Psychonomic Bulletin & Review
1997, 4 (3), 387–392

The domain specificity and generality of overconfidence: Individual differences in performance estimation bias

RICHARD F. WEST
James Madison University, Harrisonburg, Virginia

and

KEITH E. STANOVICH
University of Toronto, Toronto, Ontario, Canada

One hundred twenty-three college students performed a knowledge assessment task and a game of motor skill in which they had to predict their performance before each block of trials. There was a bias in the direction of overconfidence on both tasks, even though the latter involved the motor domain, did not require the use of numeric probabilities, and allowed predictions to be made by using an aggregate judgment made in a frequentist mode. An analysis of individual differences indicated that there was considerable domain specificity in confidence judgments. However, participants who persevered in showing overconfidence in the motor task—despite previous feedback revealing their overconfident performance predictions—were significantly more overconfident in the knowledge calibration task than were participants who moderated their motor performance predictions so as to remove their bias toward overconfidence. The latter finding is consistent with explanations of overconfidence effects that implicate mechanisms with some degree of domain generality.

Overconfidence in performance estimation has been observed in a variety of different paradigms and domains. For example, overconfidence has been observed in perceptual judgments (Baranski & Petrusic, 1995), prediction of sports outcomes (Ronis & Yates, 1987), reading-comprehension monitoring (Glenberg & Epstein, 1987), judging the sex of handwriting samples (Schneider, 1995), prediction of one's own behavior or life outcomes (Hoch, 1985), economic forecasts (Braun & Yaniv, 1992), and of course in the much-investigated knowledge assessment paradigm (e.g., Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein, Fischhoff, & Phillips, 1982). The effect has been obtained from subject samples in a variety of different countries (Lee et al., 1995; Yates et al., 1989) and in subject samples of adolescents as well as adults (Newman, 1984). Although the finding of an overconfidence bias can be substantially modified by variables such as item difficulty,[1] task type, and expertise (Keren, 1991; Ronis & Yates, 1987; Schneider, 1995), it is so ubiquitous that it is routinely classified as a pervasive cognitive bias (Arkes, 1991; Baron, 1994; Fischhoff, 1982).

Despite this overall finding of overconfidence on many tasks, performance across a sample of participants is almost always characterized by enormous variability. It is almost always the case that some participants show no global bias toward overconfidence. Why do these subjects not display the characteristic bias? Do they not share a widely prevalent motivation to make self-serving attributions? Do they not share a characteristic bias in translating subjective uncertainties into an appropriately calibrated probability scale? The answers to such questions are completely unknown, because although overconfidence effects have been found in a variety of different domains, individual differences in the effect and their correlates have generally remained unexplored (see Stanovich, in press).

The relative neglect of individual differences is unfortunate, because patterns of individual differences displayed across tasks may well have implications for contending theories of the overconfidence effect. For example, cognitively based (as opposed to motivationally based) theories have recently been ascendant as explanations of the overconfidence effect in knowledge calibration. Theories emphasizing anchoring and adjustment effects (Block & Harper, 1991), random error statistical models (Erev, Wallsten, & Budescu, 1994; Pfeifer, 1994), response contraction toward a reference magnitude on a .50–1.00 probability scale (Poulton, 1994), and unrepresentative stimulus sampling (Gigerenzer, Hoffrage, & Kleinbolting, 1991; Juslin, 1994) form only a partial list of the cogni-

This research was supported by Grant 410-95-0315 from the Social Sciences and Humanities Research Council of Canada to K.E.S. and a James Madison University Program Faculty Assistance Grant to R.F.W. The authors thank Jamie Pegher, Douglas Tees, Jason Mott, Robin Sidhu, and Penny Chiappe for their assistance in data collection and coding. Art Glenberg, Asher Koriat, and Chuck Weaver are thanked for their comments on an earlier version of the manuscript. Requests for reprints should be sent to K. E. Stanovich, Department of Applied Psychology, Ontario Institute for Studies in Education, University of Toronto, 252 Bloor Street West, Toronto, ON, Canada M5S 1V6 (e-mail: kstanovich@oise.utoronto.edu).

tively based theories that have recently been the subject of intense investigation.

