Journal of Personality and Social Psychology
2006, Vol. 90, No. 1, 60–77

0022-3514/06/$12.00 DOI: 10.1037/0022-3514.90.1.60

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

People are inaccurate judges of how their abilities compare to others'. J. Kruger and D. Dunning (1999, 2002) argued that unskilled performers in particular lack metacognitive insight about their relative performance and disproportionately account for better-than-average effects. The unskilled overestimate their actual percentile of performance, whereas skilled performers more accurately predict theirs. However, not all tasks show this bias. In a series of 12 tasks across 3 studies, the authors show that on moderately difficult tasks, best and worst performers differ very little in accuracy, and on more difficult tasks, best performers are less accurate than worst performers in their judgments. This pattern suggests that judges at all skill levels are subject to similar degrees of error. The authors propose that a noise-plus-bias model of judgment is sufficient to explain the relation between skill level and accuracy of judgments of relative standing.

Keywords: calibration, above-average effect, better-than-average effect, judgment errors, unskilled–unaware

Research on overconfidence has found that subjective and objective measures of performance are poorly correlated (see Alba & Hutchinson, 2000, for a comprehensive review). Whereas most of this research compares confidence in one's estimates with one's actual performance, one particular line focuses on people's accuracy in estimating their ability compared with their peers. Such judgments are important in many contexts. In many societies, success in school, jobs, entrepreneurship, sports, and many other activities is largely a function of how one's ability and performance compare to those of 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 in many social and intellectual domains. However, the inaccuracy of this perception is not uniform. Empirically, Kruger and Dunning (1999, p. 1132) found across four studies that the "bottom quartile participants accounted for the bulk of the above average effects observed," with only a small amount of inflation accounted for by the rest of the participants. Figure 1 summarizes these results. Kruger and Dunning (1999) attributed the pronounced overestimation by the worst performers to a lack of metacognitive skill—the worst performers lack the knowledge required both to perform well and to evaluate whether they have performed well. 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 article, we refer to this as the "unskilled–unaware hypothesis." In explaining the above-average effect, Kruger and Dunning (1999) therefore proposed that "focusing on the metacognitive deficits of the unskilled may help to explain this overall tendency toward inflated self-appraisals" (p. 1122).

Explanations for the Unskilled–Unaware Pattern

The unskilled–unaware hypothesis has logical and intuitive appeal. As Kruger and Dunning (1999) pointed out, the skills required to write a grammatically correct sentence are similar to the skills required to recognize a grammatically correct sentence. The most incompetent individuals overstate their abilities in many contexts. One of this article'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 looked at only one judgment context—one in which participants on average believe that they are above average. In fact, research by Kruger (1999) showed that this condition is not as universal as it was once thought to be. Kruger 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 argued that participants anchor on their perception that they will perform well or poorly in

Katherine A. Burson, Ross School of Business, University of Michigan; Richard P. Larrick, Fuqua School of Business, Duke University; Joshua Klayman, Graduate School of Business, University of Chicago.

We thank Joachim Krueger, Justin Kruger, and John Lynch for their useful insights about this work and Fred Feinberg for his advice on statistical analyses.

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

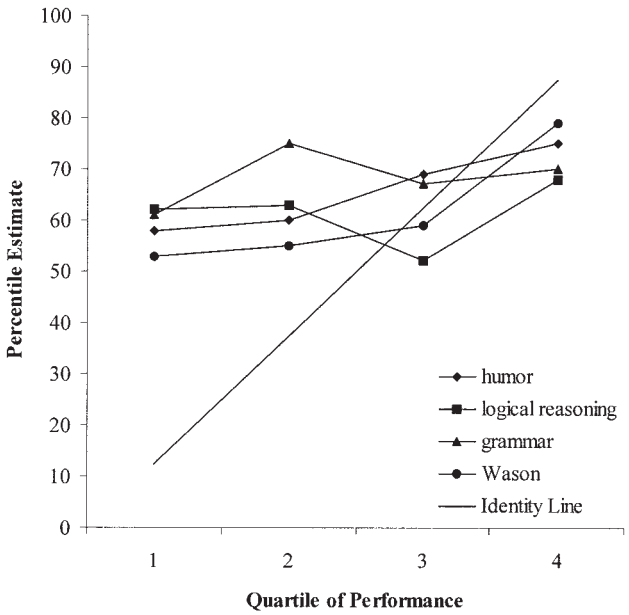

Figure 1. Participants' estimates of the percentiles of their performances relative to their peers, by quartile of actual performance in four experiments from Kruger and Dunning (1999). This pattern of results suggests that unskilled participants are more miscalibrated than are skilled participants.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

an absolute sense and adjust insufficiently for the fact that the task may be easy or hard for everyone (see also Chambers & Windschitl, 2004).

Do the unskilled contribute the bulk of erroneous assessments in more difficult tasks as well, when the average percentile estimate 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; see also Ackerman, Beier, & Bowen, 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 their estimates of relative performance regress toward the mean.[1] Additionally, people at all skill levels make percentile estimates that are biased upward. In other words, regardless of skill level, people do not have much knowledge about how they compare with 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) examined judgments of percentile on tasks with overall positive biases. The two explanations are difficult to distinguish in that context. In a published exchange, the two sets of authors focused on the question of whether metacognitive skills can be shown to mediate the difference between good and poor performers, and they disagreed. Kruger and Dunning argued that Krueger and Mueller's participants showed regression because of the modest reliability of the task. 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 or the lack of insight among all participants.

In the present studies, we take a different approach to investigate the cognitive processes underlying judgments of percentile. We vary task difficulty by selecting easier and harder domains and by varying the criteria for success. Our tasks include some for which there is no overall bias and some for which there is a negative (*worse than average*) bias.[2] We also address the reliability issue raised by Kruger and Dunning (1999, 2002) and ultimately use a split-samples method that removes regression effects due to task ambiguity and luck. These approaches permit us to test an extension and generalization of Krueger and Mueller's (2002) basic hypothesis, which we call a *noise-plus-bias* model. We propose that people at all performance levels are equally poor at estimating their relative performance (e.g., their judgments are noisy) and equally prone to overestimating their percentile on tasks that are perceived to be easy and underestimating it on tasks that are perceived to be hard (e.g., their judgments reflect task-induced bias). The results expected under this hypothesis are illustrated in Figure 2 (see the Appendix for the simulation model that generated these results).

An important implication of the noise-plus-bias account is that higher skilled performers are better judges of their percentile only for easy tasks. For difficult tasks, the opposite is true: The most skilled are the least accurate. Although poor performers account for the bulk of the above-average effect in easy tasks, good performers account for the bulk of the below-average effect in difficult tasks. We wish to emphasize that in the noise-plus-bias account, the apparent accuracy of good performers on easy tasks and poor performers on difficult tasks does not reflect insight, but is an accident of the match between actual percentile on a task and task-induced bias in perceived percentile. (This is why the most accurate participants in Kruger and Dunning's 1999, 2002, studies are those in the upper-middle percentiles. Such participants, by accident as much as by insight, report their position more accurately than do those in the top percentiles.)

This noise-plus-bias argument parallels an earlier critique made of the depressive realism literature. The initial finding in this research was that moderately depressed people make accurate judgments of their degree of control in situations with low control, whereas the nondepressed overestimated their control (Alloy & Abramson, 1979; Martin, Abramson, & Alloy, 1984). However, the question was raised, Do depressed people better discriminate degrees of control, or do they just possess a different mean level of bias in their estimates? When depressives and nondepressives were

[1] A similar explanation has been offered for miscalibration in confidence judgments in which overconfidence is greatest for those who are least accurate and for items that are most difficult (Erev, Wallsten, & Budescu, 1994; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1993, 1994; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Soll, 1996; Wallsten, Budescu, Erev, & Diederich, 1997).

[2] Krueger and Mueller (2002) also manipulated difficulty on their task, but their manipulation did not produce percentile estimates below the 60th percentile.

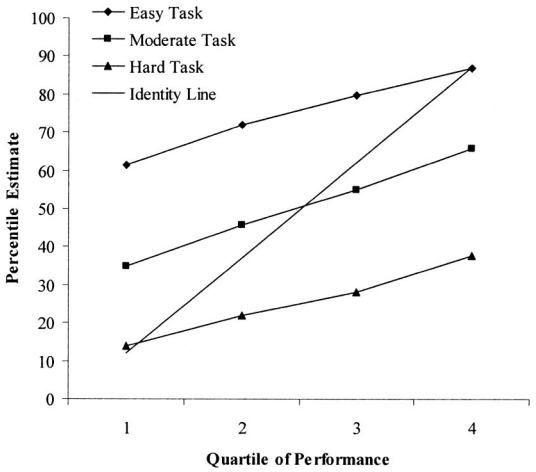

Figure 2. Hypothetical estimates of performance percentile by actual quartile of performance on tasks of varying difficulty, assuming that everyone is equally unaware of his or her ability and equally prone to the overall biasing effects of task difficulty.

compared in situations that permitted control, depressives underestimated their degree of control and nondepressives appeared more accurate. Alloy and Abramson concluded in a later collaboration (Dykman, Abramson, Alloy, & Hartlage, 1989; see also Coyne & Gotlib, 1983) that "neither depressed nor nondepressed subjects displayed differential accuracy in terms of being able to vary their judgments to achieve accuracy across changing situations" (p. 442), and they found no evidence of "a characteristic tendency for either group to process information in either a biased or unbiased way" (p. 442). In this literature, it turned out that the accuracy of perceived control was largely an accident of whether personal dispositions (a chronic tendency to estimate high or low levels of control) matched the degree of actual control available in a task.

