Mem Cogn (2014) 42:164–173
DOI 10.3758/s13421-013-0351-4

The contribution of judgment scale to the unskilled-and-unaware phenomenon: How evaluating others can exaggerate over- (and under-) confidence

Marissa K. Hartwig · John Dunlosky

Published online: 12 July 2013

Abstract The unskilled-and-unaware phenomenon occurs when low performers tend to overestimate their performance on a task, whereas high performers judge their performance more accurately (and sometimes underestimate it). In previous research, this phenomenon has been observed for a variety of cognitive tasks and judgment scales. However, the role of judgment scale in producing the unskilled-and-unaware phenomenon has not been systematically investigated. Thus, we present four studies in which all participants judged their performance on both a relative scale (percentile rank) and an absolute scale (number correct). The studies included a variety of performance tasks (general knowledge questions, math problems, introductory psychology questions, and logic questions) and test formats (multiple-choice, recall). Across all tasks and formats, the percentile-rank judgments were less accurate than the absolute judgments, particularly for low and high performers. Furthermore, in Studies 1–3, the absolute judgments were highly accurate, even when the percentile-rank judgments were not. Thus, differences in the accuracy of percentile-rank judgments across skill levels do not always represent differences in *self*-awareness, but rather they may arise from difficulties that performers have at evaluating how well *others* are performing. Most importantly, the unskilled-and-unaware phenomenon on a relative scale does not guarantee inaccurate self-evaluations of absolute performance.

Keywords Unskilled and unaware · Judgment scale · Judgment accuracy

After completing a cognitive task—such as a standardized college admissions test (e.g., SAT)—students may attempt to estimate or judge how well they performed. For example, a student who felt confident in the responses that he or she selected may predict a high score or high percentile ranking. But how accurate are students' judgments of their performance? The answer to this question partly depends on whether the student is a low or a high performer. Low performers tend to overestimate their global (overall) performance on a cognitive task, whereas high performers tend to be more accurate, or even to underestimate their performance. Because low performers both (1) perform poorly and (2) do not recognize how poorly they have performed, they have been described as "unskilled and unaware" and "doubly cursed" (Kruger & Dunning, 1999). According to Kruger and Dunning (1999), the lack of knowledge (or skill) that produces their poor performance also makes them unable to recognize when their responses are incorrect—thereby causing their overestimation.

The unskilled-and-unaware phenomenon has been observed in numerous studies using a variety of cognitive tasks, including tests of trivia (Burson, Larrick, & Klayman, 2006), logical reasoning (Kruger & Dunning, 1999), grammar (Kruger & Dunning, 1999), gun safety (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008), humor (Kruger & Dunning, 1999), psychology course materials (Dunning, Johnson, Ehrlinger, & Kruger, 2003), and debate (Ehrlinger et al., 2008; Kruger & Dunning, 1999), among others (e.g., Burson et al., 2006; Hodges, Regehr, & Martin, 2001; Mattern, Burrus, & Shaw, 2010). In these studies, participants are asked to judge their performance relative to others (e.g., on a scale of percentile rank). Some studies have also demonstrated the phenomenon with scales of absolute performance, on which participants judge the number (or percentage) of items answered correctly (Dunning et al., 2003; Ehrlinger et al., 2008; Kruger & Dunning, 1999; Miller & Geraci, 2011). Thus, it seems that unskilled performers are overconfident on some tasks, regardless of whether performance is judged on a relative or absolute scale. However,

M. K. Hartwig · J. Dunlosky (✉)
Psychology Department, Kent State University,
Kent, OH 44242, USA
e-mail: jdunlosk@kent.edu

whether the type of judgment scale moderates the size of the unskilled-and-unaware phenomenon has not been systematically investigated. That is, might performers' ability to judge their performance partly depend on the judgment scale being used? In the present studies, we answered this question across several different cognitive tasks, and the results have important implications for the scope of this phenomenon, as well as for possible causes of it.

In the remainder of the introduction, we provide a brief review of previous studies that have demonstrated the unskilled-and-unaware phenomenon, focusing on what judgment scales have been used. We then consider reasons to expect that either (1) the judgment scale would not influence the size of the unskilled-and-unaware phenomenon, or (2) the judgment scale would moderate the size of the effect. Finally, we describe the present studies, which evaluate these alternative hypotheses.

The unskilled-and-unaware phenomenon: Are performers' judgments relative or absolute?

A variety of studies have demonstrated overestimation by low performers and slight underestimation by high performers—that is, the unskilled-and-unaware pattern. Across these studies, participants have judged their performance on various scales, including scales of relative performance such as percentile rank (Burson et al., 2006; Dunning et al., 2003; Ehrlinger et al., 2008; Hodges et al., 2001; Kruger & Dunning, 1999), scales of absolute performance such as number correct or percent correct (Dunning et al., 2003; Ehrlinger et al., 2008; Kruger & Dunning, 1999; Miller & Geraci, 2011), or more complex scales such as points earned according to some system of scoring (Ehrlinger et al., 2008; Kruger & Dunning, 1999). Thus, the pattern has been observed using several types of judgment scales, including both relative and absolute scales. Regardless of which judgment scale was used, overestimation by low performers is usually taken as evidence that low performers are unable (or less able) to recognize their mistakes (e.g., Kruger & Dunning, 1999; but see Burson et al., 2006; Krajc & Ortmann, 2008; and Krueger & Mueller, 2002, for alternative hypotheses).

