The Journal of General Psychology, 2002, *129*(3), 257–299

The Role of Individual Differences in the Accuracy of Confidence Judgments

GERRY PALLIER
REBECCA WILKINSON
VANESSA DANTHIIR
SABINA KLEITMAN
School of Psychology
The University of Sydney, Australia

GORAN KNEZEVIC
Department of Psychology
University of Belgrade, Yugoslavia

LAZAR STANKOV
RICHARD D. ROBERTS
School of Psychology
The University of Sydney, Australia

ABSTRACT. Generally, self-assessment of accuracy in the cognitive domain produces overconfidence, whereas self-assessment of visual perceptual judgments results in underconfidence. Despite contrary empirical evidence, in models attempting to explain those phenomena, individual differences have often been disregarded. The authors report on 2 studies in which that shortcoming was addressed. In Experiment 1, participants (N = 520) completed a large number of cognitive-ability tests. Results indicated that individual differences provide a meaningful source of overconfidence and that a metacognitive trait might mediate that effect. In further analysis, there was only a relatively small correlation between test accuracy and confidence bias. In Experiment 2 (N = 107 participants), both perceptual and cognitive ability tests were included, along with measures of personality. Results again indicated the presence of a confidence factor that transcended the nature of the testing vehicle. Furthermore, a small relationship was found between that factor and some self-reported personality measures. Thus, personality traits and cognitive ability appeared to play only a small role in determining the accuracy of self-assessment. Collectively, the present results suggest that there are multiple causes of miscalibration, which current models of over- and underconfidence fail to encompass.

Key words: calibration, cognitive ability, confidence judgments, metacognition, self-assessment

SOME 25 YEARS AGO, Lichtenstein and Fischhoff (1977) presented psychologists with an intriguing question: "Do those who know more also know more about how much they know?" (p. 159). Although interest in the psychology of self-assessment was hardly new (see, e.g., Fullerton & Cattell, 1892; Griffing, 1895), their article was pivotal in reviving interest in metacognitive processes, a component of which is self-assessment (Stankov, 1999). The impetus for research into the so-called confidence paradigm derives from experimental cognitive psychology (see Harvey, 1997, for a review) and thus far has mainly led to examinations of the possible cause (or causes) of confidence bias, that is, the miscalibration of people's confidence in the accuracy of their responses to various stimuli. After a brief description of the methods typically used in such studies, we review in this article the major conceptualizations of confidence bias and point to the need to consider individual differences within those frameworks.

Expressing the Accuracy of Confidence Judgments

Confidence bias is one of several robust research findings associated with the confidence paradigm. The term refers to a systematic error of judgment made by individuals when they assess the correctness of their responses to questions relating to intellectual or perceptual problems. One obtains a person's self-assessment of the accuracy of his or her response simply by asking for a rating of confidence, on a percentage scale. Typically, confidence ratings are grouped into discrete categories; for example, 66–75% would be the 70% category, and 76–85% the 80% category, and so on.

The correspondence between subjective probability (i.e., a personal assessment of accuracy) and the actual probability of a correct response (i.e., objective or empirical result) provides a measure of calibration (Phillips, 1973). The statistic used in the current study was the over- and underconfidence rating, commonly known as the *bias score*, which is referred to as "calibration-in-the-large" by Yates (1990, p.79).[1] The overall bias score represents the mean difference between the confidence ratings over all categories and the average of correct responses for each test sequence. A positive bias score represents overconfidence,

Two pilot studies and Experiment 2 were conducted at The University of Sydney (assisted by a Departmental Research Grant). Experiment 1 was performed at the Human Resources Laboratory, Brooks Air Force Base, TX, where the last author was a National Research Council Fellow. Due acknowledgment is given to those institutions.

The authors thank Dragoco (Australia) P/L for generously supplying the odorant used in Experiment 2 and Sensonics Inc. (USA) for allowing a discount on the purchase price of the smell identification labels. The authors also thank the editor of this journal and 2 anonymous reviewers for their helpful and insightful comments on a previous version of this article.

Address correspondence to Gerry Pallier, School of Psychology, The University of Sydney, Sydney, NSW, 2006, Australia; gerryp@psych.usyd.edu.au (e-mail).

and a negative bias score represents underconfidence. A bias score of zero indicates accurate self-assessment.

Research findings suggest that people are usually overconfident (positive bias score) on tests of acculturated ability, such as vocabulary and general knowledge tasks (Juslin, 1994; Kleitman & Stankov, 2001; Stankov, 1998; Stankov & Crawford, 1996a, 1997). In contrast, people are more often underconfident (negative bias score) on sensory and perceptual tasks (Baranski & Petrusic, 1999; Björkman, Juslin, & Winman, 1993; Juslin, 1994; Stankov, 1998; Stankov & Crawford, 1996b, 1997). However, the latter finding has recently been questioned in a study undertaken by Stankov and Pallier (2002). Examination of confidence ratings obtained from judgments that participants made when they used the five possible modalities of sensory input led Stankov and Pallier to conclude that, in general, only visual stimuli consistently elicit the underconfidence phenomenon.

Global Theoretical Models of Overconfidence

In several theoretical models, investigators have attempted to provide an explanation for the overconfidence and underconfidence phenomena. Currently, the most prominent are the heuristics and biases approach (see, e.g., Kahneman, Slovic, & Tversky, 1982) and the ecological approach (see, e.g., Gigerenzer, 1991). Those models are built on different probabilistic theories, and they propose different psychological explanations for the overconfidence phenomenon.

Heuristics and biases approach. The main claim in the heuristics and biases approach is that error in confidence judgments occurs because of general cognitive biases, heuristics, or both, that are assumed to mediate intuitive predictions and judgments (Kahneman & Tversky, 1996). That position has been formalized in the *strength–weight model* (see Griffin & Tversky, 1992). The term *judgmental heuristics* denotes a "small number of distinctive mental operations" (Kahneman & Tversky, 1996, p. 582), a technique commonly used for problem solving, but one that does not guarantee a correct solution. Thus, according to those theorists, although it is often useful to use judgmental heuristics, that technique might lead to characteristic errors (or biases).

Ecological approach. According to the probabilistic mental model (PMM; Gigerenzer, Hoffrage, & Kleinbölting, 1991), each item in a test has its own reference class and target variable. It is suggested in the PMM that environmental knowledge provides cues that are used to solve mental problems. Each cue has a "cue validity" that provides the basis for unique confidence ratings. Confidence ratings are thus determined by knowledge of the relative frequency of events in the natural environment, and researchers assume that when a person answers a question, both the person's choice of response and his or her confidence in that choice are generated by the same cue (Gigerenzer et al.).

According to proponents of the PMM, however, many questions forming typical general knowledge tests are misleading (see, e.g., Juslin, 1993; Winman & Juslin, 1993). Consequently, their cue validities do not correspond to their ecological validities (i.e., the real state of affairs in the natural environment). Therefore, people are tricked by those questions into believing that they have a high probability of being correct when, in fact, the most popular cue for a particular question will lead to an incorrect response. Proponents of the PMM believe that the disparity between cue and ecological validities results in the overconfidence phenomenon. In contrast to the heuristics and biases approach, in the ecological model, overconfidence is seen to be derived from the procedures involved in the creation of traditional general knowledge items rather than from judgmental biases within the individual.

Confidence judgments: An individual-differences perspective. Stanovich and West (1998) pointed out that consistent individual differences have been found in examinations of confidence judgments. Those differences have also been acknowledged within the experimental field. For example, Soll (1996, p. 133) remarked on "potentially important individual differences among subjects" that were observed during his empirical investigation of the PMM. To that end, he noted that some individuals have a disposition toward overconfidence, whereas others show the opposite trend. That finding is by no means trivial. One group of Soll's participants had an average bias score approximating 30%, whereas another group had an average bias score of less than 10%.[2] Similarly, Stankov and Crawford (1996a) reported that although the mean bias score for a test might reveal over- or underconfidence, up to 30% of participants might show the opposite trend.