We propose that if judgments of percentile show the pattern displayed in Figure 2 (i.e., parallel lines with modest upward slopes), then noise and bias across all performers provide a sufficient explanation for the unskilled–unaware pattern in Figure 1. There is no need to appeal to metacognitive differences between better and worse performers. On the other hand, significant departures from this pattern would suggest that such differences may be important. Figure 3, for example, shows one plausible instantiation of the unskilled–unaware hypothesis when it is extended to difficult tasks. Here, worse performers—lacking metacognitive insight—are more prone to noise and task-induced bias, whereas better performers are relatively immune to these errors. (See the Appendix for details of the simulation.)

Although Kruger and Dunning did not explicitly extend their theory of metacognitive differences to tasks that yield worse-than-average effects, they did propose that "for the incompetent to overestimate themselves, they must satisfy a minimal threshold of knowledge, theory, or experience that suggests to themselves that they can generate correct answers" (p. 1132). As counterexamples, the authors listed tasks that are impossible for most people to perform: translating Slovenian proverbs, reconstructing 8-cylinder engines, and diagnosing acute disseminated encephalomyelitis. Kruger and Dunning expected that most people would recognize that they are poor performers on these tasks and, if they showed a bias, would rate themselves as worse than their peers (presumably as a kind of floor effect). They did not, however, explicitly discuss how the minority of competent performers would assess themselves on these tasks.

One interpretation of these claims is that there is a boundary to the unskilled–unaware hypothesis such that the relationship weakens in domains where average perceived percentile estimates are near or below the 50th percentile. Thus, for domains where the average perception is above the 50th percentile, a performance–metacognition association holds such that those who perform worse at a task are less able to assess their own performances and those of others, and the bulk of the error in judging relative performance is produced by poor performers. However, as the average perceived percentile for a task decreases, so does the performance–metacognition association. We believe that this bounded version of the unskilled–unaware account is an interesting possibility but would argue that the noise-plus-bias account generalized from Krueger and Mueller (2002) is a sufficient and more parsimonious explanation for the pattern shown in Figure 2. We return to this issue in the General Discussion.

Alternative Measures of Inaccuracy

The noise-plus-bias account suggests some specific methodological improvements for evaluating inaccuracy in perceived performance. The main measure of accuracy used by Kruger and Dunning is the difference between perceived percentile and actual percentile, which they term *miscalibration.* However, the use of miscalibration as a measure has limitations. One of the terms in the calculation, perceived percentile, is sensitive to task-induced bias (Kruger, 1999): Higher percentiles are reported on tasks that feel easy than on tasks that feel hard. Task-induced bias creates a discrepancy between perceived percentile and actual percentile that is not attributable to metacognitive differences. Thus, on tasks that induce an upward bias, poor performers appear more miscali-

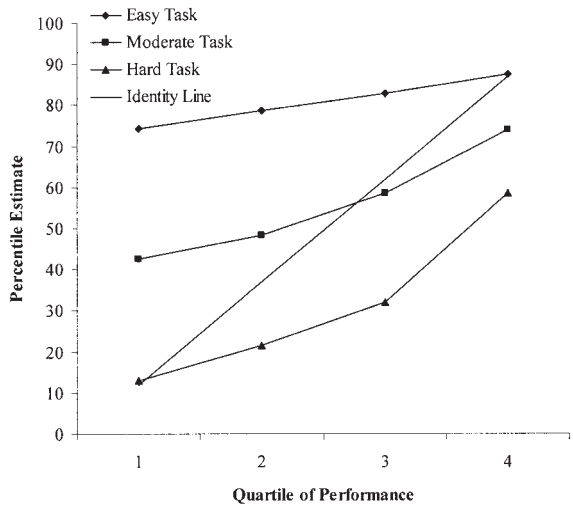

Figure 3. Hypothetical estimates of performance percentile by actual quartile of performance on tasks of varying difficulty, assuming that less skilled participants are simply more error-prone in estimating their relative performance. Less skilled participants' estimates will regress more, and the mean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

brated than top performers; the opposite pattern occurs on tasks that induce a downward bias. We present an additional measure of accuracy, which we term *sensitivity,* that generally is not influenced by overall task-induced bias. We test whether the correlation between percentile judgments and actual percentiles differs between bottom and top performers.

The noise-plus-bias account implies a second methodological consideration. Participants found in the lowest quartile or the highest quartile on a given test fall within these segments of a distribution partly because of true 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. Consequently, participants' estimates of their performance will tend to be regressive, but regressiveness will be counted as error in their perceptions. That is, estimates of miscalibration for the best and worst performers will be exaggerated. In our final experiment, we adapt a split-sample method developed by Klayman, Soll, Gonzalez-Vallejo, and Barlas (1999) in studies of overconfidence and also used by Krueger and Mueller (2002). This method reduces the contribution of statistical regression to miscalibration, removing this source of bias from the measure of judgmental error. Specifically, we separated participants according to their actual percentile on one subtask and measured their degree of miscalibration on the other subtask.

In the remainder of this article, we describe three experiments that manipulated the perceived difficulty of tasks and hence participants' beliefs about their percentile. By sampling a wider range of judgment contexts and taking regression effects into account, these studies provide evidence on a fundamental question about the psychological processes underlying comparative judgments: 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 were paid two dollars for participating in this 15-min 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 questions were answered well above chance level (as, of course, were the easier trivia questions).

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 in the study, and the difficulty of the task for themselves and for the average participant on a scale ranging from 1 (*very easy*) to 5 (*very difficult*). The use of the percentile scale was explained in detail. For half of the participants, performance estimates appeared first, followed by the difficulty questions, and for half, these segments were presented in the reverse order.

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 did the easier trivia (M = 10.62 vs. M = 14.64), $t(87) = 7.53$, $p < .001$, $d = 1.60$, and performance on both tasks was better than the 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 vs. 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) where 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 seem to have been only moderately difficult, whereas the easier trivia were indeed easy. Therefore, we call our harder condition the "moderate" condition.

Asymmetry by quartiles. We looked for a performance–metacognition relationship in two ways. The first involved looking at miscalibration measures, as introduced by Kruger and Dunning (1999), measuring the difference between estimated performance and actual performance. The second method involved looking at sensitivity to relative standing, measuring the correlation between estimated performance and actual performance.

To examine how estimated percentiles varied with skill level, we divided the participants in each condition into four groups on the basis of actual performance (following Kruger & Dunning, 1999). These groups represented four quartiles of performance relative to other participants in that condition. As shown in Figure 4, percentile estimates were fairly uniform across quartiles on both the easy and the moderate task and were lower, on average, on the moderate task. An analysis of variance (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, $F(3, 81) = 2.60$, $p =$

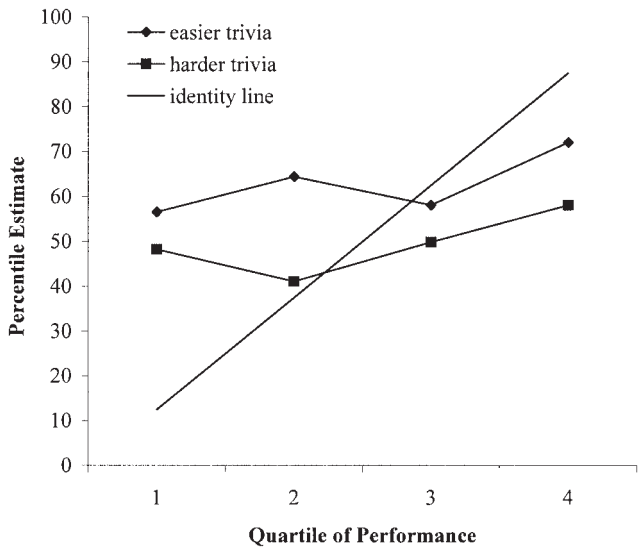

Figure 4. Participants' estimates of performance percentile by quartile of actual performance on easier and harder tests of University of Chicago trivia in Study 1.

.058, $\eta^2 = .09$, but no significant interaction. The main effect of quartile was explored by regressing percentile estimates on actual percentiles and revealed a significant linear relationship, $B = .142$, $SE = .068$, $\beta = .219$, $t(87) = 2.10$, $p = .039$.

Paired t tests confirmed some of Kruger and Dunning's (1999) findings. In both conditions, those in the lowest quartile overestimated their mean percentile, and those in the highest quartile underestimated theirs. Participants in the lowest quartile on the easy trivia were actually in the 12th percentile but thought they would be in the 57th, $t(11) = 8.35$, $p < .01$, $d = 2.32$; on the moderate trivia, these participants 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 highest quartile on the easy trivia were actually in the 89th percentile but thought they would be in the 72nd, $t(9) = -3.09$, $p = .013$, $d = -.93$; on the moderate trivia, these participants 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 miscalibration between highest and lowest performers in a direct statistical test, we coded errors as estimated percentile minus actual percentile for the lowest quartile and actual percentile minus estimated percentile for the highest 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 (1999); the lowest quartile was much more miscalibrated than was the highest ($M_{\text{lowest}} = 44.34$ vs. $M_{\text{highest}} = 16.84$), $t(20) = 3.59$, $p = .002$, $d = 1.54$. However, in the moderate condition, the first and fourth quartiles did not differ significantly ($M_{\text{lowest}} = 39.23$ vs. $M_{\text{highest}} = 28.13$), $t(17) = 1.03$, $p = .32$, $d = .48$.