However, whether low performers' overestimation on each judgment scale represents the same metacognitive phenomenon has not been established. If the unskilled-and-unaware pattern represents the same phenomenon regardless of which judgment scale is used, the pattern should be observed consistently when more than one scale type is used in the same study. Only three studies have allowed for such a comparison (Dunning et al., 2003; Ehrlinger et al., 2008; Kruger & Dunning, 1999), but none of them evaluated differences in the pattern on the basis of scale type. Kruger and Dunning (1999) asked participants to judge their performance on both percentile-rank and number-correct scales for tests of logical reasoning (Studies 2 and 4) and grammar (Study 3). They concluded that both scales produced the same qualitative pattern in which low performers were overconfident. However, not all of their studies included absolute scales to allow for this comparison, nor were absolute judgments reported for each quartile of performers. For those that were reported, we note that the magnitude of judgment inaccuracy was typically diminished on the absolute scale in comparison to the relative scale (e.g., in Study 3, low performers were approximately 19% overconfident on the number-correct scale vs. 57% overconfident on the percentile-rank scale), suggesting that meaningful differences in judgment accuracy could exist between relative and absolute scales. One possibility is that the patterns may be qualitatively consistent yet may differ substantially in magnitude, depending on the judgment scale. This possibility has not been systematically investigated.

In summary, the role played by judgment scale in producing the unskilled-and-unaware pattern remains unclear. Further research will be required to evaluate whether the pattern is scale-invariant (i.e., observed consistently when both relative and absolute scales are used) or scale-dependent (i.e., the pattern may differ substantially depending on the type of scale used).[1] The present studies will test these opposing hypotheses, but first we will highlight the reasons to expect each outcome.

Reasons to expect scale invariance or scale dependence

One possibility is that both relative and absolute scales would produce the same basic pattern of judgment accuracy, providing evidence that the unskilled-and-unaware phenomenon is scale-invariant. Scale invariance would be consistent with Kruger and Dunning's (1999) hypothesis regarding the basis of the phenomenon—that is, that low performers (who lack knowledge or skill, and thus perform poorly) are less able to recognize when test items have been answered incorrectly, which produces unawareness of their overall performance. Most importantly, this hypothesis assumes that a major source of low performers' unawareness is their inability to judge their *absolute* performance. If so, the phenomenon should occur consistently, and to similar degrees, for both number-correct judgments (i.e., judging one's absolute score, regardless of how anyone else scored) and percentile-rank judgments (i.e., judging how one's absolute score compares to others' scores). Such an observation would provide support for the hypothesis that unskilled participants' overconfidence largely arises from a lack of self-awareness.

[1] The terms *scale-dependent* and *scale-invariant* are used here simply to represent the alternative possibilities that the unskilled-and-unaware pattern may or may not depend on the kind of judgment scale being used, and they are not meant to refer to issues pertaining to scale dependence from measurement theory.

Alternatively, the number-correct judgments might not produce the same pattern as the percentile-rank judgments. This idea—that different judgment scales may elicit different patterns of results—has some intuitive appeal. For example, one mechanism that could contribute to poorer accuracy of percentile-rank judgments (as compared to number-correct judgments) concerns what information people must retrieve or construct to make a given judgment. In particular, number-correct judgments primarily require knowledge of one's own performance, regardless of anyone else's. In contrast, percentile-rank judgments also require knowledge about the performance of the reference group, as well as about the variability in scores for the group. If judgment inaccuracy is caused primarily by unawareness of the reference group's performance (or variability in scores), or from a failure to adequately incorporate knowledge about the reference group (and variability) into one's judgment, then the unskilled-and-unaware phenomenon should be larger for percentile-rank judgments than for number-correct judgments. Such an observation would support the hypothesis that inaccuracy among percentile-rank judgments reflects difficulty in judging the performance of the comparison group (either with respect to judging others' mean performance or the variability in their performance, both of which would be needed to make accurate percentile-rank judgments).