The confidence paradigm has also been examined in research programs originating from within differential psychology. For example, Schraw and his colleagues found that confidence ratings assigned to measures of cognitive ability correlate more highly among themselves than do accuracy scores from the same test batteries (Schraw, 1994, 1997; Schraw & Dennison, 1994; Schraw & Roedel, 1994). Similarly, Stankov and his associates have produced evidence supporting the existence of a confidence factor (Crawford & Stankov, 1996a, 1996b; Kleitman & Stankov, 2001; Stankov, 1998, 2000; Stankov & Crawford, 1996a, 1996b, 1997). Both sets of findings are highly suggestive of an independent metacognitive trait that mediates the accuracy of self-assessment. Thus, it is argued in current differential models that the cause of miscalibration is a tendency for individuals to express a consistent confidence level, irrespective of their accuracy level. However, proponents of the differential approach have so far failed to satisfactorily examine the relation of confidence bias to the traditional individual difference variables of intelligence and personality.

Rationale for the Current Studies: Toward an Integrated Approach

Experimentally derived evidence has provided at least partial support for the

three theoretical propositions just outlined (i.e., heuristics and biases, ecological, and individual-differences approaches). Accordingly, overconfidence might be caused (to some extent) by reasons offered in any of those (seemingly disparate) perspectives, either separately or synergistically. Following the tradition of multivariate research, in the current studies we incorporated promising avenues of evidence stemming from the three theoretical positions. Thus, because the present experimental design allowed for the examination of propositions regarding the effects of both question content and format, derived respectively by ecological and by biases and heuristics theorists, those propositions were incorporated in the analyses that follow.

In planning the two studies reported herein, consideration was given to two important points. First, would it be possible to predict confidence bias on one test from the data obtained from another task? Second, could the generality of individual differences in confidence bias reported by Stankov and Pallier (2002) be replicated across a wider than customary battery of both intellectual and perceptual tests? An affirmative answer to those questions would strengthen the proposition that consistent individual differences in confidence levels provide a meaningful, causal, explanation for at least part of the miscalibration phenomenon. Indeed, Stankov (1999) proposed that such a metacognitive trait mediates the accuracy of confidence judgments, but the possible relation of that trait to variations in cognitive ability and personality remains uncertain. Our major aim in the current studies was thus to elucidate the nature of the purported trait within the domains of intelligence and personality.

EXPERIMENT 1

Rationale and Aims

Confidence bias and intelligence. Although the role of individual differences in the resolution of confidence bias appears indisputable, a number of questions remain to be answered. A major point of contention lies in the answer to the intriguing aforementioned question posed by Lichtenstein and Fischhoff (1977). Indeed, there has been some evidence of a possible correlation between overconfidence (i.e., positive bias score) and the accuracy of test scores. For example, Zakay and Glicksohn (1992) asked 43 undergraduates to express their confidence in the answers that they supplied to a "real-life" psychology examination. The analysis of data obtained from the individuals' self-assessment and their university results suggested that overconfident individuals achieve lower grades than do those whose scores and confidence ratings are better calibrated.

To address the role of individual differences in confidence bias, it seems critical for the investigator to eliminate experimental dependency. That dependency results when researchers derive bias scores from accuracy and confidence measures by using the same test instrument. Thus, a possible criticism of the corre-

lational approach to confidence ratings is that an experimental artifact caused by the use of three measures obtained from the same test contaminates the ensuing statistical analyses (Judd, Smith, & Kidder, 1991). In the current study, we used parallel versions of tests presented in both open-ended and multiple-choice formats. By implementing that method, one can compare bias scores from one test with accuracy scores and confidence ratings obtained from a second (parallel) version. Consequently, in many of the analyses that follow, concern regarding experimental dependency was minimized or eliminated.

Determining a confidence factor by using confirmatory analysis. In the present study, we used confirmatory factor analytic techniques to examine evidence that there is a self-confidence trait. The use of confirmatory factor analyses (CFA) was appropriate in the present study because those methods can provide evidence to support constructs previously identified by exploratory methods (Stankov, 2000). Furthermore, Carroll (1993) noted that to obtain meaningful results by using factor analytic methods, one must ensure that a sufficient number of tests for specific abilities (i.e., markers) are administered to an adequate number of participants. One of our intentions in the present study was to ensure that those conditions were met. Thus, eight measures, derived from established markers of cognitive ability, were completed by a larger than usual sample of participants (N = 520).[3]

The role of question format. In studies conducted by Koehler (1994) within the heuristic and biases approach, one group of participants was asked to generate answers to open-ended questions (choosing-answer condition). Another group was asked to state their confidence in the accuracy of those answers (evaluation condition). Confidence ratings were significantly lower in the former condition than they were in the latter (Koehler). However, the findings were reversed when multiple-choice questions were used: People in the choosing-answer condition were more confident than were those in the evaluation condition. Those findings suggest that when one is answering a question, one's level of confidence is sensitive to question format. That proposition was also examined in Experiment 1. Both open-ended and multiple-choice question formats for a number of different cognitive abilities were included in the design. Given the evidence just presented, we expected that the open-ended-format tests would elicit less confidence bias than would the multiple-choice-format tests.

Method

Participants

The sample consisted of 520 United States Air Force recruits (80 were women) from Brooks Air Force Base, TX, who were undergoing their 6th week

of basic training. The mean age of the participants was 20.03 years (*SD* = 2.52 years), and 96% spoke English as a first language.

Test Descriptions

Eight tests measuring four cognitive domains within the theory of fluid and crystallized intelligence (Gf/Gc; Cattell, 1943; Horn & Cattell, 1966) were administered to all participants. The cognitive domains assessed were vocabulary (Gc), general knowledge (Gc), visualization (Gv), and abstract reasoning (Gf). For each ability construct, both multiple-choice and open-ended tests were used; the open-ended test of general knowledge and the open-ended test of abstract reasoning were newly devised forms of existing instruments.[4] All tests were given in paper-and-pencil format. Before each test, the participants were presented with instructions on the particular task and an example. After every test item was a scale, marked from either 20% (multiple-choice items) or 0% (open-ended items) to 100% (both formats), in 10% intervals. Participants were told that *just guessing* was indicated by 0% (for the open-ended questions) and by 20% (for the multiple-choice items), and that a confidence rating of 100% indicated that one was *absolutely sure* that the correct response had been chosen. A time limit determined from a pilot study was placed on all tests. The time imposed was found to be sufficient for approximately 95% of participants in the pilot study to complete the tests (under the instruction to work as quickly and accurately as possible). A detailed description of each test follows:

1. *Multiple-Choice Synonyms Vocabulary Test.* This test was taken from the Factor Reference Kit of French, Ekstrom, and Price (1963). The participants were presented with a key word and were then asked to choose (from among five alternatives) the word with the meaning closest to that of the key word. In total, 18 items were given, and a 4-min time limit was imposed.

2. *Open-Ended Synonyms Vocabulary Test.* The items used for this test were derived from the Gf/Gc Quickie Battery (Stankov, 1997). Test 2 was similar to the vocabulary subtest of the Wechsler Adult Intelligence Scale-III (WAIS-III; Psychological Corporation, 1997). Most important, this test has also been found to have close structural concordance with items in Test 1 (see, e.g., Pallier, Roberts, & Stankov, 2000; Roberts, Pallier, & Stankov, 1996). The participants were presented with a key word and were asked to write (on a line next to the key word) either a synonym or a short explanation of the word's meaning. The test consisted of 18 items and had a time limit of 4 min.