However, many of the cognitively based theories are tied to stimulus-sampling and probability-scale issues that are most likely to arise in the classic knowledge assessment paradigm where subjects must make trial-by-trial subjective probability estimates on a .50–1.00 scale (see Keren, 1991; Poulton, 1994). Several of these cognitive explanations in essence argue that the overconfidence effect is artifactual (e.g., Pfeifer, 1994; Poulton, 1994). Likewise, proponents of ecological models view overconfidence in this particular paradigm as artifactual (Gigerenzer et al., 1991; Juslin, 1994). Interestingly, many of these explanations—in being relatively closely tied to the knowledge assessment paradigm—would need to posit a different mechanism to account for overconfidence in other domains[2] and in paradigms employing aggregate rather than single-case probability judgments. Such explanations would thus be called into question by data suggesting that a common mechanism is operating across tasks involving vastly different domains and having vastly different response and cognitive requirements (i.e., tasks not sharing characteristics such as .50–1.00 probability scales, anchoring opportunities, biased stimulus sampling, use of a numerical probability scale, reliance on a single-case probability mode rather than an aggregate judgment). The presence of correlated individual differences among individuals who have performed two tasks—one not containing the critical task features—would undermine cognitive theories requiring certain domain and paradigm characteristics to explain overconfidence.

In the experiment reported here, we investigated performance in two very different domains: the much-investigated knowledge assessment paradigm, and the domain of motor performance. Confidence/accuracy relationships in the motor domain have been examined before (see Harvey, 1994) but not in the context of parallel performance on a knowledge assessment measure. In addition to tapping skills totally different from those required by the traditional almanac-type knowledge assessment experiment, confidence in the motor task is assessed without the use of a numerical probability scale and in an aggregate frequentist mode rather than in a singular-event probability mode (see Gigerenzer & Hoffrage, 1995)—that is, without many of the features that have been found to artifactually magnify overconfidence (see Keren, 1991; Poulton, 1994).

METHOD

Subjects

The subjects were 123 undergraduate students (33 males and 90 females) recruited through an introductory psychology subject pool at a medium-sized state university. Their mean age was 19.3 years (SD = 4.4 years).

Tasks

Knowledge assessment. The methods and analyses used in this task were similar to those employed in the extensive literature on knowledge calibration (Fischhoff, 1982; Koriat et al., 1980; Lichtenstein, & Fischhoff, 1977; Ronis & Yates, 1987). Subjects answered 70 general knowledge questions in a two-choice choice format. Questions were drawn from Zahler and Zahler's (1988) book, *Test Your Cultural Literacy*. An example question is the following: "What did Anton van Leeuwenhoek invent? a. the microscope, b. the telescope."

After answering each question, the subjects indicated their degree of confidence in their answer on a six-point scale. The points on the scale were .5, .6, .7, .8, .9, and 1.0, with the leftmost point (.5) labeled *just guessing* and the rightmost point (1.0) labeled *absolutely certain*. The subjects were given the following instructions:

> In this task, we would like you to answer a series of multiple choice questions on a variety of different topics. After choosing your answer, indicate on the scale provided what you think the probability is that you answered the item correctly. If you feel that you are absolutely certain your answer is correct, circle probability level 1.0. If you are just guessing, and have no idea what the answer is, circle probability .5 (indicating 50/50 chance of being correct). If you are not certain, but think that there are 9 chances out of ten that you are correct, circle .9. If you are less certain than this, and think that there are 8 chances out of ten that you are correct, circle .8. If you are still less certain than this, and think that there are 7 chances out of ten that you are correct, circle .7. If you are very uncertain, but still think that you are not just guessing, and think that there are 6 chances out of ten that you are correct, circle .6.

Penny slide task. The penny slide task was a simple test of motor skill, loosely modeled on a table shuffleboard game. The subjects sat at the narrow end of a table (91.4 × 182.9 cm) and attempted to slide pennies onto a strip on the table (marked with magic marker) located 154.2 cm from the end of the table and 112.3 cm from a "foul" line. This strip was colored red with a magic marker, was 2.5 cm wide, and extended for a length of 45.7 cm across the table. A diagram of the physical layout of the table is available from the authors. Pennies landing on this strip were scored as 9 points (and demarcated by black magic marker lines). In front of and in back of the 9-point strip were two 8-point strips each 2.5 cm wide. Subsequent 2.5-cm strips representing 7 points, 6 points, 5 points, etc. were placed in front of and in back of each previous strip. Pennies landing in front of the front 1-point strip and in back of the back 1-point strip were scored as zero. When a penny landed on a line, the strip containing the majority of the penny's area determined the score.