With respect to the unskilled-and-unaware phenomenon, difficulty judging the comparison group has previously been proposed to explain the underestimates of high performers (Kruger & Dunning, 1999), but it has not been proposed as an explanation for the overestimates of low performers. Nevertheless, plenty of prior research has indicated that people in general have difficulty making accurate judgments of relative performance (e.g., Burson et al., 2006; Kruger, 1999; Moore & Cain, 2007; Moore & Small, 2007). In these studies, whether people tended to overestimate or underestimate their relative performance depended on factors including actual or perceived task difficulty. However, the mechanism involved (i.e., difficulty judging the reference group's scores, in terms of mean and variability) pertains to *both* low and high performers (as well as to performers in between). If this mechanism is a key contributor to the unskilled-and-unaware pattern, the unawareness observed in low performers may not be unique to them. Furthermore, both low and high performers might demonstrate number-correct judgments that are substantially more accurate than their percentile-rank judgments—that is, scale dependence. If the scale-dependence hypothesis is supported, it would indicate that the overconfidence of low performers' judgments (or the underconfidence of high performers) when using a percentile-rank scale does not arise entirely from their lack of self-awareness, but instead from their lacking awareness of how well (and how variable) the reference group is performing.

Overview of the present studies

A major goal of the present studies was to evaluate whether the unskilled-and-unaware phenomenon is scale-invariant or scale-dependent. To do so, we performed four studies in which participants made global performance judgments for several different cognitive tasks. In Studies 1–3, the tasks included tests of various contents (general knowledge questions, math problems, and introductory psychology questions) and formats (multiple-choice tests and recall tests). Most importantly, in each study, participants judged their performance on both relative (percentile-rank) and absolute (number-correct and percent-correct) scales. To foreshadow the results, the difference between relative and absolute scales was quite dramatic in Studies 1–3, and unexpectedly, the absolute scales showed no evidence of the unskilled-and-unaware pattern. Because the unskilled-and-unaware pattern has been observed using absolute scales in some previous studies, in Study 4 we selected another content area (logical reasoning) that has previously demonstrated the unskilled-and-unaware pattern (Ehrlinger et al., 2008; Kruger & Dunning, 1999). This new test in Study 4 did show the standard pattern on an absolute scale, and thus, we were able to investigate whether the phenomenon is scale-dependent in this context as well.

Studies 1–3

The studies reported here differed primarily in the cognitive tasks on which participants were tested. Because the methods were otherwise nearly identical, we present the three studies together for brevity.

Method

The cognitive tasks included general-knowledge trivia questions in Study 1, algebra-based math problems in Study 2, and introductory psychology questions in Study 3. Studies 1 and 2 included two test formats—multiple-choice and recall (response generation)—that were administered to different sets of participants, whereas Study 3 included only the multiple-choice format.

Participants and design

Undergraduate students at a Midwestern university participated for course credit. They were part of a participant pool, which consisted mainly of students enrolled in introductory psychology or other lower-division psychology courses. In Study 1, participants were randomly assigned to either the trivia multiple-choice group ($n = 81$) or the trivia recall group ($n = 80$). In Study 2, participants were randomly assigned to either the math multiple-choice group ($n = 102$) or the math response-generation group ($n = 94$). Finally, Study 3

(psychology multiple-choice, n = 103) included only one group. The studies were conducted in the laboratory, and the samples did not share any participants.

Materials

In Study 1, the materials were 40 general-knowledge questions (e.g., "What kind of metal is associated with a 50th wedding anniversary?") from Nelson and Narens's (1980) norms. Each question could be answered with a one-word response (e.g., "gold"). The 40 questions were selected to represent a range of difficulty. In Study 2, the materials were 20 brief, algebra-based story problems and computational problems drawn from the math section of an SAT study guide. In Study 3, the materials were 40 general psychology questions taken from introductory psychology materials. Because participants did not study for this psychology test, the questions were selected to be easier than a typical in-class exam would be, to keep performance levels above the performance floor. Examples of the questions used in each study are provided in Table 3.

Procedure

The groups in the three studies were all designed to be analogous, differing only in their test materials and format. The participants worked individually at computers. They first read instructions on the computer screen that described what they would be asked to do. Next, all participants took the relevant test, followed by a global judgment phase.

Test phase After participants had read the instructions, 40 trivia questions (Study 1), 20 math questions (Study 2), or 40 psychology questions (Study 3) were presented, one at a time, in random order. For the multiple-choice groups, each question was presented with four choices (trivia and math) or five choices (psychology; the order of presentation was randomized anew for each participant for all studies). One choice was the correct answer, and the other choices were plausible but incorrect responses. For the recall/answer-generation groups, participants were not given choices, but instead were required to type a one-word response for each trivia question (Study 1) or a numerical response for each math question (Study 2). Even if participants could not remember the correct answer or were unable to solve the math problem, they were required to make a guess. After each response, participants rated their confidence in the response that they had selected or generated (from 0 = *not confident at all* to 10 = *completely confident* that their response was correct); these judgments were not relevant to our present aims and so will not be discussed further, except to note that the presence versus absence of these judgments did not influence the accuracy of the global judgments of task performance. For instance, Study 3 originally included two groups: one group in which participants rated their confidence in their responses to each test question, and one that was identical, except that participants did not make these item-by-item confidence judgments. The outcomes for these groups did not differ, so they were collapsed into one. Testing was self-paced.