A similar comparison of miscalibration was made with respect to judgments of absolute performance, in which we compared estimated minus actual number of correct answers for participants in the lowest quartile with actual minus estimated number of correct answers for participants in the highest quartile. The easier condition showed no significant difference in miscalibration between the lowest and highest quartiles ($M_{\text{lowest}} = 1.08$ vs. $M_{\text{highest}} = 3.10$, $t(20) = -1.48$, $p = .156$, $d = -.63$). However, in the moderate condition, the lowest quartile was significantly less miscalibrated than was the highest quartile ($M_{\text{lowest}} = -0.38$ vs. $M_{\text{highest}} = 4.73$, $t(17) = -3.47$, $p = .003$, $d = -1.60$. This pattern of miscalibration on estimated scores suggests that participants of both skill levels had only modest insight into their scores.[3]

Participants' awareness may be manifested in ways other than the difference between perceived performance and true performance. Perhaps there are skill-level differences in how sensitive participants are to their relative standing on these performance measures. In other words, skilled performers' estimates might be better correlated with the truth than unskilled performers' estimates, setting aside any general task-induced bias. To examine a potential metacognitive difference in this way, we compared the statistical relationship between estimated and actual performance separately among bottom-half and top-half performers. The first variable we examined was the relationship between actual and perceived percentile. When percentile estimates were regressed on actual percentiles for all performers, the standardized coefficients were .253 in the easy condition and .215 in the moderate condition. We then calculated the same regressions for bottom-half and top-half performers separately and compared them using Chow's test for differences between coefficients drawn from independent samples (Chow, 1960; Pindyck & Rubinfeld, 1998, pp. 133–134). The standardized regression coefficients for the bottom-half performers were not significantly lower than those for the top-half performers on either trivia task, although the pattern of coefficients was consistent with a metacognitive difference favoring the skilled performers, $\beta_{\text{bottom}} = .134$ vs. $\beta_{\text{top}} = .526$, $F(1, 44) = 2.03$, $p = .161$, for the easier trivia and $\beta_{\text{bottom}} = -.206$ vs. $\beta_{\text{top}} = .235$, $F(1, 39) = 2.05$, $p = .160$, for the moderately difficult trivia.[4] Turning now to estimates of number correct, the relation between estimated and actual scores in the sample as a whole was positive but not significant, $\beta_{\text{easy}} = .267$, $p = .069$, and $\beta_{\text{moderate}} = .240$, $p = .126$. Chow tests showed that the standardized coefficient for the

[3] It is possible to argue that we have violated the "minimum threshold" standard required by Kruger and Dunning in this experiment. We point out that the more difficult trivia here were not particularly difficult, as measured by either objective performance or subjective assessment. However, to look more carefully at the possibility that participants were faced with an impossible task in the "moderate" condition of Study 1, we conducted the following analyses. The trivia in this experiment have a benchmark for chance performance of 6.67 correct out of 20 questions. We assigned a threshold of competence somewhat above chance, since it is likely that some of the participants who performed slightly better than 6.67 correct were actually incompetent but lucky. Though we inevitably discarded participants who were competent but performed near chance due to bad luck, we conservatively eliminated all participants who got fewer than nine correct answers on the quizzes. We believe this is a sufficiently rigorous test of whether the unskilled (but not "incompetent") are unaware on tasks of varying difficulty. However, one might regard the minimum threshold of competence to be the subjective rather than the objective measure of ability. Therefore, we also repeated our analysis eliminating all participants who *believed they would* score fewer than nine points on the trivia. These analyses showed no qualitative change in our results. When we limited our tests to participants who got nine or more correct, the easier trivia test showed the less skilled as more miscalibrated, $M_{\text{lowest quartile}} = 42.95$ vs. $M_{\text{highest quartile}} = 16.91$, $t(20) = 3.76$, $p = .001$, but the moderate trivia test showed little difference between the lowest and highest quartile's miscalibration, $M_{\text{lowest quartile}} = 29.19$ vs. $M_{\text{highest quartile}} = 35.58$, $t(14) = -0.72$, $p = .485$. Similarly, if we limited our tests to participants who merely *thought* they would get nine or more correct, the less-skilled were significantly more miscalibrated on the easy trivia test, $M_{\text{lowest quartile}} = 44.22$ vs. $M_{\text{highest quartile}} = 14.23$, $t(18) = 3.77$, $p = .001$, but not on the moderate trivia test, $M_{\text{lowest quartile}} = 38.45$ vs. $M_{\text{highest quartile}} = 28.35$, $t(9) = 1.26$, $p = .239$. Thus, even among the demonstrably competent or those who think they are competent, the less skilled are not more unaware of their standing than the skilled on more difficult tasks.

[4] The Chow (1960) test examines whether a regression coefficient calculated in one subgroup differs significantly from the same coefficient calculated in a second subgroup. The test is comprised of the residual sum of squares (RSS) from three regressions: one performed within each subgroup and one using the pooled data. To examine only the differences due to slope, we fixed the constants for the regressions within subgroup at 0 by standardizing the variables within subgroup first. The data used in the pooled regression, therefore, were composed of the transformed data. With one predictor variable, these sums of squares are combined to generate an F ratio using the formula $[RSS_{\text{pooled}} - (RSS_1 + RSS_2)]/[(RSS_1 + RSS_2)/(n_1 + n_2)]$. The degrees of freedom are 1, $(n_1 + n_2)$. (See Pindyck and Rubinfeld, 1998, pp. 133–134.) Piecewise linear regressions (Pindyck & Rubinfeld, 1998, pp. 136–137) yielded conclusions identical to those from the Chow tests.

bottom-half performers was not significantly different from that for top-half performers on either the easy or the moderate trivia, and the direction of the effects differed between the two conditions (for easy trivia, $\beta_{bottom} = .288$ vs. $\beta_{top} = .478$, $F(1, 44) = .48$, $p = .491$; for moderate trivia, $\beta_{bottom} = .172$ vs. $\beta_{top} = -.197$, $F(1, 39) = 1.41$, $p = .242$). The observed patterns of sensitivity suggest the possibility of a metacognition–performance relationship but are not conclusive.

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 were similarly accurate in their percentile estimates. These results are consistent with the hypothesis that estimating one's percentile is difficult regardless of skill level.

For both perceived percentile and perceived score, who appears most accurate in terms of miscalibration depended on the difficulty of the task; the moderately difficult task made unskilled participants look about as aware as skilled participants in terms of percentile estimates and more aware than skilled participants in terms of score estimates. The easier task made skilled participants look more aware than unskilled participants in terms of percentile estimates and equally aware as unskilled participants in terms of score estimates. Other tests found only suggestive evidence that top-half performers were more sensitive to relative standing than were bottom-half performers. This possible difference is interesting. However, it does not result in consistently better accuracy for skilled performers. Rather, the difference in calibration between skilled and unskilled participants depends mostly on the biasing effects of task difficulty that span the spectrum of skills.

Study 2

The second experiment looked 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 percentile on 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 calibrated and high performers would appear less calibrated. 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 line of Figure 2). We do not suggest that poor performers are actually more perceptive than high performers in these tasks. 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 9 dollars for this 45-min experiment.

Design. Three variables were manipulated within participant: domain, question set, and difficulty. There were five domains: college acceptance rates, dates of Nobel prizes in literature, 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 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 for the harder version vs. within 30 years for the easier 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 5 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 of richest people and games won by hockey teams) included tests that were so difficult or so easy that almost all of the participants performed at the same level, making it hard to assign meaningful percentiles 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 scale ranging from 1 (*very easy*) to 10 (*very difficult*). 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 the following passage:

> 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 5 years of the actual year.

Results

Manipulation check. For each of the three dependent measures—actual performance, estimated performance, and estimated difficulty—we performed a multivariate analysis of variance (MANOVA) with domain and difficulty as within-participant variables and order (harder first or easier first) as a between-participants variable. The difficulty manipulation worked: The

harder conditions were perceived as significantly more difficult ($M = 7.94$) than were 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 vs. $M = 68.77\%$ correct), $F(1, 35) = 808.15$, $p < .001$, $\eta^2 = .96$).

Percentile estimates. Overall, the mean percentile estimate was 37.04. This was significantly less than 50, $t(39) = -4.68$, $p < .001$. A MANOVA with domain, difficulty, and order as independent variables showed that some domains were perceived as 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 more difficult the domain seemed to participants, the lower the percentile estimate ($M_{\text{colleges}} = 45.98$, $M_{\text{pop songs}} = 39.47$, and $M_{\text{Nobel Prize}} = 26.98$, $F(2, 70) = 14.25$, $p < .001$, $\eta^2 = .29$). Additionally, percentile estimates were lower in the harder (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 exacting evaluation standards or through 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 the two panels of Figure 5, the overall picture was one of a fairly uniform level of percentile estimates across quartiles within each domain. For those in the highest quartile, estimated percentiles were significantly lower than actual percentiles in each of the combinations of domain and difficulty. For those in the lowest quartile, estimated percentiles were significantly higher than actual percentiles in most cases.

Our prediction was that when mean percentile estimates were near 50, there would be no difference in miscalibration between lowest and highest performers. As mean percentile dropped below 50, lowest performers would show better calibration than highest performers. The results bear this out, as shown in Table 1. There were three sets of questions with mean percentile estimates greater than 40 (averaged across the lowest and highest performers): easier colleges, harder colleges, and easier pop songs. In these three tasks, lowest and highest performers did not differ significantly in miscalibration (ts < 1.04, ps $> .32$, ds $< .25$). In one other task (harder pop), mean percentile estimates were between 30 and 40. Lowest performers were marginally better calibrated than highest performers, $t(20) = -2.00$, $p = .06$, $d = -.82$. The two domains with a mean percentile estimate less than 30, easier and harder Nobels, showed lowest performers being significantly better calibrated than highest performers, $t(20) = -3.11$, $p = .006$, $d = -1.30$ and $t(13) = -5.00$, $p < .001$, $d = -2.48$, respectively.

