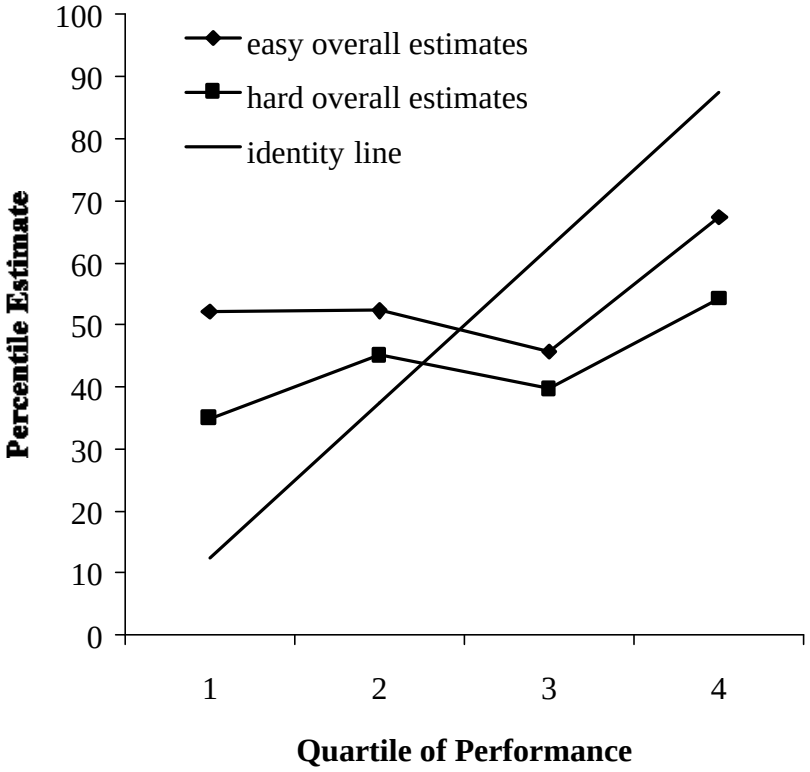
4


100
90
80
70
60
50
40
30
20
10
0
easy trivia
moderate 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
easy college
hard college
easy pop
hard pop
easy nobels
hard nobels
identity line
1
2
3
4
Quartile of Performance

6


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