Global judgments At the conclusion of the test phase, participants were asked to make global judgments about their performance on the test as a whole. First, they judged the *number* of items (out of 40 total items in Studies 1 and 3, or out of 20 total items in Study 2) that they had answered correctly. Second, they judged the *percentage* of items that they had answered correctly on the test (from 0% to 100%). These two global judgments produced similar qualitative and quantitative patterns of judgment accuracy in all of the analyses below; thus, only judgments of number correct will be presented when reporting absolute-scale judgments. Finally, participants were asked to judge their own percentile rank relative to other undergraduates who had taken the same test. The instructions read as follows:

For the [40, 20][trivia, math, psychology] questions you were tested on, what do you think your percentile rank would be when comparing your performance to the performance of other students who took this test? In other words, please estimate the percentile rank of your performance by typing any number from 1 to 99. Examples:

- A percentile rank of 99 would indicate that you think you performed better than 99% of all students who took this test.
- A percentile rank of 50 would indicate that you think you performed better than 50% of all students who took this test.
- A percentile rank of 1 would indicate that you think you performed better than only 1% of all students who took this test.

And so forth.

Note that we chose to have participants make absolute-scale judgments before percentile-rank judgments for two reasons: (1) Prior research had demonstrated that the order in which these judgments were made did not influence their accuracy (Kruger & Dunning, 1999), and (2) given that the scale-dependent hypothesis predicts that the number-correct judgments may be more accurate than percentile-rank judgments, any (unexpected) reactive effects would work against this main hypothesis.

Results and discussion

In each study presented here, our goal was to evaluate the presence and magnitude of the unskilled-and-unaware phenomenon when participants judged their performance on relative (percentile-rank) versus absolute (number-correct) scales. To do so, we first computed the accuracy of each individual's percentile-rank and number-correct judgments by subtracting their actual scores from their judged scores. Positive difference scores indicated overestimation, whereas negative difference scores indicated underestimation. The

means of these difference scores are shown in Table 1 for each quartile of performer. In the analyses that follow, these difference scores will be used to evaluate the patterns of judgment accuracy for both relative and absolute scales.

When participants judged their performance on a percentile-rank scale, low performers overestimated their scores and high performers underestimated their scores, in all studies and for all groups within these studies (Table 1). (For interested readers, we report the overall mean levels of judgments and performance in Table 2.) Furthermore, the difference scores of low and high performers differed significantly from each other (ps < .01 for all groups). Thus, for percentile-rank scales, the classic unskilled-and-unaware pattern was evident. Absolute scales, however, did not produce the same pattern: Specifically, when participants judged their performance on a number-correct scale, low performers did not demonstrate any greater overestimation than did high performers (all ps > .05, except in the recall group of Study 1, which unexpectedly demonstrated the opposite of the unskilled-and-unaware pattern). Furthermore, the mean difference scores among number-correct judgments rarely differed from zero (Table 1), except in the response-generation group of Study 2 (math), which showed a tendency for overestimation among all quartiles. In contrast, the mean difference scores among percentile-rank judgments often differed from zero, particularly for low and high performers.

Table 1 Means of judgments (estimated score), actual scores, and differences between judged and actual scores for each quartile of performers in all studies

			Percentile-Rank Scale			Number-Correct Scale		
Study	Group	Quartile	Estimate	Actual	Diff.	Estimate	Actual	Diff.
1	Trivia Multiple-Choice	1 (low)	29.3	13.0	16.3*	15.7	17.7	−2.0
		2	44.1	40.1	4.0	22.7	21.2	1.6
		3	44.6	64.2	−19.6*	25.6	23.9	1.7
		4 (high)	67.2	87.0	−19.8*	28.4	28.5	−0.1
			High vs. low: $t(40) = 5.7$, $p < .001$**			High vs. low: $t(40) = 0.8$, $p = .43$		
	Trivia Recall	1 (low)	14.6	11.3	3.3	4.7	5.2	−0.5
		2	17.2	35.0	−17.8*	8.8	10.4	−1.6
		3	34.3	59.4	−25.1*	12.9	15.4	−2.5
		4 (high)	52.8	85.6	−32.8*	22.6	20.0	2.6*
			High vs. low: $t(39) = 6.7$, $p < .001$**			High vs. low: $t(39) = 2.1$, $p = .04$		
2	Math Multiple-Choice	1 (low)	20.9	11.3	9.6	7.0	5.3	1.7
		2	38.0	34.3	3.7	8.2	8.5	−0.3
		3	48.4	57.4	−9.0	12.4	11.6	0.8
		4 (high)	73.3	84.3	−11.0*	15.8	15.6	0.2
			High vs. low: $t(53) = 4.0$, $p < .001$**			High vs. low: $t(53) = 1.7$, $p = .09$		
	Math Response-Generation	1 (low)	21.5	13.8	7.6	5.9	3.5	2.4*
		2	36.9	40.4	−3.6	9.5	7.2	2.3*
		3	47.0	66.0	−18.9*	11.3	10.0	1.3
		4 (high)	74.6	89.4	−14.8*	15.6	14.1	1.5*
			High vs. low: $t(44) = 3.4$, $p = .002$**			High vs. low: $t(44) = 0.8$, $p = .41$		
3	Psychology Multiple-Choice	1 (low)	29.9	11.2	18.7*	13.7	11.4	2.3
		2	32.0	34.0	−2.0	15.8	16.1	−0.3
		3	46.0	59.7	−13.7*	19.9	18.9	0.9
		4 (high)	53.0	86.9	−33.9*	23.8	23.6	0.3
			High vs. low: $t(48) = 8.9$, $p < .001$**			High vs. low: $t(48) = 1.1$, $p = .28$		
4	Logic Multiple-Choice	1 (low)	50.2	13.6	36.7*	10.8	5.8	5.0*
		2	50.4	40.8	9.6*	10.7	8.7	2.0*
		3	56.1	64.6	−8.5	12.8	11.7	1.1
		4 (high)	62.2	87.5	−25.3*	14.4	15.0	−0.5
			High vs. low: $t(46) = 11.4$, $p < .001$**			High vs. low: $t(46) = 6.1$, $p < .001$**		