Table 1
Perceived Percentiles, Actual Percentiles, and Miscalibration for Each Trivia Quiz in Study 2, by Quartile of Performance on that Quiz

		Quartile							
		Lowest				Highest			
Subdomain and measure	Overall *M*	*M*	(*SD*)	*t*	*df*	*M*	(*SD*)	*t*	*df*
Easier college acceptance rates									
Perceived percentile	43.07	41.89	(26.21)			44.83	(30.28)		
Actual percentile		10.89	(5.61)			91.17	(4.65)		
Miscalibration		31.00		3.62**	8	46.33		3.63*	5
Harder college acceptance rates									
Perceived percentile	42.09	49.22	(29.28)			37.15	(25.77)		
Actual percentile		10.89	(5.02)			83.19	(8.44)		
Miscalibration		38.33		3.61**	8	46.04		6.04**	12
Easier pop songs on charts									
Perceived percentile	42.00	30.85	(23.59)			60.13	(27.95)		
Actual percentile		15.85	(8.45)			83.88	(8.10)		
Miscalibration		15.00		2.23*	12	23.75		2.90*	7
Harder pop songs on charts									
Perceived percentile	38.18	28.00	(24.07)			46.67	(24.21)		
Actual percentile		12.50	(7.15)			82.83	(7.93)		
Miscalibration		15.50		2.19	9	36.17		4.93**	11
Easier year of Nobel Prize									
Perceived percentile	29.14	19.78	(20.71)			35.62	(27.63)		
Actual percentile		11.26	(6.59)			80.67	(6.31)		
Miscalibration		8.52		1.07	8	45.06		5.59**	12
Harder year of Nobel Prize									
Perceived percentile	28.53	24.00	(15.17)			35.33	(18.62)		
Actual percentile		12.00	(3.75)			92.00	(3.87)		
Miscalibration		12.00		2.59*	8	56.67		6.63**	5

Note. Overall *M* is the average perceived percentile across lowest and highest quartiles combined. Miscalibration = perceived percentile minus actual percentile for the lowest quartile and actual percentile minus perceived percentile for the highest quartile. The *t* test is a paired *t* test on actual versus perceived percentile testing whether miscalibration is significantly different from 0. Tests of the miscalibration measure between lowest and highest quartiles are reported in the text.
* $p < .05$. ** $p < .01$.

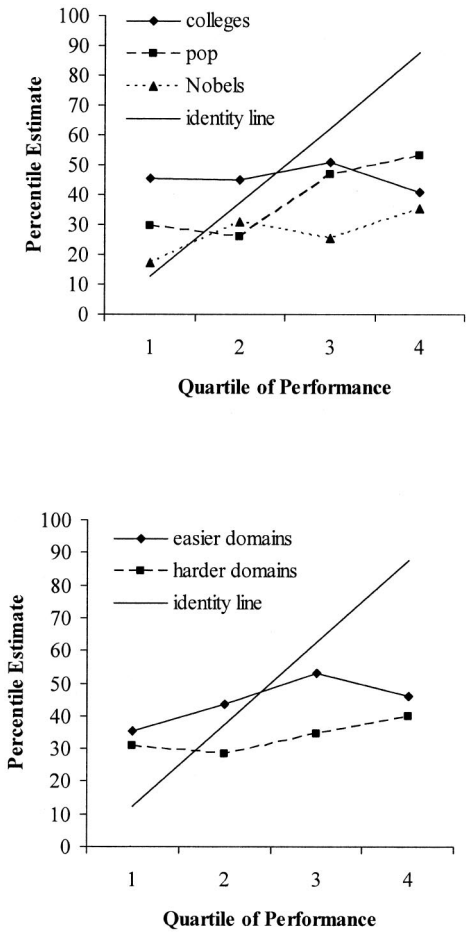

Figure 5. (Top) Participants' estimates of performance percentile by quartile of actual performance by domain in Study 2. (Bottom) Participants' estimates of performance percentile by quartile of actual performance on easier and harder tests in Study 2. College refers to college acceptance rates; pop refers to popular songs; Nobels refers to Nobel Prizes in Literature.

Next, we examined participants' sensitivity to relative standing by looking at the correlation between actual and estimated percentile. In the overall sample, percentile estimates were not significantly related to actual percentiles in either condition of colleges or Nobel Prizes ($\beta_{\text{easier colleges}} = .183$, $\beta_{\text{harder colleges}} = -.181$, $\beta_{\text{easier Nobels}} = .150$, $\beta_{\text{harder Nobels}} = .145$, *ns*), but were significantly related to actual percentiles in both conditions of pop songs ($\beta_{\text{easier pop}} = .483$, $p = .002$ and $\beta_{\text{harder pop}} = .389$, $p = .013$). We then regressed estimated percentile on actual percentile separately for top-half performers and bottom-half performers. There were no significant differences between bottom-half and top-half performers. The direction of the differences was not consistent across domains and conditions, although the two strongest positive coefficients did occur among the top-half performers. For easier colleges, $\beta_{\text{bottom}} = .193$ versus $\beta_{\text{top}} = -.237$, $F(1, 36) = 1.79$, $p = .189$; for harder colleges, $\beta_{\text{bottom}} = -.267$ versus $\beta_{\text{top}} = -.090$, $F(1, 37) = 0.31$, $p = .583$. For easier pop, $\beta_{\text{bottom}} = -.048$ versus $\beta_{\text{top}} = .438$, $F(1, 37) = 2.49$, $p = .123$; for the harder version, $\beta_{\text{bottom}} = .139$ versus $\beta_{\text{top}} = .008$, $F(1, 37) = .16$, $p = .688$. For easier Nobel Prizes, $\beta_{\text{bottom}} = .267$ versus $\beta_{\text{top}} = .013$, $F(1, 36) = 0.60$, $p = .442$; on the harder version, $\beta_{\text{bottom}} = -.129$ versus $\beta_{\text{top}} = .391$, $F(1, 37) = 2.56$, $p = .118$.

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 was driven by the task difficulty: With domains that feel harder (e.g., Nobel Prizes) and with more exacting criteria, low performers were better-calibrated than high performers, producing an unskilled–*aware* pattern. As with the apparent advantage of skilled performers in easy domains, the apparent accuracy of the *un*skilled in hard domains also is an accident of task difficulty, not an indication of greater awareness. Tests of sensitivity to relative standing found no consistent pattern of differences between top-half and bottom-half performers in this study. In Study 2, as in Study 1, we found only a weak positive relation between objective and subjective measures of relative performance. A potential concern is that these might be tasks for which relative performance is for some reason 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) made a similar point in their critique of Krueger and Mueller's (2002) studies, suggesting that tasks with low reliability have a large random component and are thus unpredictable. (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 split-half reliabilities ranged from −.24 on one set of easier pop music estimates to .52 on one set of harder Nobel Prize estimates). We note that the apparent unskilled–*aware* pattern held within the latter, most reliable subdomain, $t(7) = -3.73$, $p = .008$, where the lowest quartile showed better calibration ($M_{\text{lowest}} = 14.00$) than did the highest quartile ($M_{\text{highest}} = 65.33$). However, one might wish to have more evidence about the relation between skill level and estimates of percentile in more reliable, predictable tasks.

Study 3

In this study, we used a task that is more amenable to prediction of one's percentile 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 percentile. 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,* and so forth. 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 percentile is of course the performance of others. Here, too, par-

ticipants 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 (such as SAT Verbal percentiles), 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 participants two different word-prospector problems of similar difficulty and asked them for estimates about their percentiles on each word individually and overall. This facilitated two approaches for comparing predicted to actual performance at different levels of ability. The first approach was to separate participants 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 (from Klayman et al., 1999; see also Krueger & Mueller, 2002) was to separate participants according to their performance on one subtask, and measure their miscalibration on the other subtask. In other words, one task was used only to classify participants as having high or low task skills, ignoring their subjective estimates of performance for that task. Next, we took the difference between the actual and estimated performance on the other task to obtain our measure of accuracy. This second sample provides a noisy measure of true ability, which makes the resulting estimate of miscalibration conservative, but it eliminates the bias introduced when perceived and actual performance are measured within the same sample. As an illustration of the same-sample bias, consider that those found in the lowest quartile or the highest quartile on a given test fall in these segments of the distribution partly because of ability and partly because of bad and good luck in that sample, 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 in a given sample. Thus, their estimates of their performance will justifiably be regressive, but this will be counted as miscalibration when compared with their actual performance. Consequently, those with very low actual percentiles will appear to greatly overestimate their perceived percentile and those with very high actual percentiles will appear to greatly underestimate their perceived percentile. In the split-sample method, 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 actual percentile on one subtask and measuring miscalibration (estimated vs. actual percentile) on another subtask provides a luck-neutral way of comparing the estimates of good and poor performers. (For a more detailed explanation of how this method removes the biasing effects of regression to the mean, see Klayman et al., 1999.)

Method

Participants. Seventy-six University of Chicago students were recruited with advertisements posted around campus and were paid 5 dollars for their participation, which required approximately 15 min.