3. *Multiple-Choice General Knowledge Test.* This test also was derived from the Gf/Gc Quickie Battery (Stankov, 1997). The participants were presented with questions assessing their knowledge of history, geography, current events, science, and technology, along with five response alternatives. Test 3 consisted of 18 such items and had a time limit of 4 min.

4. *Open-Ended General Knowledge Test.* The items used for Test 4 were open-ended versions of the Multiple-Choice General Knowledge Test. As such, this test is similar to the information subtest of the WAIS-III (Psychological Corporation, 1997). We constructed this test by choosing questions that were considered equivalent to those in Test 3 in terms of item difficulty. In total, 18 items were given; the participants were required to write their answers in the response booklet. Five minutes were given to complete this test.[5]

5. *Multiple-Choice Visualization Test.* The test used here was the Hidden Figures Test (French et al., 1963), which is a measure of the Flexibility of Closure primary factor. At the top of the test page were five shapes, labeled *A* through *E*; beneath the shapes were nine rectangles. Intersecting lines ran through each rectangle so that one of the five shapes was hidden in each rectangle. The participants were instructed to identify the shape hidden in each rectangle. A time limit of 8 min was placed on this test.

6. *Open-Ended Visualization Test.* The Concealed Words Test (French et al., 1963) was used for this part of the battery. That test is a measure of the Speed of Closure primary factor. A list of 18 words was presented, all with parts of the script degraded so that the word was not immediately identifiable. The participants were required to write (on a line provided next to each word) what they thought the degraded word represented. The participants were allowed 8 min to complete Test 6.

7. *Multiple-Choice Reasoning Test.* A mixture of the standard and advanced versions of the Raven's Progressive Matrices (RPM) test (Raven, Court, & Raven, 1979) was included in this 18-item test. The participants were presented with a 3×3 array of symbols, with the bottom right-hand symbol missing. They were instructed to choose, from five symbols given below the matrix, which was the correct symbol (logically) to complete the matrix. A time limit of 6 min to complete the test was imposed.

8. *Open-Ended Reasoning Test.* As in Test 7, 18 items selected from the standard and advanced RPM test were used in this free-response test. The description of the test requirements is similar to that of Test 7. However, no alternative solutions were provided. Instead, the participants were instructed to deduce the symbol necessary to complete the matrix and to draw that shape in a blank box. To that end, we chose only items that we considered simple enough to draw, with the items representing, to the same proportion, each of the varying levels of difficulty in Test 7. To fit that dual purpose, we chose items for this test that fell immediately after (or before) equivalent items used in the open-ended version. The participants were allowed 6 min to complete this test.

Procedure

Before undertaking the tests, the participants were informed that the study was confidential and they were given a brief rationale for the experiment. Before

each particular test, time was allowed for the participants to read the instructions that were specific to that test. Because the participants were told that they had to answer every question, even if they had to guess, the proctor informed them when half the allotted time had elapsed so that they could pace their remaining responses as required. When the time limit had lapsed, the proctor told the participants to cease writing and to read the instructions for the next test. The battery of tests took approximately 1 hr to complete.

Results

Reliabilities of the Calibration Measures

Reliabilities (Cronbach's coefficient alpha) for the various measures obtained from Experiment 1 are reported in Table 1. All reliabilities were considered satisfactory for an experimental study, following the guidelines of Guilford and Fruchter (1978; see also, Gregory, 1996).

Descriptive Statistics: Open-Ended Versus Multiple-Choice Responses

The mean and standard deviation of accuracy, confidence, and bias scores (obtained from the eight cognitive ability tests) are presented in Table 2. It is important to note that all the tests except RPM (Tests 7 and 8) were rather difficult—all arithmetic means for accuracy scores fell below the 50% mark.

Of interest to the current study, for every pair of tests (except RPM) there was less confidence bias for open-ended questions (see Table 2). Paired sample *t* tests of differences in bias scores between the two relevant question formats indicated that those differences were significant at the .01 level (see Table 3). From the results presented herein, it was apparent that question format (i.e., open-ended versus multiple choice) does play a role in confidence bias, at least for certain cognitive abilities.

Contrary to expectations, further inspection of Tables 2 and 3 reveals that although virtually no difference was found between the RPM bias scores for ques-

TABLE 1
Averaged Reliabilities (Cronbach's alpha) for the Measures Extracted in Experiment 1

Measure	Vocabulary	General knowledge	Visualization	RPM
Confidence	.87	.90	.83	.86
Accuracy	.62	.68	.62	.49
Bias score	.65	.67	.65	.56

Note. RPM = Raven's Progressive Matrices.

TABLE 2
Means and Standard Deviations of Accuracy, Confidence, and Bias Scores for the Eight Cognitive Ability Tests

	Accuracy (% correct)		Confidence (Average %)		Bias score	
Test	*M*	*SD*	*M*	*SD*	*M*	*SD*
Vocabulary (multiple-choice)	41.05	10.95	63.77	12.93	22.72	14.74
Vocabulary (open-ended)	36.59	14.71	53.88	18.44	17.29	15.86
General knowledge (multiple-choice)	39.32	14.71	63.22	16.26	23.90	17.69
General knowledge (open-ended)	31.59	17.30	45.93	19.09	14.33	15.58
Visualization (multiple-choice)	36.65	23.54	49.47	23.21	12.83	21.66
Visualization (open-ended)	44.66	16.07	49.82	19.17	5.16	13.54
RPM (multiple-choice)	75.27	14.63	88.37	11.67	13.10	14.09
RPM (open-ended)	68.61	12.13	81.86	13.11	13.25	13.69

Note. RPM = Raven's Progressive Matrices.

TABLE 3
Paired Sample *t* Tests of Differences in Bias Scores for the Alternative Versions of the Cognitive Ability Tests

Test	*M*	*SD*	*t*	*df*	*p*
Vocabulary MC–OE	5.43	16.49	7.51	519	0.001
General knowledge MC–OE	9.57	16.11	13.55	519	0.001
Visualization MC–OE	7.66	22.10	7.91	519	0.001
RPM MC–OE	–0.15	15.77	–0.21	519	0.832

Note. MC = multiple-choice. OE = open-ended. RPM = Raven's Progressive Matrices.

tion format, overconfidence bias was exhibited on that test. Note also that the bias score for RPM was lower than that of any Gc measure used in Experiment 1, especially the various multiple-choice versions. However, previous studies in our laboratory have indicated that very good calibration is the norm when undergraduate participants take the RPM.

Correlations Among the Tests

Pearson product–moment correlations between the accuracy and the confidence scores of Table 2 were also computed. Those correlations are presented in Table 4. Inspection of Table 4 reveals two salient features.

First, except for the visual tasks, there was a reasonably good association between the two versions of the same test in accuracy levels (*r*s between .49 and .61). That finding indicates that people who performed well on one version of the test performed well on the other version of the test. One exception to that general finding is worth noting. To assess broad visualization, we used two different markers to create open-ended and multiple-choice versions. Usually, both tests might be expected to define a second-order, General Visualization (Gv) factor. That did not appear to be a likely outcome with the present sample, however, because the correlations obtained from the variables used in Tests 5 and 6 were rather low (accuracy, $r = .18$; confidence, $r = .29$).

Second, all correlations (*r*s ranged .21 to .72) between confidence levels for the different tests were significant at the .01 level. That consistent, and at times substantial, level of correlation across all tests suggests that confidence ratings might contain a common component that is independent of either question content or question format.