* Mean difference score differed from zero, $p < .05$. ** Mean difference scores of high versus low performers differed in the direction that is consistent with the unskilled-and-unaware phenomenon.

Table 2 Mean judgments (estimated score) and actual scores, averaged across quartiles of performance in all studies

		Percentile-Rank Scale		Number-Correct Scale	
Study	Group	Estimate	Actual	Estimate	Actual
1	Trivia multiple-choice	46.3	50.0	23.0 (57.5%)	22.7 (56.8%)
	Trivia recall	30.9	50.0	12.8 (32.0%)	13.2 (33.0%)
2	Math multiple-choice	47.6	50.0	11.2 (56.0%)	10.7 (53.5%)
	Math response-generation	43.2	50.0	10.3 (51.5%)	8.4 (42.0%)
3	Psychology multiple-choice	41.0	50.0	18.6 (46.5%)	17.8 (44.5%)
4	Logic multiple-choice	54.5	50.0	12.1 (60.5%)	10.1 (50.5%)

Study 4

In Studies 1–3, participants' absolute judgments were highly accurate overall. This result was surprising, given prior demonstrations of the unskilled-and-unaware pattern with absolute scales, but it nevertheless indicates that the pattern is not inevitable. We will consider possible reasons for these disparate results in the General Discussion. In Study 4, however, we sought materials that would produce the unskilled-and-unaware pattern on an absolute scale, to provide a further test of scale dependence. Thus, in Study 4, we selected a type of test material—logical reasoning—that has produced the unskilled-and-unaware pattern with absolute judgments in previous research (Ehrlinger et al., 2008; Kruger & Dunning, 1999). To foreshadow, these materials did produce the unskilled-and-unaware pattern for absolute judgments, so our focal question was whether percentile-rank judgments would show the unskilled-and-unaware phenomenon to either the same or a greater degree than judgments on an absolute scale.

Also, in Studies 1–3, we did not counterbalance the order of judgment types; that is, number-correct judgments always preceded percentile-rank judgments. Order was not a concern, because prior research had demonstrated that the order of the two judgments does not influence their accuracy (Kruger & Dunning, 1999), and furthermore, any carryover effects would be expected to make judgments on the second scale more similar to those on the first, thereby working against the scale-dependent discrepancies that we observed. Also note that the unexpected lack of the unskilled-and-unaware pattern in Studies 1–3 occurred for the first judgment, which was not affected by the subsequent judgment, whereas the second judgment demonstrated the typical unskilled-and-unaware pattern. Nevertheless, in Study 4, we counterbalanced the order of number-correct and percentile-rank judgments to verify that order did not influence the outcomes.

Method

The participants (N = 92) were undergraduate students at a Midwestern university who participated for course credit via a participant pool. All participants received a multiple-choice test of logical reasoning. The logical reasoning questions and response choices were drawn from an LSAT (Law School Admission Test) study guide. The procedure (including a testing phase, followed by global judgments) was identical to those of the previous studies, except that participants made only two global judgments—number correct and percentile rank—and the order of these global judgments was counterbalanced. Approximately half of the participants (n = 48) judged their absolute performance before judging their relative performance, whereas the other participants (n = 44) judged relative performance before judging absolute performance.

Results and discussion

As in the previous studies, we first computed the accuracy of each individual's percentile-rank and number-correct judgments by subtracting their actual scores from their judged scores. We then evaluated whether the order of judgments influenced the patterns of accuracy observed for each judgment scale. For both judgment scales, order did not affect the size of the difference scores in any quartile (all ps > .05); for difference scores as a function of order, see Table 4 (Appendix). Given the lack of an order effect, the main analyses presented next involved collapsing across orders.