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 3 min to work on each. Those in the easier condition received two easier words (*typewriter* and *overthrown*) and were given 5 min 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 also repeated at the top of the page containing the 10-letter word. Participants received points for each letter of each correct word they spelled and lost points for each letter of nonexistent, repeated, or misspelled words. For example, if a participant looking at the word *gargantuan* spelled the word *grant,* 5 points would be counted toward the overall score. However, if the participant spelled the nonexistent word *naut,* 4 points would be subtracted from the overall score.

After reading the page of instructions, the experimenter repeated the instructions and the rules for scoring. Next, 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 3 or 5 min, 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 ranging 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 5-min, unrelated questionnaire. Next, they were given 3 min (harder condition) or 5 min (easier condition) to repeat the task using a different 10-letter word. Lastly, after the experimenter stopped them, the participants 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 ("how good are you at finding 4-, 5-, and 6-letter words in 10-letter words?").

Results

The present design affords several variations in how to measure performance and accuracy of estimates. One can examine overall performances across the two tasks completed by each participant, take separate measures for each task, or take the average of the two tasks measured separately. We found no theoretically important differences among the three variations. We report the results from the first, that is, from measures and estimates of overall performance across each participant's two word-prospector problems.

Manipulation checks. First, we checked the reliability of the task by comparing the first half with the second half. The split-half reliability was very high for both the easier and harder versions (.74 and .78, respectively). Next, we checked the difficulty manipulation using MANOVAs, with difficulty as a between-participants variable and first versus 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 that the word *typewriter* was more difficult 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

Table 2
Performance Scores and Ratings of Difficulty on Each Word Prospector Problem in Study 3

Domain and word	Score		Difficulty rating	
	M	(*SD*)	*M*	(*SD*)
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)

effect of difficulty, $F(1, 68) = 5.07$, $p = .028$, $\eta^2 = .07$, and an interaction between difficulty and estimated versus 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 was inevitable, given that quartile was determined by the same performance that determined actual percentiles. However, follow-up tests using regression showed that there was also a positive linear relationship between estimated percentiles and actual percentiles, $B = .224$, $SE = .072$, $\beta = .343$, $t(73) = 3.118$, $p = .003$; participants in higher quartiles of performance gave higher estimates of performance than did participants in lower quartiles (see Figure 6). There was also an interaction between quartile and estimated versus actual percentile, $F(3, 68) = 42.77$, $p < .001$, $\eta^2 = .65$. Those in the highest quartile underestimated their percentile ($M_{\text{easier estimate}} = 67.33$ vs. $M_{\text{easier actual}} = 87.00$ and $M_{\text{harder estimate}} = 54.20$ vs. $M_{\text{harder actual}} = 87.00$), whereas those in the lowest quartile overestimated theirs ($M_{\text{easier estimate}} = 52.22$ vs. $M_{\text{easier actual}} = 12.00$ and $M_{\text{harder estimate}} = 35.00$ vs. $M_{\text{harder actual}} = 11.90$). There was no three-way interaction between quartile, difficulty, and estimated versus actual measures (Fs < 1). That is, there was no evidence to contradict the hypothesis that the estimate lines for easier and harder tasks are parallel.[5]

Asymmetry by quartiles. As in Studies 1 and 2, we calculated miscalibration as estimated overall percentile minus actual overall percentile for the lowest quartile and actual overall percentile minus estimated overall percentile for the highest quartile. In the easier condition where the average percentile estimate across the lowest and highest quartiles was 59.78, miscalibration was significantly greater in the lowest quartile ($M = 40.22$) than in the highest ($M = 19.67$), $t(16) = 3.32$, $p = .004$, $d = 1.49$. As predicted, in the harder condition where the average percentile estimate across the lowest and highest quartiles was 44.60, miscalibration was nonsignificantly less in the lowest quartile ($M = 23.10$) than in the highest ($M = 32.80$), $t(18) = 1.20$, $p = .25$, $d = .51$.

We also looked at calibration on estimates of absolute performance (i.e., points scored), rather than relative standing. In the easier condition, there was no difference in miscalibration between the lowest and highest quartiles ($M_{\text{lowest}} = 6.78$ vs. $M_{\text{highest}} = 12.11$), $t(16) = -0.24$, $p = .812$. In the harder condition, the lowest quartile was directionally more miscalibrated than the highest ($M_{\text{lowest}} = 31.90$ vs. $M_{\text{highest}} = 4.80$), $t(18) = 1.76$, $p = .095$. It is interesting to note that, consistent with Kruger and Dunning's (1999) metacognitive hypothesis, the poorest performers on the hardest task seem to lack the skill to perceive how badly they are doing. (They received about 5 points on average and thought they were earning 37.) Nevertheless, these poor performers were not worse than the top performers at estimating the percentile of their performance.

We next examined sensitivity to relative standing, measuring the correlation between estimated and actual performance. In the sample as a whole, percentile estimates were significantly related to actual percentiles for *petroglyph, gargantuan,* and *overthrown* ($\beta_{\text{petroglyph}} = .474$, $p = .002$, $\beta_{\text{gargantuan}} = .350$, $p = .029$, $\beta_{\text{overthrown}} = .445$, $p = .007$) and not significantly related for typewriter ($\beta_{\text{typewriter}} = .216$, $p = .205$). When the same regressions were performed separately for bottom- and top-half performers, the standardized coefficients for the top-half were not significantly stronger than the standardized coefficients for the bottom-half performers on any of the four words. However, in every case, sensitivity directionally favored the top-half performers. For the easier words, $\beta_{\text{bottom}} = .270$ versus $\beta_{\text{top}} = .395$, $F(1, 33) = 0.15$, $p = .702$, on *typewriter,* and $\beta_{\text{bottom}} = .185$ versus $\beta_{\text{top}} = .316$, $F(1, 33) = 0.16$, $p = .695$, on *overthrown*. For the harder words ($\beta_{\text{bottom}} = .169$ vs. $\beta_{\text{top}} = .607$), $F(1, 37) = 2.24$, $p = .143$, on *petroglyph,* and $\beta_{\text{bottom}} = -.227$ vs. $\beta_{\text{top}} = .034$, $F(1, 37) = 0.66$, $p = .423$ on *gargantuan*.

We also examined sensitivity to absolute performance (points scored). In the sample as a whole, score estimates were significantly related to actual scores for all four words ($\beta_{\text{petroglyph}} = .603$, $p < .001$; $\beta_{\text{gargantuan}} = .373$, $p = .018$; $\beta_{\text{typewriter}} = .455$, $p = .005$; $\beta_{\text{overthrown}} = .644$, $p < .001$). On the easier words, the Chow test showed that the standardized coefficient for the bottom-half performers was marginally lower than that of the top-half performers on *overthrown* ($\beta_{\text{bottom}} = .110$ vs. $\beta_{\text{top}} = .652$), $F(1, 33) = 3.19$, $p = .083$, and not significantly different on *typewriter* ($\beta_{\text{bottom}} = .245$ vs. $\beta_{\text{top}} = .718$), $F(1, 33) = 2.62$, $p = .115$. On the harder words, the standardized coefficient for the bottom-half performers were significantly lower than for top-half performers ($\beta_{\text{bottom}} = -.111$ vs. $\beta_{\text{top}} = .739$), $F(1, 37) = 9.30$, $p = .004$, on *petroglyph,* and $\beta_{\text{bottom}} = -.292$ vs. $\beta_{\text{top}} = .555$, $F(1, 37) = 8.49$, $p = .006$, on *gargantuan*.

Split-sample tests of accuracy. The previous studies in this article and in Kruger and Dunning (1999) sorted participants on actual percentile to create quartiles and then calculated miscalibration within each quartile. As we discussed earlier, effects of regression toward the mean necessarily exaggerate the degree of

[5] Though there is no chance-level of performance that we can use to argue that this experiment meets a minimum threshold of ability, we believe that it does in fact clear this hurdle. Participants' average estimates of raw score were fairly large, especially given that they could even be negative numbers ($M_{\text{typewriter}} = 58$, $M_{\text{overthrown}} = 65$, $M_{\text{petroglyph}} = 29$, and $M_{\text{gargantuan}} = 21$). Furthermore, participants' performance on these words were positive and high ($M_{\text{typewriter}} = 57$, $M_{\text{overthrown}} = 65$, $M_{\text{petroglyph}} = 26$, and $M_{\text{gargantuan}} = 17$).

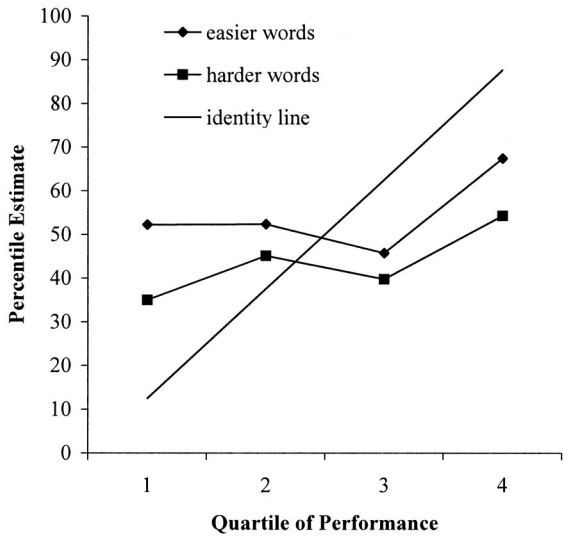

Figure 6. Participants' overall estimates of performance percentile by quartile of overall actual performance on an easier and harder word prospector task in Study 3.

miscalibration in the extreme quartiles. To avoid this problem, we divided participants according to their quartile of performance on one word and measured the difference between their estimated and actual percentiles on the other word. We again calculated miscalibration as estimated percentile minus actual percentile for those in the lowest quartile and as actual percentile minus estimated percentile for those in the highest quartile.