The subjects were told that they would be attempting to slide pennies onto the grid and have them stop on the center red stripe. The scoring system was explained, and they were told to try to accumulate as many points as possible. The subjects were given two practice slides to get accustomed to the difficulty of the game. After each practice trial (and after each subsequent trial), the experimenter announced the score for that trial, removed the penny from the table, and recorded the score. The subjects completed two blocks of 30 trials. The score for each block was the total number of points accumulated across the 30 slides. Scores for each block could thus potentially range from 0 to 270; but the observed range for Block 1 was 43–160, and the observed range for Block 2 was 57–159. As for the observed score, the predicted score for each block could potentially range from 0 to 270. The range of the predicted scores for Block 1 was 38–229, and the range of the predicted scores for Block 2 was 23–193.

Prior to each block of 30 trials, the subjects predicted their performance for that block by arranging 30 pennies on the table onto the strips in an effort to mimic their subsequent performance for that block. The sequence of tasks was thus: practice trials, predictions for Block 1, the 30 Block 1 trials, predictions for Block 2, and the 30 Block 2 trials.

Procedure

Participants completed the tasks in a single session that lasted less than an hour. They first filled out a demographics sheet, then completed the knowledge assessment task, and then completed the penny slide task.

RESULTS

Knowledge Calibration

Performance on the knowledge assessment task mirrored that previously reported in the published literature

in all essential aspects. The mean accuracy rate across the 70 questions was 64.9%, and the mean confidence rating was 75.2%. Several measures of knowledge calibration were calculated (Yates et al., 1989, Ronis & Yates, 1987, and Schneider, 1995, should be consulted for discussions of the computational and conceptual details of these indices). The first such index—computationally the simplest—will be our main focus, because it directly assesses the phenomenon on which we wish to focus: overconfidence. Termed the measure of *over/underconfidence* by Lichtenstein and Fischhoff (1977) and *bias* by Yates et al. (1989), it is simply the mean percentage confidence judgment minus the mean percentage correct. The mean bias score in our sample was 10.3% (SD = 8.8), significantly different from zero [$t(122) = 13.04$, $p < .001$]. The positive sign of the mean score indicates that the sample as a whole displayed overconfidence, the standard finding with items of this type. A positive bias score (in the direction of overconfidence) was displayed by 108 (87.8%) of the 123 participants. The calibration curve for the task was prototypical (Lichtenstein & Fischhoff, 1977; Yates et al., 1989). The mean proportion of questions answered correctly in the 1.0 confidence category was .875. Likewise, overconfidence was displayed in the .9 category (.708 correct), .8 category (.649 correct), .7 category (.575 correct), and .6 category (.581 correct). In the .5 confidence category, the proportion of questions answered correctly was .519.

Calibration-in-the-small (see Yates et al., 1989), resolution (Yates et al., 1989), and a normalized discrimination index (Yaniv, Yates, & Smith, 1991) were calculated for each subject and averaged across subjects. These values, .041, .033, and .149, respectively, were in line with those observed in previous studies (Schneider, 1995; Yates et al., 1989). *Calibration-in-the-small* (termed simply *calibration* by Lichtenstein & Fischhoff, 1977) refers to the mean squared difference between the probability label of the category and the percentage correct in that category across all the categories (see Yates et al., 1989). Resolution is the mean squared difference between the percentage correct in each category and the overall mean correct percentage summed across all the categories (see Yates et al., 1989). It reflects "the ability of the responder to discriminate different degrees of subjective uncertainty by sorting the items into categories whose respective percentages correct are maximally different from the overall percentage correct" (Lichtenstein & Fischhoff, 1977, p. 162). The normalized discrimination index (see Yaniv et al., 1991) is a resolution measure that takes into account the total variance in the outcome variable. In summary, on every index of calibration performance, this group of subjects mirrored the trends in the published literature.