Confirmatory Factor Analysis: Evidence for a Broad Confidence Factor

To examine and simplify interpretation of correlational matrices, investigators commonly use factor analytic procedures. To investigate whether there was a separate confidence factor, we used the maximum likelihood method from the AMOS program (Arbuckle & Wothke, 1999) to conduct a confirmatory factor analysis. The first model tested was based on the findings from several studies conducted in our laboratory (and elsewhere) in which similar batteries of tests had been used (see, e.g., Stankov, 2000). The results of the previous research suggested that the present study would yield (a) a single Self-Confidence factor (with loadings from all the confidence rating scores) and (b) four further factors corresponding to ability measures (i.e., Vocabulary, General Knowledge, Raven's Progressive Matrices, and Visualization). Several modifications to that initial model were carried out. Finally, a six-factor model produced the most acceptable measures of fit, with a chi-square goodness-of-fit statistic equal to 169.10 ($df = 78$; $p = .000$). The root mean square error of approximation (RMSEA) equaled 0.047 (the 90% confidence interval for

TABLE 4
Pearson Product–Moment Correlations Between the Accuracy and Confidence Scores Reported in Table 1

Test	1	2	3	4	5	6	7
			Vocabulary				
1. Accuracy (MC)	—						
2. Confidence (MC)	.246	—					
3. Accuracy (OE)	.485	.247	—				
4. Confidence (OE)	.357	.585	.562	—			
			General knowledge				
5. Accuracy (MC)	.408	.189	.503	.344	—		
6. Confidence (MC)	.160	.516	.260	.527	.351	—	
7. Accuracy (OE)	.413	.234	.600	.444	.614	.366	—
8. Confidence (OE)	.289	.463	.441	.633	.462	.723	.638
			Visualization				
9. Accuracy (MC)	.126	.086	.160	.111	.131	.087	.157
10. Confidence (MC)	.066	.274	.142	.257	.091	.302	.119
11. Accuracy (OE)	.221	.124	.225	.204	.270	.178	.231
12. Confidence (OE)	.116	.306	.163	.396	.219	.440	.195
			Raven's Progressive Matrices				
13. Accuracy (MC)	.243	.104	.259	.204	.242	.188	.236
14. Confidence (MC)	.111	.249	.050	.213	.092	.388	.109
15. Accuracy (OE)	.265	.143	.327	.221	.266	.163	.312
16. Confidence (OE)	.125	.293	.127	.298	.105	.426	.135

Note. MC = multiple-choice. OE = open-ended.

the RMSEA was 0.038–0.057). The Tucker–Lewis index was .963, complemented by a comparative fit index of .976, with the goodness-of-fit index yielding .961. Because all aforementioned indices were within the most conservative acceptable levels, the results indicated a reasonably good degree of model fit. The results of the CFA are shown in Table 5, and interpretation of the six-factor model follows.

Factor 1: Confidence. This factor was clearly defined by confidence ratings from all eight ability tests, with no loadings from any other source.

Factor 2: Crystallized Intelligence (Gc). Whereas the highest loadings on Gc came from both accuracy and confidence on the general knowledge tests, salient loadings from the accuracy scores on both vocabulary measures were also evidenced. Hence, we interpreted this factor as representing acculturated knowledge.

8	9	10	11	12	13	14	15	16
—								
.080	—							
.252	.571	—						
.231	.178	.180	—					
.417	.102	.288	.718	—				
.200	.254	.263	.232	.215	—			
.296	.166	.326	.241	.386	.444	—		
.229	.267	.223	.276	.238	.484	.326	—	
.356	.141	.293	.226	.398	.368	.649	.414	—

Factor 3: Raven's Progressive Matrices. The only loadings here came from both accuracy and confidence scores on the RPM.

Factor 4: Vocabulary. Loadings on the 4th factor were exclusively from accuracy and confidence scores obtained from the two vocabulary tests.

Factor 5: Open-Ended Visualization. This factor was defined by the open-ended visualization questions of the Concealed Words Test, with high salient loadings from both confidence and accuracy.

Factor 6: Multiple-Choice Visualization. As for Factor 5, the only loadings on Factor 6 came from one source. The factor in that case was defined by accuracy and confidence on the Multiple-Choice Visualization Test (i.e., Hidden Figures).

As noted previously, the unexpected appearance of two separate visualiza-

TABLE 5
Results of a Confirmatory Factor Analysis of the Accuracy Scores and Confidence Ratings Obtained From the Cognitive Ability Tests

Measure	Factor 1 Conf	Factor 2 Gc	Factor 3 RPM	Factor 4 Voc	Factor 5 Gv-OE	Factor 6 Gv-MC
Vocabulary						
1. Accuracy (MC)		.168		.445		
2. Confidence (MC)	.642			.598		
3. Accuracy (OE)		.240		.616		
4. Confidence (OE)	.630			.934		
General knowledge						
5. Accuracy (MC)		.703				
6. Confidence (MC)	.739	.665				
7. Accuracy (OE)		.863				
8. Confidence (OE)	.594	.887				
Visualization						
9. Accuracy (MC)						.710
10. Confidence (MC)	.375					.870
11. Accuracy (OE)					.802	
12. Confidence (OE)	.415				.923	
Raven's Progressive Matrices						
13. Accuracy (MC)			.666			
14. Confidence (MC)	.474		.710			
15. Accuracy (OE)		.129	.608			
16. Confidence (OE)	.499		.714			

Note. MC = multiple-choice. OE = open-ended. Conf = confidence. Gc = crystallized intelligence. RPM = Raven's Progressive Matrices. Voc = vocabulary. Gv = general visualization ability.

tion factors was probably the result of the difference in the nature of the two tests we used to demarcate the Gv factor in the design. Nonetheless, it is interesting to note that the lowest bias scores of all tests occurred with the visualization measures (see Table 2). In previous studies on self-assessment of accuracy in spatial tasks, relatively simple perceptual measures have been used that have (generally) produced underconfidence. Plausibly, there is mediation between the cognitive demands of those two tests and more elementary perceptual demands, which appears to lead to good calibration.

The intercorrelations between those six factors are presented in Table 6. The largest correlation (r = .75) was between the Vocabulary and the Crystallized Intelligence factors. That outcome was in accordance with expectations because the vocabulary accuracy scores loaded on Gc (vocabulary being a component of acculturated knowledge). As has been reported previously (e.g., Pallier et al., 2000), there were also relatively high (around .40) correlations between the RPM factor and the two Visualization factors.

Principal Component Analyses: The Role of Cognitive Ability in the Accuracy of Confidence Judgments

It is important to reiterate that the results of the above analyses were indicative of the existence of an independent confidence trait. The empirical evidence so far presented provided strong support for that claim. However, the results did not sufficiently indicate the relationship between the confidence trait and cognitive abilities.

To that end, we performed a principal components analysis on accuracy, confidence, and bias scores obtained from the four open-ended tests. We conducted a second analysis with the four multiple-choice tests, again assessing the accuracy, confidence, and bias scores. Only the first principal component was retained in each case, and intercorrelations between the principal component for accura-

TABLE 6
Intercorrelations for the Factors Identified in Table 5

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Factor 1: Conf	—					
Factor 2: Gc	–.291	—				
Factor 3: RPM	–.182	.420	—			
Factor 4: Voc	–.390	.753	.379	—		
Factor 5: Gv-OE	–.067	.394	.426	.365	—	
Factor 6: Gv-MC	–.175	.288	.444	.321	.265	—

Note. MC = multiple-choice. OE = open-ended. Conf = confidence. Gc = crystallized intelligence. RPM = Raven's Progressive Matrices. Voc = vocabulary. Gv = general visualization ability.