The results revealed 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. Highest performers were marginally better calibrated when the task was perceived as easier (i.e., average percentile estimates across lowest and highest quartiles were slightly above 50; $M_{\text{lowest}} = 28.59$ versus $M_{\text{highest}} = 6.38$), $t(16) = 1.75$, $p = .099$, $d = .79$. As expected given that the average percentile estimates were slightly below 50, the first and fourth quartiles did not differ significantly on the harder word ($M_{\text{lowest}} = 11.98$ vs. $M_{\text{highest}} = 26.10$), $t(20) = 1.38$, $p = .184$, $d = .56$. Next, 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 the bottom half of Table 3. Because the perceived percentile for both words (*typewriter* and *petroglyph*) averaged to roughly 50 across the lowest and highest quartiles (suggesting that it was moderately difficult just as in the moderate trivia condition of Study 1), there were the expected nonsignificant differences in miscalibration by quartile (ts $< .13$, ps $> .89$).

The key result is that the overall magnitude of miscalibration was substantially lower once quartiles were defined on a different task (compare the means in Table 3 with the means reported in the *Asymmetry by quartile* section). This reflects the removal of the bias induced by regression to the mean. Of course, the total amount of error across all participants on a given task is a constant.

Table 3
Perceived Percentiles, Actual Percentiles, and Miscalibration for Each Word Prospector Problem in Study 3, by Quartile of Performance on the Other Problem

		Quartile							
		Lowest				Highest			
Word and measure	Overall *M*	*M*	(*SD*)	*t*	*df*	*M*	(*SD*)	*t*	*df*
Easier word: *overthrown*[a]									
Perceived percentile	56.67	55.56	(19.11)			57.78	(25.01)		
Actual percentile		26.97	(25.84)			64.16	(34.79)		
Miscalibration		28.59		2.77*	8	6.38		0.87	8
Harder word: *gargantuan*[a]									
Perceived percentile	43.77	37.83	(15.34)			50.90	(20.80)		
Actual percentile		25.85	(23.96)			77.00	(22.30)		
Miscalibration		11.98		1.59	11	26.10		3.94**	9
Easier word: *typewriter*[b]									
Perceived percentile	52.88	40.56	(17.04)			66.75	(18.62)		
Actual percentile		30.98	(33.06)			74.85	(24.28)		
Miscalibration		9.57		0.67	8	8.10		0.83	7
Harder word: *petroglyph*[b]									
Perceived percentile	50.74	44.11	(14.37)			56.70	(28.98)		
Actual percentile		29.78	(27.86)			72.50	(28.98)		
Miscalibration		14.33		1.63	8	15.80		2.17†	9

Note. Overall *M* is the average perceived percentile across lowest and highest quartiles combined. Miscalibration = perceived percentile minus actual percentile for the lowest quartile and actual percentile minus received percentile for the highest quartile. The *t* test is a paired *t* test on actual versus perceived percentile, testing whether the miscalibration is significantly different from 0. Tests of the miscalibration measure between lowest and highest quartile are reported in the text.

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

† $p < .10$. * $p < .05$. ** $p < .01$.

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-assessment.

Discussion

It is clear that the word-prospector task allows participants to estimate how well they have done compared with others to a moderate degree. On these tasks, better performers do show greater sensitivity to where they stand relative to others. However, as in our previous studies, this does not mean that they are better calibrated. Rather, we find once again that who is miscalibrated is mostly a function of the overall bias in judgments across people. For good and bad performers alike, overall bias varies according to task difficulty such that in easier tasks the unskilled seem unaware of their percentile; in harder tasks, the skilled seem unaware.

General Discussion

People have a difficult time judging how their performance compares to the average performance of their peers. Judgments of relative standing are noisy. Accordingly, estimates of percentile 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. Judgments of relative standing are biased. People give lower estimates of their percentile 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.

In our studies, we replicated, eliminated, or reversed the association between task performance and judgment accuracy reported by Kruger and Dunning (1999) depending on 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 additional tests using the current data. However, our analyses indicate that the primary drivers of *miscalibration* in judging percentile are general inaccuracy due to noise and overall biases that arise from task difficulty. Thus, it is important to know more about sources of each class of error in order to better understand and ameliorate them.

Sources of Noise

The results of our three studies indicate that there is often a weak positive relation between objective and subjective measures of percentile. This suggests that people have limited insight into their skills and performance. 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 that we think are worthy of further study in this regard are task randomness and diagnosticity of feedback.

Task 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. Performance on any given task is subject to random variation, holding ability constant. In a testing situation, this kind of randomness stems from luck as to which particular test items are included, transient effects on performance such as distraction or fatigue, and so forth. Judges can be expected to have only incomplete awareness of such effects. Thus, estimates of performance will inevitably be regressive when conditioned on actual performance. In Study 3, we eliminated these effects by using one sample of performance to segregate low and high performers and then used an independent sample of performance to measure perceived and actual percentile. This reduces the degree to which the extreme quartiles appear biased although there is still substantial error in estimates of percentile across performance levels. This indicates that even the stable components of relative performance are hard for people to predict. The less predictable one's performance, the less performers can be expected to guess what their percentile will be. In terms of a graph such as that portrayed in Figure 1, the less predictability, the flatter the lines relating predicted to actual performance.[6] In future work, it will be useful to investigate the task and person characteristics that affect predictability and how different aspects of predictability affect judgments of percentile.

Diagnosticity of feedback. People's ability to estimate their percentile is also determined by the kind of feedback they receive and how they use it. 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,

[6] 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, tossers are 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 and so forth. 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.

If the coin tossers' blindfolds are put back on but different, biased coins are given to each tosser, the following situation is likely to occur: 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.

participants were not told whether their quiz answers were correct 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 will need to take into account differences in specific feedback conditions. The original unskilled–unaware hypothesis of Kruger and Dunning (1999) 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 Bias

Judgments of percentile are not only noisy because of random error and poor feedback but 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 provides evidence that judges anchor on their own absolute performance and adjust insufficiently for the knowledge they have of other people when they estimate their own percentile (see also Chambers & Windschitl, 2004, and Moore & Kim, 2003). Other unexplored factors are likely to bias percentile estimates, and we briefly sketch some possibilities.

First, there may be additional factors that affect perceived percentile by influencing the perceived difficulty of a task. For example, people use subjective feelings of cognitive effort as a cue to performance (Schwarz et al., 1991). We speculate that tasks for which it is easy to produce a response (e.g., a multiple-choice recognition test) are likely to lead to upwardly biased estimates of relative performance compared to tasks for which producing a response is difficult (e.g., uncued recall). However, we would also note that the association between perceived difficulty and estimated percentile may not be a "bias" at all. Instead, it may represent the fact that, in the natural ecology, absolute performance and percentile 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 (Burson & Klayman, 2005). When experimenters deliberately select tasks that are hard or easy *for everyone*, the participants naturally seem to overestimate the extent to which their success or failure is individual rather than universal.

Second, factors other than perceived task difficulty are likely to contribute to systematic bias in percentile estimates. One possibility is that people are 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 their high school 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 psychology 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, the skilled–unaware pattern we observed in our studies could be produced. (Accordingly, in our studies, we manipulated task difficulty by varying characteristics of the task, not by selecting more or less talented subpopulations.)

Finally, motivation may also play an important role in systematic misestimation of abilities, typically yielding an upward bias in percentile estimates. Self-enhancement is undoubtedly a major contributor to overestimation. When judges have a nearly costless opportunity to maintain some optimal level of optimism, most will undoubtedly take it and give higher estimates of performance than are warranted (Baumeister, 1989). On the other hand, many studies have shown that self-enhancement is less likely when people are confronted with reality constraints (Kunda, 1990) such as when performance feedback is expected to be diagnostic (Dunning, Meyerowitz, & Holzberg, 1989; Larrick, 1993) or temporally near (Gilovich, Kerr, & Medvec, 1993; Shepperd, Ouellette, & Fernandez, 1996). In other words, it is difficult to hold inflated images of oneself when that inflated image is about to be confronted with the truth. The defensive pessimism and self-handicapping literatures also suggest that some performers may strategically report or even manufacture deflated views of themselves on a task as they strive to protect general feelings of competence in a domain (Berglas & Jones, 1978; Hirt, Deppe, & Gordon, 1991; McCrea & Hirt, 2001; Norem & Cantor, 1986). Overall, self-enhancement contributes to the bias of overestimation of performance in many domains (Krueger & Mueller, 2002), and constraints on self-enhancement may reduce the bias. Even if errors in percentile estimates are only due to a noise-plus-bias process, as we (and Krueger & Mueller, 2002) have proposed, there are many interesting psychological sources of bias to incorporate into such a model.

Sensitivity Measures for the Unskilled–Unaware Hypothesis

Kruger and Dunning (1999, 2002) argued that those who are less skilled at a task are also less able to judge their relative performance, as measured by perceived percentile. This argument is based on two hypotheses: (a) 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, and (b) because of this, the bulk of the error in judging relative performance is produced by poor performers. Although we disagree with the latter interpretation, we do not reject the former.