There was a moderate degree of internal consistency within the knowledge calibration task. The overconfidence effect (bias statistic) was calculated for both the even and odd items separately, and the split-half reliability (Spearman-Brown corrected) for the bias score was .71, a figure somewhat higher than that obtained by Schraw, Dunkle, Bendixen, and Roedel (1995). The higher reliability of overconfidence bias in our study probably resulted from the larger number of items used in our investigation.

Penny Slide Performance

In Block 1 of the penny slide task, the mean predicted score was 128.1 (SD = 40.2) and the mean attained score was 98.2 (SD = 23.9). Thus, there was a substantial 29.9-point overconfidence effect [$t(122) = 8.13$, $p < .001$]. The predicted scores of 95 of the 123 participants (77.2% of the sample) were higher than their obtained scores. Several indices other than mean difference also indicated a strong overconfidence effect. For example, subjects predicted that 16.6% of their slides would result in scores of zero, whereas 36.9% of the actual trials resulted in slides scoring zero.

On the second block of penny slide trials, the mean predicted score was 115.0 (SD = 31.6) and the mean attained score was 108.6 (SD = 22.6). Thus, the predicted score on Block 2 was significantly lower than that on Block 1 [$t(122) = 3.70$, $p < .001$], and the attained score was significantly higher [$t(122) = 4.97$, $p < .001$]. Nevertheless, on Block 2, there was still a statistically significant overconfidence effect of 6.4 points [$t(122) = 2.11$, $p < .05$]. Although participants adjusted their performance expectations subsequent to Block 1, the adjustment was not enough to eliminate the overconfidence effect. For example, although subjects predicted that more trials would receive a score of zero in Block 2 (25.4%, vs. the 16.6% predicted in Block 1), they still underpredicted the number of such trials that actually occurred in Block 2 (30.5%). Figure 1 presents the predicted and observed frequencies for each of the score categories (0 through 9) on Block 1 (top) and Block 2 (bottom). The shift toward lower overconfidence from Block 1 to Block 2 is apparent (especially in score category 0), as is the failure to entirely eliminate overconfidence.

Table 1 presents the intercorrelations of the predicted and observed performance on each of the two blocks of trials. Individual differences in performance were moderately stable from Block 1 to Block 2 ($r = .51$, $p < .001$), as were predicted scores ($r = .43$, $p < .001$). There was a substantial correlation ($r = .63$, $p < .001$) between actual performance on Block 1 and the performance predicted for Block 2.

Although in an aggregate prediction task such as this there are no strict parallels to calibration-in-the-small and resolution (which require trial-by-trial prediction), and although our focus was on the domain specificity or generality of the overconfidence effect, alternative indices of accuracy can be computed for the penny slide task. For example, the correlation between expected category frequencies and observed frequencies were computed as were the summed absolute deviations between predicted and observed category frequencies. Finally, the sum of the squared deviations between predicted and observed category frequencies was computed. All three of these indices indicated improved accuracy of prediction in