TABLE 7
Correlations Between the Principal Components Identified in the Analysis

	1	2	3	4	5	6
			Multiple-choice			
1. Accuracy	—					
2. Confidence	.42	—				
3. Bias	–.46	.61	—			
			Open-ended			
4. Accuracy	.70	.38	–.23	—		
5. Confidence	.46	.76	.35	.61	—	
6. Bias	–.13	.55	.66	–.23	.62	—

Note. See text for details.

cy, confidence, and bias scores from the two question formats were computed. The resulting correlations suffered less from the experimental dependency that was present in the results of the factor analysis in which accuracy and confidence scores from the same tests had been used. Those correlations, which are presented in Table 7, clarified a possible relation between ability and confidence bias. For that reason, the correlations of primary interest were those between test accuracy (as assessed by each question format) and the level of bias (from the alternative format). In each instance, a small negative correlation was present (*r*s = –.13 and –.23). Those individuals who performed poorly on cognitive ability tests appeared to be slightly less likely to express accurate, unbiased, confidence judgments. That result is in agreement with the conclusions of Zakay and Glicksohn (1992). However, further consideration of Table 7 reveals that there was a considerably greater correlation between the confidence ratings and bias levels derived from alternative question formats (*r*s = .55 and .35). Whereas cognitive ability appeared to play some role in determining the accuracy of confidence judgments, the confidence trait per se was a more important determinant of bias.

Discussion

The Confidence Factor

The confirmatory factor analysis reported in Table 5 identified a factor that captured common variance and was attributable solely to the confidence ratings ascribed to all eight tests of cognitive ability. It is difficult to describe that factor in any way other than as a trait apparently mediating confidence in decision making. Because the factor loadings stemmed from a variety of abilities identi-

fied within Gf/Gc theory, the Confidence factor appeared to transcend any single facet of cognitive ability. It therefore appeared to be a substantive, independent factor. Those findings replicated similar results reported by Schraw, Stankov, and their respective colleagues. Indeed, Stankov and Pallier (2002) have recently demonstrated that such a factor cuts across a range of perceptual tasks. Thus, there can now be little reason to question that humans maintain a consistent relationship in their expression of confidence in the accuracy of their responses across a wide array of cognitive capabilities.

Confidence Bias and Intelligence

The evidence currently presented suggests that those who are able to perform well on tests are less likely to be overconfident. Those results imply that low confidence bias is to some (likely small) extent a result of superior test-taking ability, whereas overconfidence on cognitive tasks is partially the result of poor test-taking ability. On the other hand, consideration of the correlations presented in Table 7 suggests that accurate responses alone do not determine the accuracy of confidence judgments. The levels of correlation and amount of variance explained by the confidence component in the analyses support that argument.

Confidence and Question Format

The present finding—that bias scores were lower for open-ended tasks than they were for the multiple-choice tests that assess acculturated knowledge and visualization—is consistent with the prediction derived from the work of Koehler (1994). Possible explanations for that finding can be drawn from both the heuristics and biases and the ecological schools of thought. For the former, one might argue that multiple-choice questions invoke a bias that causes test takers to neglect to consider alternatives to the focal hypotheses. People are perhaps less likely to consider reasons why the chosen response might be incorrect when answering forced-choice questions, especially when one of the answers is known to be correct. Open-ended problems, on the other hand, provide more space for the consideration of alternatives and might thus elicit more appropriate self-assessment.

In the latter approach, following PMM theory, one could argue that the discrepancy between the conditional probability that the answer is correct (cue validity) and the true relative frequency of that answer's accuracy (ecological validity) varies according to the nature of the question (representative vs. nonrepresentative). Presumably, in multiple-choice questions, there might be what Stankov (1999, p. 324) has called "familiar attractors," for example, possible responses that appear extremely viable when, in fact, they are incorrect. In other words, there might be so-called tricky questions with misleading alternative answers. Open-ended questions, on the other hand, do not contain tricky answers and might thus be less misleading than multiple-choice questions. That does not mean that open-

ended questions cannot be misleading or nonrepresentative; rather, they are perhaps less likely to be so than are multiple-choice questions. Therefore, for all tests (except Raven's Progressive Matrices), a feasible explanation for the observed differences in confidence bias might be provided by either theoretical position.

Confidence and Question Content

A surprising outcome of Experiment 1 was the presence of a substantial bias score on the RPM. RPM is an abstract reasoning task intended to enable investigators to examine the eduction of relations and the cognition of figural relations (see, e.g., Carpenter, Just, & Shell, 1990; Horn & Noll, 1994). By definition, such tests cannot contain misleading questions; otherwise, they would lack construct validity. The cue and ecological validities must be the same.

Previous studies in our laboratory have indicated that very low confidence bias is the norm when undergraduate participants answer the RPM. Those results support the predictions of the PMM. The fact that there was virtually no difference between bias scores obtained from the two versions of that test suggests that the cue and the ecological validities remained unaffected by question format. Given the earlier reasoning, however, there should be almost perfect assessment of accuracy on those two tests. In fact, the mean bias scores reported in Table 2 revealed that assessments were not perfect in the current study. The main difference between the results of the present sample population and those of previous undergraduate participants was a lower test-accuracy score of the present group, a finding that adds further support for a proposed (but relatively small) relation between ability and overconfidence.

Time Constraints

All the tests in Experiment 1 were administered under time constraints. Although the results of a pilot study indicated that the imposed restrictions were generous, it was possible that they affected the outcome. In Experiment 2, to be described next, the participants were allowed unlimited time to complete the tasks (with one exception, to be noted shortly). This possible confound was thus eliminated from the second study.

EXPERIMENT 2

Rationale

Personality and confidence. Although the outcome of Experiment 1 helped to clarify the role of intelligence in the confidence trait, in Experiment 2, we investigated the possible intrusion of personality factors. Because there is a paucity of

information regarding the personality correlates of confidence judgments, scales that, at least intuitively, might be expected to reveal those conjectured relationships were chosen for Experiment 2. To that end, we included in the research design the Extraversion subscale of the NEO PI-R (Costa & McCrae, 1992), the Self-Monitoring Scale (Lennox & Wolfe, 1984), and the Proactiveness subscale of the True Self-Report Inventory (Irvine, 1999).

The NEO PI-R assesses personality within the framework of the Big-Five model of personality, which is currently regarded as the most efficacious theoretical description of personality constructs (Davies, Stankov, & Roberts, 1998). Persons who score high on the Extraversion subscale are variously claimed to lead active, fast paced lives; to be dominant and forceful; to speak without hesitation; and to be high spirited and optimistic (Costa & McCrae, 1992). The factorial structure of the Extraversion subscale includes the following personality constructs: warmth, gregariousness, assertiveness, activity, excitement-seeking, and positive emotions. The Extraversion scale was chosen because those subcomponents appear, on face value, to be reasonable candidates for assessing a personality–confidence relation.

Lennox and Wolfe's (1984) Self-Monitoring Scale has been used in the investigation of proposed relations between confidence in personality tests and the trait of self-monitoring. For example, Cutler and Wolfe (1989) found a significant positive correlation between self-monitoring and an individual's level of confidence in self-reporting his or her personality type. We therefore included that scale in the present design so that we could investigate the possibility that self-monitoring might be related to confidence ratings obtained from more general ability tests.

Finally, it is claimed that Irvine's (1999) proactiveness scale assesses such personality features as determination and decisiveness. Because the questionnaire contains items asking how self-confident and assured a person is, the scale appeared to tap tendencies of relevance to the current investigation.