Miscalibration, we argue, is a poor measure of accuracy because one of its two terms—perceived percentile—is subject to task-induced bias (Kruger, 1999), making it inappropriate to compare directly to actual percentile (Krueger & Mueller, 2002). We used a different measure to test for metacognitive differences that is less vulnerable to task-induced bias. We tested sensitivity by regressing perceived percentile on actual percentile among bottom-half performers and top-half performers. Results were mixed across studies and conditions, but the results generally favored top-half performers more than bottom-half performers. In this section, we

provide a small-scale meta-analysis of sensitivity, aggregating data from all of our experimental tasks.

Our studies included 12 tasks altogether (2 quizzes from Study 1, 6 subsets of easy and hard trivia from Study 2, and 4 words from Study 3). In each study, we regressed estimated percentile on actual percentile separately for bottom-half and top-half performers and obtained the standardized coefficients. We transformed those 24 coefficients using Fisher's *r*-to-*z* and then compared top-half and bottom-half performers by a paired samples *t* test with tasks as cases.[7] The average standardized coefficient was .23 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 pattern also held for the relationship between estimated and actual absolute scores in the studies that included those measures (Studies 1 and 3). The average coefficient between estimated and actual score was .45 for top-half performers and .05 for bottom-half performers, $t(5) = 2.19$, $p = .08$.

To look at these relationships using a more powerful test, we standardized perceived and actual percentile within each task and then aggregated all of the data into one large data set. This provides an analysis at the level of the individual rather than the task. In an analysis parallel to those conducted in the individual studies, we regressed perceived percentile on actual percentile among bottom-half performers and among top half performers. A Chow test confirmed at a marginal level of significance that bottom-half performers were less sensitive to their actual percentile ($\beta_{bottom} = .031$) than were top-half performers ($\beta_{top} = .206$), $F(1, 473) = 3.67$, $p = .056$. Because participants contributed several data points each in Studies 2 and 3, however, this test may overstate the appropriate degrees of freedom. We reduced the degrees of freedom to match the number of total participants ($N = 206$), and the difference in coefficients was still marginally significant, $F(1, 205) = 3.67$, $p = .057$.

We offer one final test of differences in sensitivity. When the data from all studies are pooled, and unstandardized perceived percentiles are regressed on actual percentiles, there is a significant positive relationship, $B = .188$, $SE = .037$, $\beta = .226$, $t(474) = 5.04$, $p < .001$, constant = 34.85, $R^2 = .05$. If top-half performers are more sensitive than bottom-half performers, there should also be a significant quadratic component. Indeed, regressing perceived percentile onto actual percentile and actual percentile squared showed a significantly better fit, R^2 change = .010, $F(1, 473) = 5.01$, $p = .026$. The resulting regression equation is graphed in the solid line of Figure 7. In this model (constant = 40.23), actual percentile was no longer a significant predictor of estimated percentile ($B = -.138$, $SE = .151$, $\beta = -.165$), $t(473) = -.918$, $p = .359$, but actual-percentile-squared was significant ($B = .003$, $SE = .001$, $\beta = .404$), $t(473) = 2.238$, $p = .026$). (This pattern held when dummy variables were added for task and for level of difficulty.)

Both this regression model and the earlier test of correlation asymmetries between top- and bottom-half performers are subject to a potential concern. As mean percentile estimates drop below 50, the correlation among the bottom-half performers may be depressed because of a floor effect that restricts the range of the dependent variable (percentile estimates cannot go below zero). The correlation among the top-half performers would not be similarly restricted, creating an artificial asymmetry.[8] This is an important concern that should be addressed in future research. However, we find in our studies that the asymmetry held in many instances when mean perceived percentile was at 50 or above. It is interesting to note that an opposite problem could arise on easy tasks: As mean perceived percentile rose above 50, ceiling effects would reduce the correlation among top-half performers and artificially create an asymmetry in favor of bottom-half performers. Kruger and Dunning's (1999) original data provided a test of this, because mean percentile estimates were well above 50. Nevertheless, the means shown in Figure 1 also suggest greater sensitivity among top-half performers. In fact, if perceived percentile is regressed on actual percentile and actual percentile squared using the 16 data points in Figure 1, the resulting model (shown as the dotted line in Figure 7) is almost identical to ours except, as expected, with a substantially higher constant.

Figure 7 provides a simple graphical summary of our main findings. It shows that above-average performers have better sensitivity to their relative standing than do below-average performers (cf. Figure 3). However, the metacognitive advantage suggested by greater sensitivity does not imply better calibration. The solid line in Figure 7 shows that in our tasks, in which feedback is ambiguous and the net task-induced bias is negative, it is the judgments of the better performers that deviate more from the truth. On average, participants in the 85th percentile underestimated their actual percentile by 35 percentile points; participants in the 15th percentile overestimated their actual percentile by only 25 percentile points. Ultimately, who deviates more from the truth is more a function of task-induced bias than of metacognitive advantage.

Debates on Accuracy in Other Literatures

Similar debates about the accuracy of perceptions have occurred in other literatures. We noted in the introduction that a debate in the depressive realism literature ended in the conclusion that "neither depressed nor nondepressed subjects displayed differential accuracy in terms of being able to vary their judgments to achieve accuracy across changing situations" (Dykman et al., 1989, p. 442). Instead, who appeared more accurate was an accident of the match between a dispositional bias (chronic perceptions of low or high control) and the degree of control actually available in a given task. In our studies, accuracy was driven by stable task-related biases rather than by individual differences in bias. Both cases share, however, the caveat that straightforward comparisons between perceptions and observed outcomes can be misleading with regard to apparent cognitive differences between "accurate" and "inaccurate" performers.

It is interesting also to compare our findings to those from research on overconfidence, which compares subjective confi-

[7] Strictly speaking, these tasks should not all be regarded as independent cases given that the six tasks of Study 2 were presented within participants. However, we checked the extent to which different sets of estimates provided independent tests of relative ability. Recall that each participant received six sets of questions, one in each combination of domain and difficulty. We compared participants' quartiles on each set of questions in the study to their quartiles on each of their other sets. The sets proved to be largely independent in terms of relative performance. Correlations between pairs of sets ranged from −.39 to .31, with a median of .02.

[8] We thank Joachim Krueger for raising this concern.

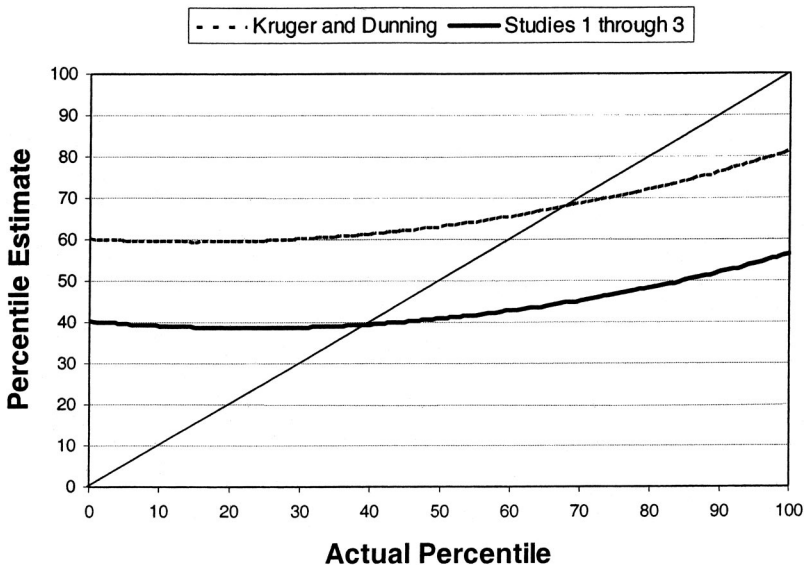

Figure 7. Plot of regression equation predicting perceived percentile from actual percentile, with data from Studies 1–3 (solid line) and Kruger and Dunning's (1999) means (dashed line).

dence estimates with absolute performance (e.g., the probability of answers being correct). There are some clear parallels. In both cases, the correlation between actual and predicted performance is only weakly positive, and in both the effects of noise in judgments can look like systematic biases. Consider the "hard–easy effect" in overconfidence. When questions are sorted according to how frequently they are answered correctly and then compared directly to noisy subjective estimates of confidence, a systematic pattern is observed: Questions that are rarely answered correctly show overconfidence, but those that are frequently answered correctly show underconfidence. Initially, psychological explanations were offered for this pattern. Eventually, however, it was argued that given the weak relationship between confidence and accuracy, regression effects are sufficient to produce this pattern (Erev, Wallsten, & Budescu, 1994; Juslin, 1993; Juslin, Winman, & Olsson, 2000). This argument directly parallels the one we have offered here, but with a different unit of analysis. With the hard–easy effect in overconfidence, the unit of analysis is a question; in tests of the unskilled–unaware hypothesis, the unit of analysis is a person. When people are sorted according to an objective measure of relative performance and then compared directly to noisy subjective estimates, a systematic pattern is observed: Poor performance on the objective measure is associated with overestimation (e.g., the "unskilled" will overestimate their percentile) and superior performance will be associated with underestimation (the "skilled" will underestimate their percentile). The implication in both literatures is that the effects of noise in judgment must be carefully accounted for before drawing conclusions about systematic cognitive biases. (See also our earlier discussion of the reasons for task-induced biases.)

There is another interesting connection between the current work and the overconfidence literature. Researchers often casually speak of overconfidence and above-average effects as demonstrating similar optimistic biases. In this view, it might be surprising that hard–easy effects in overconfidence seem to produce a different pattern of results than difficulty-induced biases in percentile estimates. For example, hard tasks produce overconfidence but worse-than-average perceptions. This seeming paradox is reconciled when it is recognized that performance on hard tasks, by definition, is quite poor such that even low average confidence will often exceed average performance levels. However, low perceived percentiles on hard tasks must fall short of true percentiles (which necessarily average to 50). Larrick, Burson, and Soll (2005) and Moore and Small (2005) have demonstrated that increases in task difficulty do in fact increase overconfidence while simultaneously reducing better-than-average effects.