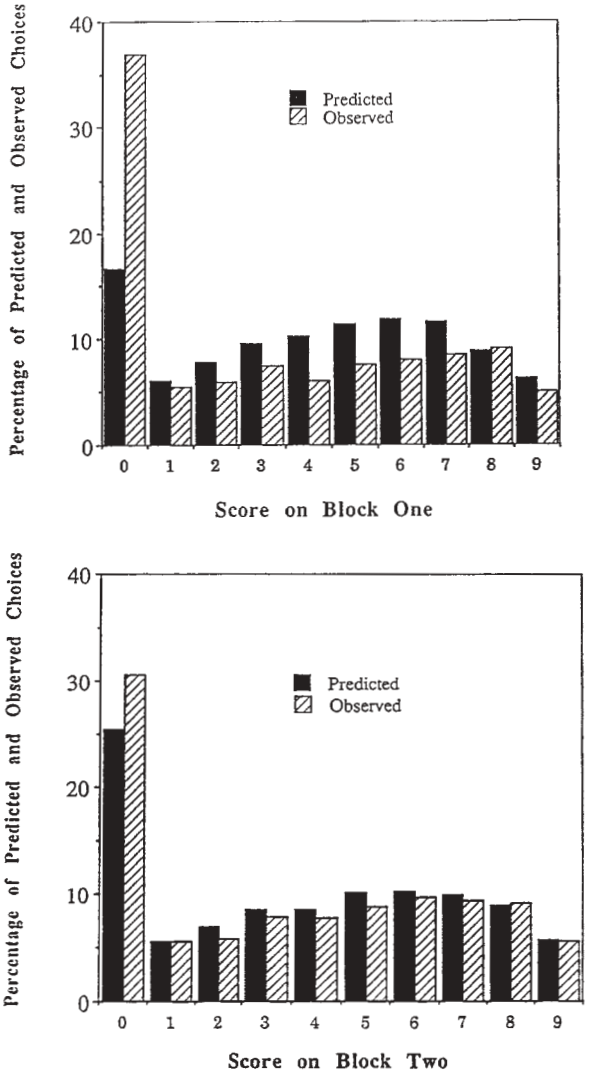

Figure 1. Proportion of predicted and observed scores in each score category on the penny slide task for Block 1 (top) and Block 2 (bottom).

Block 2 as compared with Block 1. The correlation between predicted and observed category frequencies increased (from .23 to .48), and the sum of the absolute deviations decreased (from 24.54 to 20.40), as did the sum of squared deviations (from 125.85 to 77.77).

Bias Across Tasks

Cross-task commonalities in performance were explored by dividing the sample on the basis of a median split of their bias scores on the knowledge assessment task.[3] As is indicated in Table 2, the low-overconfidence group ($n = 61$) answered significantly more questions correctly than did the high-overconfidence group ($n = 62$). Despite their superior performance, the low-overconfidence group gave significantly lower ($p < .001$) confidence ratings. As a result, the overall calibration bias shown by the low overconfidence group (3.2%) was substantially lower than that of the high overconfidence group (17.2%).

The lower part of Table 2 presents the indicators of performance on the penny slide task. Both groups performed similarly on the penny slide task in both blocks (both improved approximately 10 points from Block 1 to Block 2). However, despite their similar performance, the subjects who displayed high overconfidence on the knowledge assessment task made higher predictions in both blocks of the penny slide task than did the low-overconfidence group, although the difference was not statistically significant in Block 1. Predictably, given these trends, the bias was larger for the high-overconfidence group in both blocks (but significantly so only in Block 2). Interestingly, by Block 2, the subjects who displayed low overconfidence on the knowledge assessment task had completely eliminated their overconfidence bias (their predictions displayed a slight underconfidence), whereas the subjects who displayed high overconfidence on the knowledge assessment task still displayed a 13-point overconfidence effect on Block 2 that was significantly different from zero [$t(61) = 3.30$, $p < .01$].

The final two variables listed in Table 2 illustrate two further significant differences. The first variable is the combined overconfidence effects from Block 1 and Block 2, and the last variable is the difference between the prediction for Block 2 and the observed performance on Block 1. The latter variable, on which the two groups displayed a significant difference ($p < .01$), reflects the extent to which the subjects thought their performance on the upcoming block would exceed their performance on the previous one. The subjects who displayed low overconfidence on the knowledge calibration task thought that they would improve by 10 points in the next block, and this was almost exactly the actual extent of their improvement. In contrast, the subjects who displayed high overconfidence on the knowledge calibration task thought that they would improve by 24 points on the next block. However, their actual improvement was no more than that of the other group (10 points).

The correlations displayed on the bottom row of Table 1 converge with the results from the median split displayed in Table 2. There was a significant correlation ($p < .01$) between the magnitude of the overconfidence effect on the knowledge assessment task and the degree

Table 1
Intercorrelations Among the Primary Variables

Variable	1	2	3	4	5	6
Penny Slide Task						
1. Block 1 predicted						
2. Block 1 observed	.27†					
3. Block 2 predicted	.43‡	.63‡				
4. Block 2 observed	.14	.51‡	.26†			
5. Block 1 overconfidence	.83‡	−.32‡	.05	−.16		
6. Block 2 overconfidence	.30‡	.25†	.76‡	−.43‡	.15	
Knowledge Task						
7. Overconfidence	.10	.05	.23†	−.05	.07	.24†

*$p < .05$. †$p < .01$. ‡$p < .001$. (All are two-tailed.)