The accuracy of confidence judgments across domains. Stankov and Pallier (2002) have examined the accuracy of confidence judgments across various sensory modalities. They concluded that, in general, perceptual tasks presented only in the visual modality produce underconfidence (as expressed by the bias score). That conclusion is contrary to the view expressed by some commentators (see, e.g., Juslin, Olsson, & Winman, 1998) who have found that underconfidence is exhibited in sensory tasks in general. One of our purposes in Experiment 2 was to replicate the work of Stankov and Pallier, but with the inclusion of a greater number of cognitive tasks than the two used in their original study. To do so, we included in the battery of tests a number of cognitive ability factors identified within the framework of Gf/Gc theory but rarely (if ever) used in calibration research. The design included indices of short-term memory (short-term apprehension retrieval [SAR]), speed of mental processing (in particular, correct decision speed [CDS]), broad auditory ability (Ga), and olfactory memory (see Dan-

thiir, Roberts, Pallier, & Stankov, 2001). Markers of fluid and crystallized ability, actually the same multiple-choice general knowledge and RPM tests used in Experiment 1, were also presented to the participants.

To further assess the generality of the confidence trait identified in Experiment 1, we also included a number of discrimination tasks in the experimental design. Those measures of perceptual acuity were presented in the visual, auditory, and olfactory sensory modalities. Because that design has never been used in previous research, there was no model on which to base a confirmatory factor analysis (as was the case in Experiment 1). Therefore, an exploratory factor analysis was considered the most appropriate technique for use in Experiment 2.

Short-term apprehension retrieval, mental speed, and the confidence paradigm. Juslin and Olsson (1997) have proposed a sensory sampling model for the calibration of confidence. According to the model, there is a short-term memory window in which a number of comparison samples of perceptual judgments are stored. Horn and Noll (1994) have defined short-term memory (SAR, in the parlance of Gf/Gc theory) as the ability "measured in a variety of tasks that mainly require one to maintain awareness of, and be able to recall, elements of immediate stimulation, that is, events of the last minute or so" (p. 173). It is worthwhile to point out here that SAR does not require the manipulation of such information, the process that differentiates SAR from working memory (Kyllonen & Christal, 1990). Furthermore, it is suggested in the sensory sampling model of Juslin and Olsson that the length of time taken to reach a decision correlates negatively with the confidence in that decision. Indeed, Baranski and Petrusic (1994) found a negative correlation between perceptual confidence judgments and decision time, which they interpreted in terms of an accumulator model (cf., Vickers, 1979). That possibility is in line with the suggestion by Jensen (1993) that greater mental speed allows more comparison time before the decay of information within the short-term memory store and thus more efficacious processing of such information (cf. Roberts & Stankov, 1999). If the model proposed by Juslin and Olsson is viable, then there may be a meaningful relationship between the accuracy of confidence judgments and SAR, mental speed, or both. Because the design of Experiment 2 included several putative measures of short-term memory and mental speed, correlations between those measures and the accuracy of confidence judgments were examined.

Method

Participants

Undergraduate psychology students (N = 107; 63 women and 44 men) from The University of Sydney participated in the study as part of their course requirements. Their mean age was 19.81 years (SD = 3.99 years), and 77% spoke English as their first language.

Test Descriptions

In total, 16 measures were administered. Those included three personality questionnaires (administered in paper-and-pencil format), nine intelligence tests, and four perceptual tasks that were variously presented in different sensory modalities. We administered all the ability and perceptual tests except the olfactory tasks by computer. In contrast to Experiment 1, there was no time constraint placed on any of the tests, except for RPM, which had a very generous 12-min time limit. Instructions and an example were provided before each task in a manner similar to the procedures in Experiment 1. The times taken to provide a response on the computerized tasks were recorded.

After each test item (with the exception of the two digit-span tasks that are described shortly), the participants were instructed to indicate how confident they were that they had supplied the correct response. Because there were a number of difficulty levels in the memory tasks and one olfactory test was open-ended, the lower limit of the confidence rating scales varied depending on the chance level of a correct response. In accord with the first study, the participants were instructed that the lowest possible rating was indicative of *just guessing,* whereas 100% meant that one was *absolutely certain* that the correct response had been supplied. In questions having five alternatives, 20% was the lower limit; in questions having four possible responses, 25% was the lower bound, and so on. As discussed earlier, the confidence scale for the open-ended test ranged from 0% to 100%.

Psychometric Tests

Descriptions of the nine intelligence and four perceptual tests administered to all participants follow. For the sake of brevity, tests reported elsewhere have abbreviated descriptions, and the reader is directed to the source material for full details.

1. *General Knowledge Test.* This test was identical to Test 3 of Experiment 1, except that it contained two extra items and was administered on computer without time limits.

2. *Raven's Progressive Matrices.* This test was identical to Test 7 of Experiment 1, except that two additional items were included. In addition, the test was computerized and had a 12-min time limit.

3. *Digit Span Forward (DSF).* This test was adapted from the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981) and was administered in computer format. Participants were required to remember sequences of digits presented on a computer screen at 1-s intervals and to then type in the digits in the order in which they were presented. The score was the number of digits in the longest string answered correctly, an index of performance traditionally used with this task. This test (and the one that follows) defines a second-order, SAR factor (Horn & Noll, 1994).

4. *Digit Span Backward (DSB).* Also from the WAIS-R, this test was identical to Test 3, except that participants were required to type the numbers in the reverse order in which they were presented.

5. *Tonal Memory Test.* This 16-item test was based on the Tonal Memory Test from Seashore's Measures of Musical Talent (Seashore, Lewis, & Saetveit, 1960) and is described fully in Danthiir et al. (2001). Participants were presented with a series of between four and seven tones, and the task requirement was to identify the serial position of an altered tone in a second presentation. The test is a marker of the primary factor Discrimination Among Sound Patterns (see Horn & Stankov, 1982; Stankov & Horn, 1980).

6. *Symbol Memory Test.* This test contained 16 items and was newly devised by Danthiir et al. (2001) to be a visual analogue of the Tonal Memory Test (Test 5). We postulated that this test would measure the primary factor of Visual Memory (see Carroll, 1993) and would load on a Short-Term Memory factor analogous to SAR (see Horn & Noll, 1994).

7. *Odor Memory Test.* This 16-item test was also newly designed by Danthiir et al. (2001) to be analogous to the tonal and symbol memory tests, but in the olfactory modality. The stimuli were two series of microencapsulated odors, with one odor different in the second presentation. As in Tests 5 and 6, the task requirement was to identify the serial position of the changed stimulus.

8. *Multiple-Choice Smell Identification Test.* The stimuli used in this test were 10 microencapsulated odors taken from the 40-item version of the University of Pennsylvania Smell Identification Test (UPSIT; Doty, 1995). Participants were required to release an odor and to choose, from four alternatives, which of the alternatives the released odor smelled most like. Danthiir et al. (2001) have provided a full description of Test 8.

9. *Open-Ended Smell Identification Test.* This test was an experimental manipulation of Test 8. The stimuli consisted of 10 microencapsulated odors from the UPSIT that had not been used in the multiple-choice version, and the response format was open-ended (see Danthiir et al., 2001, who used a similar protocol).

10. *Line Length Test.* The 20-item test used here was derived from Stankov and Crawford (1996a) and consisted of the simultaneous presentation of five vertical, nonaligned lines. Participants had to determine which was the longest line. Elsewhere, Test 10 has been recognized as a marker for the second-order factor General Visualization Ability (Gv; see Horn & Stankov, 1982).

11. *Square Gaps Test.* The test used here contained 20 items and was developed by Stankov and Pallier (2002). The stimuli were five squares with gaps in the top line. Participants had to report which of the squares had the largest gap.

12. *Pitch Discrimination Test.* This task (20 items), which was derived from Stankov and Horn (1980), has been used extensively in our laboratory. Five sounds were presented in each trial—four having the same pitch and the fifth varying from the others. The task requirement was to identify the serial position of the different tone (see Danthiir et al., 2001, for a full description).

13. *Odor Discrimination Test.* The stimuli were presented as forced-choice triangle tests, and participants were required to identify which odor, at varying levels of intensity, was different from the other two (see Stankov & Pallier, 2002, who used a similar technique). Test 13 consisted of 10 trials and was conducted on a face-to-face basis, with the experimenter recording responses for both confidence and accuracy.