The means of these difference scores are shown at the bottom of Table 1. Unlike in the previous three studies, the number-correct scale did produce an unskilled-and-unaware pattern: Specifically, low performers overestimated their performance (p < .05), whereas high performers estimated accurately. Thus, the logic test produced inaccurate absolute judgments, whereas the previous tests (in Studies 1–3) did not. Possible reasons for this discrepancy will be considered in the General Discussion.

When participants judged their performance on a percentile-rank scale, low performers overestimated their scores and high performers underestimated their scores, consistent with Studies 1–3. Furthermore, the difference scores of low and high performers differed significantly from each other (p < .001). Thus, for percentile-rank scales, the classic unskilled-and-unaware pattern was again evident.

Although both judgment scales produced an unskilled-and-unaware pattern in Study 4, the magnitudes of the pattern might differ, which would be consistent with the notion of scale dependence demonstrated in Studies 1–3. To evaluate this possibility, we converted the difference scores in Table 1 to percentages, thereby allowing the two judgment scales to be compared in terms of the observed magnitude of inaccuracy. This comparison indicated that the percentile-rank scale produced greater overestimation for low performers and greater underestimation for high performers than did the number-correct scale ($ps < .01$).

General discussion

To further illustrate the discrepancy between number-correct and percentile-rank judgment accuracy, in Fig. 1 we present the patterns and magnitudes of overestimation (positive values) or underestimation (negative values) across all four studies. Percentile-rank judgments (left panel) consistently demonstrated the unskilled-and-unaware pattern—that is, the tendencies toward overestimation for low quartiles and underestimation for high quartiles. In contrast, number-correct judgments (right panel) did not consistently produce this pattern, and low performers exhibited no greater tendency to overestimate than did high performers in Studies 1–3. Only in Study 4 was the unskilled-and-unaware pattern observed for number-correct judgments. Even in Study 4, however, the magnitude of the unskilled-and-unaware pattern was larger for percentile-rank judgments than for number-correct judgments. Thus, a major conclusion of these studies is that the type of judgment scale on which performers judge their performance substantially influences their judgment accuracy. Deciding which judgment scale to use should be an important consideration for researchers who investigate individual differences in performers' skill and judgment accuracy. One recommendation from the present series of studies is to collect both kinds of judgments, because as we will discuss next, the discrepancies have implications for understanding potential causes of judgment (in) accuracy.

The difference in accuracy for the two judgment scales has implications for the causes of the unskilled-and-unaware phenomenon. For instance, one possibility was that the inaccuracies of high and low performers arise largely from different sources (e.g., Kruger & Dunning, 1999): Namely, difficulty judging the mean performance and variability of others might be the primary source of inaccuracy for high performers, whereas low performers may additionally have difficulty judging their own performance. Consistent with this possibility, Ehrlinger et al. (2008) found that the accuracy of lower performers' percentile judgments improved when they were statistically corrected for errors in their absolute judgments, which likely occurred because lower performers' absolute judgments were overconfident. Ehrlinger et al.'s regression analysis suggested that the overconfidence shown by lower performers' percentile judgments could arise from inaccurate absolute judgments.

Nevertheless, the results from the present Studies 1–3 provide evidence that is inconsistent with the aforementioned hypothesis. Namely, all levels of performers exhibited *accurate* absolute judgments, indicating that the inaccurate percentile-rank judgments were not caused by poor awareness of one's own performance. Thus, in these studies, both low and high performers' inaccurate percentile-rank judgments may have arisen from the same mechanism, such as difficulties in judging the performance of others. For example, the false-consensus effect (Ross, Greene, & House, 1977) may have led high performers to erroneously believe that others found the task to be approximately as easy as they did, whereas low performers may have erroneously believed that others found the task to be approximately as difficult as they did. By misjudging others as being similar to themselves, their percentile-rank judgments would regress toward the middle of the scale, which would result in overestimates for low performers and underestimates for high performers. Exactly why Ehrlinger et al.'s (2008) regression analysis and

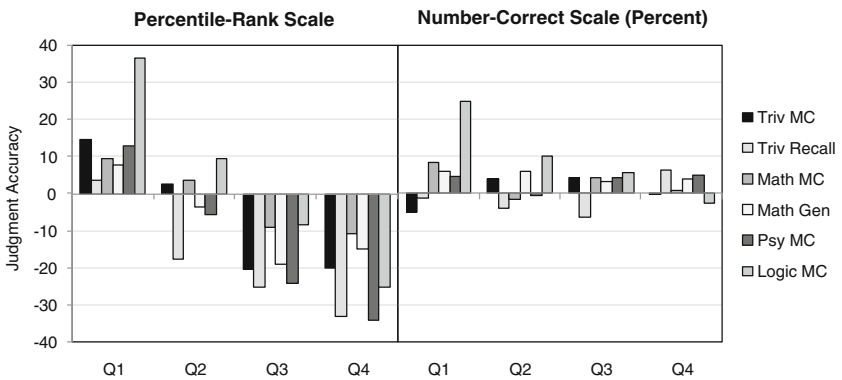

Fig. 1 Signs and magnitudes of accuracy for percentile-rank judgments versus number-correct judgments, across all four studies. For ease of comparison, number-correct judgments were converted to percent-correct judgments

the present data from Studies 1–3 support different conclusions cannot be determined without additional research (because the methods used in these studies varied, and hence the differences might arise from several factors). Certainly, overconfidence in lower performers' percentile judgments may arise from poor self-awareness in some conditions, but it is apparent from the present studies that poor self-awareness is not always responsible for the unskilled-and-unaware effect in percentile judgments.