Alternative Explanations for the Skilled–Unaware Pattern

We have proposed that a noise-plus-bias model explains both the unskilled–unaware pattern of miscalibration on easy tasks and the skilled–unaware pattern of miscalibration on hard tasks. Yet, might there be other reasons that the unskilled–unaware pattern reverses as tasks are perceived to be more difficult? Kruger and Dunning (1999) proposed that the unskilled–unaware pattern would only hold when a task permits the perception of a minimum threshold of competence, noting that on very difficult tasks, people would recognize their lack of ability and provide low percentile estimates. This explanation can account for the lower estimates and improved calibration of the worst performers on difficult tasks. However, it does not address why the best performers would dramatically underestimate their percentile on difficult tasks and thereby exhibit greater miscalibration. Other mechanisms must be added to the metacognition account to explain the pessimism of the best performers on difficult tasks. We extend Kruger and Dunning's original proposal by sketching some conjectures along these lines.

Suppose that on easy tasks, there is the expected metacognitive difference. The worst performers do poorly but do not know it and overestimate their standing. The best performers do well, know it

(almost), and only slightly underestimate their standing. Suppose, however, that on difficult tasks, a different process applies. One possibility is that the best performers have succeeded purely by good luck on hidden variables (such as good luck with guesses). They naturally will not know they have been lucky, and thus the best (luckiest) performers will underestimate their achievement. However, this explanation applies only to tasks in which luck alone determines who does well and who does poorly and in which the nature of that luck is opaque to participants. For tasks in which there are real skill differences, additional mechanisms must be postulated to explain why highly skilled judges would underrate themselves.

One possibility is that the best performers expect that their relatively good performance may not be repeatable and therefore give lower percentile estimates to match what they think they might achieve over several trials. Perhaps the best performers fear that their future performance will be worse and strategically give conservative percentile estimates for their current high performance in order to reduce future expectations (as a form of self-handicapping; McCrea & Hirt, 2001) or to motivate harder work in the future (Norem & Chang, 2002). In addition, high performers may learn through social interaction that it is generally best to be modest about their superior performance on difficult tasks (Tice, Butler, Muraven, & Stillwell, 1995).

Each of these additional mechanisms is intriguing, and it would be interesting to test whether they contribute to the miscalibration of the best performers on difficult tasks. We believe that, in the absence of direct evidence for additional mechanisms, a noise-plus-bias model is a more parsimonious account of the current results. In this model, one simple psychological mechanism—task-induced bias (Kruger, 1999)—is sufficient to explain both the underestimation of the best performers on difficult tasks and the overestimation of the worst performers on easy tasks.

Conclusions

It is a well-established and entertaining fact 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, in which direction, and when. Kruger and Dunning (1999, 2002) suggested that there is a relationship between task performance, metacognition, and judgmental accuracy. They proposed that the bulk of miscalibration in judging relative performance comes from poor performers' tendency to overestimate their abilities, which, in turn, is due to their poorer metacognitive skills. In fact, these researchers have found that poor performers substantially overestimate their percentile and better performers only slightly underestimate theirs. Our studies show, however, that poor performers are more miscalibrated than good performers only on tasks that feel easy and that the reverse pattern occurs on tasks that feel difficult.

These results indicate that the answer to the question, "Who makes errors in judging relative performance?" is, more or less, "Everyone." On the kinds of tasks that have been studied to date, the skilled and the unskilled are similarly limited in judging how their performance compares with that of others. The answer to the question, "In which direction do the miscalibrations occur?" depends on task difficulty, as Kruger (1999) has also shown. In tasks that seem easy, on average people think they are above average; on tasks that seem difficult, people generally think they are below average. This then leads to the answer to the question "When do such miscalibrations occur?" Ultimately, who appears accurate is an accident of the match between task-induced bias and actual percentile. When the task seems hard, poor performers seem perceptive, and the best performers underestimate their standing. When the task seems easy, good performers seem perceptive, and those near the bottom overestimate their standing. We propose that a noise-plus-bias account is a parsimonious explanation for this pattern of miscalibration (in line with Krueger & Mueller, 2002).

At the same time, other measures of accuracy do support the existence of metacognitive differences related to level of task skill. In our studies, though top performers were more miscalibrated than were poor performers, they also tended to be more sensitive to their relative standing than were poor performers. This illustrates our basic argument: We do not claim that there is no relation between cognitive skill and metacognitive skill but rather that such a relationship is not a primary determinant of who is miscalibrated. Better performers may be 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 percentile. Because sensitivity is less subject to task-induced bias, we propose that it is a more appropriate measure for testing metacognitive differences in the future.

Judgments of relative performance play an important role in decisions about engaging in competitive activities, purchasing goods and services, and undertaking challenging tasks (Burson, 2005; Moore & Kim, 2003; Simonsohn, 2003). Overestimates of relative performance can lead to frustration, loss, and even physical harm. (Consider, for example, mediocre horseback riders or skiers who attempt advanced trails.) On the other hand, there are also significant domains in life where relative performance may be underestimated and people fail to participate when they would have succeeded (Camerer & Lovallo, 1999; Moore & Kim, 2003). The research presented in this article provides a foundation for further exploration of how and how well people know where they are on the curve, and how people can be helped to assess their place better.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33,* 587–605.

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.

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108,* 441–485.

Baumeister, R. F. (1989). The optimal margin of illusion. *Journal of Social and Clinical Psychology, 8,* 176–189.

Berglas, S., & Jones, E. E. (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology, 36,* 405–417.

Burson, K. A. (2005). *Consumer-product skill matching: The effects of difficulty on relative self-assessment and choice. Unpublished manuscript, University of Michigan.*

Burson, K. A., & Klayman, J. (2005). *Judgments of performance: The relative, the absolute, and the in-between.* Unpublished manuscript, University of Michigan.

Camerer, C. F., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review, 89,* 306–318.

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin, 130,* 813–838.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica, 28,* 591–605.

Coyne, C. C., & Gotlib, I. H. (1983). The role of cognition in depression: A critical appraisal. *Psychological Bulletin, 94,* 472–505.

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.

Dykman, B. M., Abramson, L. Y., Alloy, L. B., & Hartlage, S. (1989). Processing of ambiguous and unambiguous feedback by depressed and nondepressed college students: Schematic biases and their implications for depressive realism. *Journal of Personality and Social Psychology, 56,* 431–445.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and under-confidence: The role of error in judgment processes. *Psychological Review, 101,* 519–528.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

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

Hirt, E. R., Deppe, R. K., & Gordon, L. J. (1991). Self-reported versus behavioral self-handicapping: Empirical evidence for a theoretical distinction. *Journal of Personality and Social Psychology, 61,* 981–991.

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.

Juslin, P., Winman, A., & Olsson, H. (2000). Naïve empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review, 107,* 384–396.

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.

Larrick, R. P., Burson, K. A., & Soll, J. B. (2005). *Social comparison and confidence: When thinking you're better than average predicts overconfidence.* Unpublished manuscript, Duke University.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Martin, D. J., Abramson, L. Y., & Alloy, L. B. (1984). The illusion of control for self and others in depressed and nondepressed college students. *Journal of Personality and Social Psychology, 46,* 125–136.

McCrea, S. M., & Hirt, E. R. (2001). The role of ability judgments in self-handicapping. *Personality and Social Psychology Bulletin, 27,* 1378–1389.

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

Moore, D. A., & Small, D. A. (2005). *Error and bias in comparative social judgment: On being both better and worse than we think we are.* Unpublished manuscript, Carnegie-Mellon University.

Norem, J. K., & Cantor, N. (1986). Anticipatory and post hoc cushioning strategies: Optimism and defensive pessimism in "risky" situations. *Cognitive Therapy and Research, 10,* 347–362.

Norem, J. K., & Chang, E. C. (2002). The positive psychology of negative thinking. *Journal of Clinical Psychology, 58,* 993–1001.

Pindyck, R. S., & Rubinfeld, D. L. (1998). *Econometric models and economic forecasts* (4th ed.). Boston: Irwin McGraw-Hill.

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). *Whether to go to college.* Paper presented at the annual conference of the Society for Judgment and Decision Making, Vancouver, British Columbia, Canada.

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

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 299–314.

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

Tice, D. M., Butler, J. L., Muraven, M. B., & Stillwell, A. M. (1995). When modesty prevails: Differential favorability of self-presentation to friends and strangers. *Journal of Personality and Social Psychology, 69,* 1120–1138.

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

Simulations of Judge Accuracy as a Function of Relative Ability and Task Difficulty

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 1,000 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 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:

$$S_j = A_j + e_j. \quad (1)$$

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 a 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 A_j.

Estimated Performance

Each participant estimates his or her performance on the basis of his or her ability plus some error and possibly some overall bias:

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

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 total 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' performance. 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, following 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 1,000 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 1,000 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 with 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 the following:

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

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 highest percentile of ability, an average amount for those of median ability, and double the average error and bias for those in the lowest 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, following 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 exploratory only. With different parameter values, the differences between the results of Equation 2 and Equation 3 can be less clearly distinct, and other models of the unskilled-unaware hypothesis are also plausible.

Received September 18, 2002
Revision received May 24, 2005
Accepted May 24, 2005 ■