Personality Measures

The personality measures were presented as questionnaires in paper-and-pencil format. Standardized instructions preceded each of those tests.

14. *Proactiveness.* This questionnaire consisted of proactiveness items from the True Self-Report Inventory (Irvine, 1999). It required self-report responses to statements such as, "I am self-confident, assured," along a 6-alternative Likert-type scale. The 6 response alternatives consisted of the following: *never, rarely, sometimes, often, usually,* and *always.* The questionnaire contained 15 items in all.

15. *Self-monitoring.* This 13-item questionnaire, devised by Cutler and Wolfe (1989), has a somewhat moderate reported reliability of .75. We therefore modified the items to remove qualifiers from the question stem, in accordance with the suggestion of Newstead and Collis (1987). For example, "I can usually tell when others consider a joke to be in bad taste" was changed to "I can tell when others consider a joke to be in bad taste." That modification allowed items to be responded to (and scored) on the same 6-point rating scale as that used in Test 14.

16. *Extraversion.* This scale contained 48 items from the Extraversion subscale of the NEO PI-R (Costa & McCrae, 1992). In addition to the second-order Extraversion factor of the NEO PI-R, six first-order factors (i.e., facets) were measured. Scores were calculated separately for the overall factor and for the facets of that well-known scale.

Procedure

The tests were administered to groups of up to 5 participants. Total testing time was approximately 4 hr; however the amount of time varied because most of the tests were self-paced. Testing was conducted in two 2-hr sessions that were usually separated by a week, with a break of 15 min after 1 hr of work. The participants were first advised of the test protocol and ethical requirements. Before each test, instructions for the particular task were presented, along with examples and practice items. Before commencing each test, the participants were encouraged to inform the experimenter of any queries. Two proctors were present at all times.

The computerized tasks were presented in the first test session, whereas the personality measures and olfactory tests were administered in Session 2. We spread olfactory discrimination tasks as far apart as possible to help minimize adaptation.

Results

Reliabilities

Where possible, reliabilities (Cronbach's alpha) for the accuracy scores, confidence ratings, and response times were calculated; they are presented in Table 8.

Inspection of Table 8 reveals that there were relatively low reliabilities for some of the tasks. In comparison to the reliability scores presented by Stankov and Pallier (2002, Table 4) some of the tasks (e.g., line length) had lower reliabilities. However, others (e.g., odor discrimination) showed considerably better reliability levels than those reported by Stankov and Pallier. Using the criteria suggested by Guilford and Fruchter (1978), we considered that all the reliability scores in Table 8, although on the low end of the acceptable range, were satisfactory for experimental purposes.

Descriptive Statistics for Accuracy, Confidence, Bias, and Response Times

The means and standard deviations of accuracy, confidence, bias, and response times of the cognitive and perceptual tasks used in Experiment 2 are presented in Table 9. Examination of Table 9 reveals that the tasks were sufficiently difficult and contained sufficient variance. Furthermore, most scores on established marker tests were similar to those previously reported in our laboratory (see, e.g., Davies, et al.,

TABLE 8
Reliabilities (Cronbach's alpha) for the Variables Obtained from Experiment 2

Measure	Accuracy	Confidence	Reaction time
RPM	.87	.90	.77
Line length	.48	.94	.89
Square gaps	.42	.94	.86
Pitch discrimination	.77	.94	.87
Odor discrimination	.48	.89	
Symbol memory	.56	.86	.59
Tonal memory	.60	.86	.66
Odor memory	.60	.89	
Multiple-choice odors	.52	.88	
Open-ended odors	.69	.89	
Proactiveness	.67		
Self-monitoring	.82		
Extraversion	.78		

Note. RPM = Raven's Progressive Matrices. It was not possible to compute reliabilities for the General Knowledge Test (Test 1) because of a computer malfunction.

TABLE 9
Means and Standard Deviations for the Ability and Perceptual Tasks

Test	Accuracy (% correct)		Confidence (Average %)		Bias score		Response time (s)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
General knowledge	58.44	10.79	68.14	13.01	9.70	12.36	9.65	3.75
RPM	70.91	18.82	75.15	15.90	4.37	15.91	30.96	13.75
Digit span forward	6.47	1.33					5.31	1.84
Digit span backwards	6.12	1.54					12.84	6.07
Tonal memory	62.63	15.69	69.07	13.43	6.49	14.59	2.10	1.02
Symbol memory	82.20	12.93	80.72	12.17	–1.18	11.32	1.84	0.08
Odor memory	52.73	17.61	60.02	14.50	7.29	19.44		
Multiple-choice smell	71.36	17.10	75.35	12.18	4.25	19.17		
Open-ended smell	28.45	18.40	51.72	17.15	16.61	21.25		
Line length	76.88	11.24	63.46	13.55	–13.41	16.30	7.10	3.60
Square gaps	61.95	12.79	56.47	13.78	–5.18	18.82	7.18	3.49
Pitch discrimination	64.48	20.37	61.31	18.55	–3.11	13.01	8.88	9.06
Odor discrimination	53.64	20.68	72.01	12.16	18.51	21.12		

Note. RPM = Raven's Progressive Matrices. Blank spaces indicate that the data were not obtained from these measures (see text). Digit span scores are for the level achieved.

1998; Pallier et al., 2000), indicating that the current participants were typical of an Australian undergraduate sample. Thus, although the DSF (low) and the DSB (high) scores were somewhat unusual, they were not without precedent. Similarly, the mean accuracy score for Test 9 (28.45%) indicated that the test was rather difficult, but results fell within the range reported in Larsson's (1997) review of studies of unaided odor identification by young adults.

Comparison between the two studies of the bias scores for general knowledge and Raven's Progressive Matrices indicated that the student sample was better calibrated than was the military sample. Furthermore, the visualization tasks used in Experiment 2 exhibited underconfidence, whereas those of Experiment 1 (which contained a cognitive component) showed overconfidence. That outcome suggests that the intrusion of cognitive demands in visual-perceptual tasks mediates the accuracy of confidence judgments, a possibility mentioned in Experiment 1.

Underconfidence was markedly present in the Line Length Test (Test 10), whereas pitch discrimination (Test 12) showed reasonably good calibration. Conversely, odor discrimination (Test 13) had the highest bias score of all the tasks in this battery. Most important, the findings of Stankov and Pallier (2002) were generally replicated in Experiment 2, which suggests primarily that only perceptual judgments made in the visual modality are likely to produce a consistent underconfidence phenomenon.

Correlations Between the Tasks

In Table 10, we present the correlation matrix for the variables used in the ensuing factor analysis. Those correlations were as predicted, except for the low correlations between odor discrimination and the other olfactory measures and the exceptionally high correlation (r = .78) between accuracy and confidence on pitch discrimination. Such large matrices are difficult to interpret, and the data-reduction technique of factor analysis provided a convenient method of clarifying the results presented in Table 10.

Exploratory Factor Analysis: Evidence for the Generality of the Confidence Trait

To determine the generality of the confidence factor identified in Experiment 1 across a range of cognitive and perceptual tasks, we subjected the correlation matrix of Table 10 to an exploratory factor analysis, using principal axis factoring with Promax rotation. The root-one criterion allowed for the extraction of nine factors; however, the scree plot (Cattell, 1966) was somewhat indeterminate after seven factors. Because eight or nine factors produced Heywood cases (using maximum likelihood procedures), singlet factors, or both, a solution involving seven factors was preferred. The results are presented in Table 11 (see page 289). Interpretation of the seven-factor solution follows.