Regarding whether judging absolute performance is simply an easier task than judging relative performance, we suspect that it is: When judging absolute performance, the performer only needs to judge the self, which in this case would involve estimating the frequency of correct responses across all items on the test. By contrast, judgments of relative performance require both judging the self (i.e., frequency estimates) and estimating the performance of others. Importantly, all levels of performers seem to be challenged when judging their percentile rank, not just the low performers. The bottom line is that differences in the accuracy of percentile-rank judgments across skill levels do not always represent differences in self-awareness. When performers do exhibit mistaken absolute judgments (as in Study 4), self-awareness presumably contributes partly to the inaccuracy observed in relative judgments. However, in Studies 1–3, it appears that all inaccuracy arose from difficulties that performers had at evaluating how well others had performed.

The unskilled-and-unaware phenomenon has previously been assumed to be pervasive for cognitive tasks, regardless of judgment scale. However, Studies 1–3 demonstrated a complete absence of the unskilled-and-unaware pattern when performance judgments were made on an absolute scale. This absence—which occurred across a variety of materials (trivia, math, and psychology)—is noteworthy, because it suggests that global-judgment inaccuracy may be less common than has previously been thought, at least for absolute judgments. Given the possibility that failures to replicate the unskilled-and-unaware pattern have remained unpublished (and in the proverbial file drawer) due to null effects on absolute-judgment accuracy, the prevalence of global inaccuracy for absolute judgments may be overestimated. With this in mind, however, one may still wonder why the standard unskilled-and-unaware pattern did not occur for absolute judgments in Studies 1–3, yet did occur for these judgments in Study 4. Existing theories do not adequately explain why the tasks in Studies 1–3 would differ from the task in Study 4 or those in previous studies that have demonstrated the unskilled-and-unaware pattern with absolute judgments (e.g., Ehrlinger et al., 2008). The present studies were not designed to evaluate why the unskilled-and-unaware pattern does not appear consistently, but differences in the absolute level of performance can be ruled out (Table 2), given that the performance on the logic task (50.5%) that demonstrated the effect with an absolute scale was within the range of performance on the other tasks (33.0%–56.8%). Future research should seek a theoretical explanation to predict when, or for what types of tests or materials, absolute judgments will be accurate versus inaccurate. Below, we speculate on some possible reasons.

One possible difference between the tests, which may affect judgment accuracy, is their reliability. Poor reliability (such as is measured by split-half reliability) has been shown to contribute to global-judgment inaccuracy (Krueger & Mueller, 2002; although it may not entirely account for the unskilled-and-unaware pattern; see, e.g., Kruger & Dunning, 2002). Thus, we chose tasks that were expected to have acceptable reliability, so that any skill-level effects that we did find could not be attributed to this artifact. Importantly, to the degree that highly reliable tasks minimize the unskilled-and-unaware pattern, we must again emphasize that this pattern was large in magnitude for the percentile-rank judgments. Thus, poor reliability will provide only a partial explanation for the pattern, albeit it may account for more of the pattern when it arises for absolute than for percentile-rank judgments. We leave exploration of this possibility for future research. The nature of the test or test items could also contribute to the presence (vs. absence) of the unskilled-and-unaware pattern with absolute judgments. For example, test items that are "tricky," taking advantage of common mistakes in procedure or common errors in knowledge, might be more likely to produce overestimates for low performers. Some content areas (such as logical reasoning) or certain question types may be particularly prone to supporting such tricky items. Furthermore, some content areas or types of tests may be more likely to elicit performers' desires to achieve success or to protect their self-image in the face of low performance, such as when low-performing students overestimate their performance on a class exam (e.g., Dunning et al., 2003; Miller & Geraci, 2011).

What is clear is that unskilled performers are mistaken about their performance on some (but not all) tests, and the unskilled-and-unaware phenomenon is less robust for absolute than for percentile-rank judgments (see also Kruger & Dunning, 1999). Nevertheless, prior research (and also Study 4 presented here) has indicated that absolute judgments can sometimes produce the phenomenon. Thus, the following question becomes important: What conditions do (and do not) produce global unawareness with absolute judgments? That is, when are participants truly unable to judge that their responses are incorrect? More research will be needed to determine the scope of the unskilled-and-unaware phenomenon, because as in the present case, identifying boundary conditions for the phenomenon will contribute to our understanding of it.