Table 2
Mean Performance (With Standard Deviations) on the Penny Slide Task of Subjects With Low (n = 61) and High (n = 62) Overconfidence (OC) Scores on the Knowledge Task

	Low OC		High OC		
Variable	*M*	*SD*	*M*	*SD*	*t* value
Knowledge Task					
% Correct	67.9	6.5	61.8	5.9	−5.45‡
% Confidence	71.1	6.6	79.2	6.0	7.13‡
Overconfidence effect	3.2	4.9	17.2	5.3	15.28‡
Penny Slide Task					
Block 1 predicted	123.9	41.0	132.2	39.3	1.14
Block 1 observed	97.9	22.3	98.5	25.6	0.14
Block 1 overconfidence	26.0	38.2	33.7	43.1	1.04
Block 2 predicted	107.7	31.2	122.2	30.6	2.59*
Block 2 observed	108.3	20.3	108.8	24.9	0.12
Block 2 overconfidence	−0.6	34.6	13.4	31.9	2.32*
OC Block 1 + OC Block 2	25.5	56.4	47.0	55.8	2.13*
B2 Predicted − B1 Observed	9.9	21.8	23.7	26.2	3.18†

*p < .05. †p < .01. ‡p < .001. (All are two-tailed.)

of overconfidence on Block 2 of the penny slide task,[4] but not with the degree of overconfidence on Block 1.

DISCUSSION

Analysis of penny slide performance indicated that there was an overconfidence bias in predicting performance on this task. The effect was much larger on the first block of trials but was still significant in Block 2, after the participants had completed 32 trials of the penny slide task. The demonstration of overconfidence in calibrating this task is important because it is a motor task very different from the knowledge judgments that have predominated in the literature on overconfidence. Performance calibration on this task did not overtly employ the use of subjective probabilities or response scales that might artifactually lead to response patterns suggesting overconfidence (see Poulton, 1994). Subjects simply arrayed the pennies on the table where they thought they were going to land. To the extent that subjects operated in a probabilistic mode at all, this prediction task allowed them to operate in a distributional or frequentist mode (Gigerenzer & Hoffrage, 1995; Gigerenzer et al., 1991) and to make an aggregate judgment rather than an on-line, trial-by-trial calibration. Both of these factors have been shown to reduce overconfidence in the knowledge assessment paradigm (Gigerenzer et al., 1991; Griffin & Tversky, 1992; Schneider, 1995; but see Brenner, Koehler, Liberman, & Tversky, 1996). Thus, the subset of cognitive explanations that can be invoked to account for overconfidence on this task is much reduced in comparison with the knowledge calibration situation. Such theories provide no principled way of explaining correlations between overconfidence effects across the two tasks.

Nevertheless, more generic models—such as those which invoke domain general motivational mechanisms (Kunda, 1990)—received only mixed support from these data. The magnitude of the overconfidence effect displayed in Block 1 of the penny slide task was not significantly related to the magnitude of the overconfidence effect displayed in knowledge assessment (although the trend was in that direction). This finding is consistent with the idea that there is considerable domain specificity in the mechanisms generating overconfidence in the two situations.[5] Importantly, however, after one block's worth of feedback, overconfidence in subsequent performance on the penny slide game *was* significantly related to overconfidence displayed in the knowledge calibration task. Subjects who displayed low overconfidence in the knowledge task displayed no bias in estimating penny slide performance on Block 2, whereas subjects who displayed high overconfidence in the knowledge task still displayed an overconfidence bias in penny slide estimation even after a block of trials that had exposed their mistaken optimism. This finding may be viewed as support for a mechanism with some degree of domain generality.

Nevertheless, it must be noted that the degree of association between knowledge assessment overconfidence and penny slide overconfidence in Block 2, although significant, was low. When corrected for attenuation that was due to the imperfect reliability of the variables, the correlation (.34) is still modest. This figure is based on an estimate of the penny slide reliability because of the lack of test–retest reliability for the task. One further limitation of the study is that the fixed task order utilized may have accentuated common variance. Finally, our study does not adjudicate between models of the modest degree of domain generality that was obtained. Nevertheless, the finding of even such modest covariance is important in light of a substantial trend in the literature to treat the overconfidence effect in the knowledge calibration paradigm as largely artifactual (Gigerenzer et al., 1991; Pfeifer, 1994; Poulton, 1994).