Author Note Our thanks to the members of RADlab at Kent State University for helpful comments on these studies. This research was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award.

Appendix

Table 3 Sample items

Questions	Choices (if applicable)			
Study 1: **General-Knowledge Trivia Questions**				
Of which country is Buenos Aires the capital?	(a) Chile	(b) Brazil	(c) Argentina*	(d) Venezuela
What is the last name of the doctor who first developed a vaccine against polio?	(a) Salk*	(b) Jenner	(c) Lister	(d) Pasteur
Study 2: **Algebra-Based Math Problems**				
75% of 88 is the same as 60% of what number?	(a) 108	(b) 110*	(c) 105	(d) 103
On planet Urano, each year has 8 months and each month has 16 days. How many full Urano years will have passed after 600 days?	(a) 8	(b) 3	(c) 6	(d) 4*
Study 3: **Psychology Questions**				
Developmental psychologists use the term "instrumental aggression" to refer to behavior in which an aggressor:	(a) acts to achieve a goal*			
	(b) hurts someone by accident			
	(c) reacts to an attack with greater force than the attacker used			
	(d) attacks with a weapon			
	(e) repeatedly attacks the same person without provocation			
A sample of 50 school-aged children are given either a pill with a certain medication in it or a placebo. The children fill out a survey about their energy level before the treatment begins and again after the two-week treatment is complete. The dependent variable is:	(a) the pill with the medication in it			
	(b) the pill without the medication			
	(c) the children			
	(d) the two weeks of the experiment			
	(e) the children's energy level*			
Study 4: **Logic Questions**				
"Fifty of the 150 businesses in Cutbright Township have closed during the last calendar year. The number of businesses in a community is a sign of economic health; thus it is obvious that Cutbright Township has experienced serious economic decline."				
Which one of the following is an assumption upon which the argument depends?	(a) The businesses that closed were predominately small sole-proprietorships.			
	(b) The sites formerly occupied by the closed businesses are now public buildings or recreation centers.			
	(c) Cutbright Township has experienced similar closings in previous years.			
	(d) Fewer than fifty new businesses opened in Cutbright Township during the last calendar year.*			
	(e) All of the businesses closed in the first quarter of the fiscal year.			
"Everyone in Tom and Angie's class likes drawing or painting or both; but Angie does not like painting."				
Which one of the following statements CANNOT be true?	(a) Tom likes drawing and painting.			
	(b) Angie likes drawing.			
	(c) Tom dislikes drawing and painting.*			
	(d) Everyone in the class who does not like drawing likes painting.			
	(e) No one in the class likes painting.			

* correct response

Table 4 Study 4 means reported separately by order of judgment scale

Judgment Order	Percentile-Rank Scale				Number-Correct Scale		
	Quartile	Estimate	Actual	Diff.	Estimate	Actual	Diff.
Absolute first, then relative	1 (low)	52.2	13.5	38.7*	11.2	6.6	4.5*
	2	52.5	41.7	10.8	11.5	9.4	2.1
	3	57.2	65.6	−8.4	13.1	12.0	1.1
	4 (high)	61.7	87.5	−25.8*	15.1	15.8	−0.7
		High vs. low: $t(23) = 8.0$, $p < .001$**			High vs. low: $t(23) = 3.9$, $p = .001$**		
Relative first, then absolute	1 (low)	48.1	13.6	34.4*	10.3	4.9	5.4*
	2	47.8	39.8	8.0	9.7	7.8	1.9
	3	55.1	63.6	−8.5	12.5	11.4	1.1
	4 (high)	62.7	87.5	−24.8*	13.7	14.1	−0.4
		High vs. low: $t(21) = 7.8$, $p < .001$**			High vs. low: $t(21) = 4.8$, $p < .001$**		

* Mean difference score differed from zero, $p < .05$. ** Mean difference scores of high versus low performers differed in the direction that is consistent with the unskilled-and-unaware phenomenon

References

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90,* 60–77. doi:10.1037/0022-3514.90.1.60

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83–87. doi:10.1111/1467-8721.01235

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105,* 98–121.

Hodges, B., Regehr, G., & Martin, D. (2001). Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine, 76,* S87–S89.

Krajc, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology, 29,* 724–738.

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. doi:10.1037/0022-3514.82.2.180

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. doi:10.1037/0022-3514.77.6.1121

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. doi:10.1037/0022-3514.82.2.189

Mattern, K. D., Burrus, J., & Shaw, E. (2010). When both the skilled and unskilled are unaware: Consequences for academic performance. *Self and Identity, 9,* 129–141.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 502–506. doi:10.1037/a0021802

Moore, D. A., & Cain, D. M. (2007). Overconfidence and undercondidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes, 103,* 197–213.

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgments: On being both better and worse than we think we are. *Journal of Personality and Social Psychology, 92,* 972–989.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19,* 338–368. doi:10.1016/S0022-5371(80)90266-2

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attributional processes. *Journal of Experimental Social Psychology, 13,* 279–301.