REFERENCES

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, **110**, 486-498.

Baranski, J. V., & Petrusic, W. M. (1995). On the calibration of knowledge and perception. *Canadian Journal of Experimental Psychology*, **49**, 397-407.

Baron, J. (1994). *Thinking and deciding* (2nd ed.). Cambridge: Cambridge University Press.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior & Human Decision Processes*, **49**, 188-207.

Braun, P. A., & Yaniv, I. (1992). A case study of expert judgment: Economists' probabilities versus base-rate model forecasts. *Journal of Behavioral Decision Making*, **5**, 217-231.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior & Human Decision Processes*, **65**, 212-219.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, **101**, 519-527.

Ferrell, W. R. (1994). Calibration of sensory and cognitive judgments: A single model for both. *Scandinavian Journal of Psychology*, **35**, 297-314.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422-444). Cambridge: Cambridge University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, **102**, 684-704.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Proba-

bilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, **15**, 84-93.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, **24**, 411-435.

Harvey, N. (1994). Relations between confidence and skilled performance. In G. Wright (Eds.), *Subjective probability* (pp. 321-352). Chichester, U.K.: Wiley.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 719-731.

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

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, **77**, 217-273.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 107-118.

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

Lee, J., Yates, J. F., Shinotsuka, H., Singh, R., Onglatco, M., Yen, N., Gupta, M., & Bhatnagar, D. (1995). Cross-national differences in overconfidence. *Asian Journal of Psychology*, **1**, 63-69.

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

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration and probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge: Cambridge University Press.

Newman, R. S. (1984). Children's numerical skill and judgments of confidence in estimation. *Journal of Experimental Child Psychology*, **37**, 107-123.

Pfeifer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior & Human Decision Processes*, **58**, 203-213.

Poulton, E. C. (1994). *Behavioral decision theory: A new approach*. Cambridge: Cambridge University Press.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior & Human Decision Processes*, **40**, 193-218.

Schneider, S. L. (1995). Item difficulty, discrimination, and the confidence-frequency effect in a categorical judgment task. *Organizational Behavior & Human Decision Processes*, **61**, 148-167.

Schraw, G., Dunkle, M., Bendixen, L., & Roedel, T. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology*, **87**, 433-444.

Stanovich, K. E. (in press). *Variable rationality*. Mahwah, NJ: Erlbaum.

Yaniv, I., Yates, F. J., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, **110**, 611-617.

Yates, J. F., Zhu, Y., Ronis, D., Wang, D., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior & Human Decision Processes*, **43**, 145-171.

Zahler, D., & Zahler, K. (1988). *Test your cultural literacy*. New York: Simon & Schuster.

NOTES

1. Overconfidence is higher for item sets that are more difficult.

2. See Ferrell (1994) for an exception.

3. One consistent finding in the knowledge calibration literature that creates a potential problem for individual difference analyses is that the overconfidence effect is higher for more difficult items (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982). If a sample is partitioned on the basis of subjects' degree of overconfidence as calculated with the traditional bias score—the mean percentage confidence judgment minus the mean percentage correct—then the subjects displaying low overconfidence will most likely have attained a higher percentage correct on the knowledge measure. Thus, any variable that correlates with knowledge will almost invariably display a negative correlation with the degree of overconfidence. This was not a problem in the present study, because the percentage of questions answered correctly on the knowledge test was not associated with any component measure (either observed or predicted scores) of the penny slide task.

4. This correlation increases from .24 to .34 when corrected for attenuation that was due to unreliability under the assumption that the penny slide task is as reliable as the knowledge assessment task. Further, the correlation between overconfidence on the two tasks remained significant after the percentage correct on the knowledge questions and the actual performance on Block 2 were partialled out [partial $r = .18$, $F(1,122) = 3.92$, $p < .05$].

5. Likewise, Lee et al. (1995) found no relation between knowledge calibration bias and a nonperformance peer-comparison measure and concluded that "the two phenomena rest on different mechanisms" (p. 67).

(Manuscript received August 6, 1996;
revision accepted for publication January 10, 1997.)