Factor 1: Confidence. The only salient loadings on this factor were derived

from confidence scores obtained from all but one of the tasks in the battery. This factor clearly represented confidence, generalized across a wide range of both cognitive and perceptual tasks. That result also provided strong evidence for the role of the confidence trait in determining the accuracy of confidence judgments. Confidence ratings obtained from pitch discrimination did not contribute to this factor, probably because of the high correlation between accuracy and confidence on that task.

Factor 2: Visualization/Fluid Intelligence. This factor was defined by loadings from the visual ability tasks (Tests 10 and 11) and from the Raven's matrices test (Test 2). The failure to differentiate between measures of Gf and Gv in batteries containing a range of perceptual and cognitive tasks has been reported previously in the literature and appears to be especially the case when stimuli are presented in several sensory modalities (see, e.g., Pallier et al., 2000; Roberts, Stankov, Pallier, & Dolph, 1997). Thus, the structure of Factor 2 was in line with expectations.

Factor 3: General Knowledge. The only loadings on this factor were from measures of general knowledge (Test 1); this factor therefore clearly represented levels of this specific acculturated ability expressed by the student sample.

Factor 4: Short-Term Memory. All the tests loading on this factor were representative of short-term memory tasks, as defined within the theory of fluid and crystallized ability (see Horn & Noll, 1994). Factor 4 might therefore be unequivocally described as an SAR factor. Of interest, odor memory (Test 7) had a low, nonsalient loading on this factor, which provided support for recent suggestions in the literature that memory for odors is fundamentally different from memory processes in which other sensory modalities are used (see, e.g., Annett, 1996; Herz & Engen, 1996; White, 1998). That outcome in the present study was therefore not without precedent.

Factor 5: Mental Speed. This factor was defined solely by the time of providing responses to the computerized tasks. Thus, it clearly represented a Speed-of-Test-Taking factor, which has previously been related to Correct Decision Speed (at the first-order) and General Mental Speed at a higher-order of analysis (see Roberts & Stankov, 1999). That finding too was in line with expectations. Hence, Stankov and Pallier (2002) noted, "if accuracy, confidence ratings, and speed scores from each test are included in a single factor analysis, confidence scores will define a single factor and speed scores will also define a single factor" (p.16).

Factor 6: Olfactory Memory. The highest loadings here were from the tests presented in the olfactory modality that required participants to use some form of memory process to obtain a correct solution (i.e., Tests 7, 8, and 9). Danthiir et al. (2001) recently reported a similar finding.

Factor 7: Auditory Ability. This factor was poorly defined within the present battery because it had loadings from only two components. However, pitch discrimination (Test 12) has been identified elsewhere (Horn & Stankov, 1982) as a marker for the primary factor Discrimination Amongst Sound Patterns, which

TABLE 10
Correlation Between Variables of Experiment 2

Measure	1	2	3	4	5
1. General knowledge accuracy	—				
2. General knowledge confidence	.47	—			
3. General knowledge time	–.05	–.15	—		
4 RPM accuracy	.14	.05	.06	—	
5. RPM confidence	.15	.47	.15	.59	—
6. RPM time	.09	.15	.18	.29	.32
7. Digit span forward accuracy	.00	.01	.00	.15	.14
8. Digit span forward time	.12	–.03	.30	.11	.14
9. Digit span backward accuracy	–.12	.01	.10	.20	.20
10. Digit span backward time	.02	.03	.32	.08	.16
11. Tonal memory accuracy	–.06	–.08	.13	.42	.27
12. Tonal memory confidence	.00	.37	.10	.20	.54
13. Tonal memory time	.22	.09	.24	.01	–.04
14. Symbol memory accuracy	.03	.12	–.13	.29	.16
15. Symbol memory confidence	.11	.38	.10	.26	.52
16. Symbol memory time	–.05	–.05	.27	.08	.13
17. Odor memory accuracy	.00	.02	.11	.41	.19
18. Odor memory confidence	.04	.50	.08	.10	.41
19. Multiple-choice smell accuracy	.15	.04	–.08	.24	.11
20. Multiple-choice smell confidence	.02	.35	.08	.12	.44
21. Open-ended smell accuracy	.03	.00	.20	–.02	–.06
22. Open-ended smell confidence	–.04	.37	.08	–.06	.25
23. Line length accuracy	.05	.09	.04	.38	.19
24. Line length confidence	.03	.35	.12	–.04	.42
25. Line length time	.03	.05	.27	.16	.12
26. Square gaps accuracy	.12	–.01	.01	.39	.25
27. Square gaps confidence	.13	.41	.06	–.04	.40
28. Square gaps time	–.05	.06	.24	.20	.16
29. Pitch discrimination accuracy	–.05	–.03	.07	.23	.05
30. Pitch discrimination confidence	.05	.21	.11	.12	.23
31. Pitch discrimination time	.13	.12	.20	.29	.20
32. Odor discrimination accuracy	.12	.11	–.01	.15	.19
33. Odor discrimination confidence	–.03	.41	.13	.06	.41

Note. RPM = Raven's Progressive Matrices.

loads onto General Auditory ability (Ga) at the second-order of analysis. This factor was therefore (albeit cautiously) identified as such in the present analysis.

The solution just reported contained no loading from the odor discrimination task (Test 13) accuracy score, which is not surprising given this test's very low communality (h^2 = .11) with the other tests in the battery. Test 13 thus appeared relatively independent from the other tasks, suggesting that Factor 6 (Olfactory Memory) is not reliant on odor discrimination for its

6	7	8	9	10	11	12	13	14
—								
–.03	—							
.03	.11	—						
.13	.31	.21	—					
.30	–.01	.45	.35	—				
–.02	.46	.02	.32	–.02	—			
.05	.17	.02	.19	.03	.50	—		
.20	–.17	.17	–.09	.28	–.15	–.09	—	
–.04	.35	.00	.17	–.11	.32	.17	–.16	—
.05	.30	.11	.24	.02	.27	.54	–.08	.60
.25	–.15	.42	–.04	.38	–.15	.00	.52	–.19
.07	.11	.12	.14	.07	.20	.00	.00	.30
–.07	–.14	.05	.01	–.03	.08	.49	–.01	.14
.09	.03	.08	.02	.00	–.13	–.19	–.12	.26
–.12	–.08	.06	.09	–.12	.11	.47	–.22	.11
–.08	–.14	.00	.13	.07	–.04	–.07	.01	–.06
.00	–.12	–.03	.07	–.04	.08	.49	–.03	–.01
.04	.10	–.05	.25	–.03	.28	.11	.00	.16
–.01	.04	–.10	.13	–.07	.15	.61	–.14	.09
.26	–.14	.41	–.01	.52	–.05	–.01	.48	–.07
.20	.16	.01	.19	–.04	.14	.10	–.05	–.02
.00	–.08	–.05	.01	.06	.00	.46	–.02	.02
.41	–.10	.25	.02	.32	.02	.02	.19	–.08
–.12	.12	.00	.11	–.01	.34	.26	.02	.16
–.09	.12	–.04	.09	.00	.26	.48	.01	.08
.11	.06	.31	.14	.22	.08	.12	.11	.00
.11	–.11	.07	.17	.06	.21	.12	.02	.15
–.02	–.21	–.01	–.03	–.04	–.01	.53	–.05	.05

(table continues)

structure. Nonetheless, confidence ratings obtained from Test 13 helped to define Factor 1, indicating the robustness and independence of that metacognitive trait.

Factor Intercorrelations: Is There a Relationship Between Confidence, Mental Speed, and SAR?

Table 12 (see page 291) shows the intercorrelations between the factors reported in Table 11. Overall, those correlations were comparable with those

TABLE 10 (Continued)

Measure	15	16	17	18	19
1. General knowledge accuracy					
2. General knowledge confidence					
3. General knowledge time					
4 RPM accuracy					
5. RPM confidence					
6. RPM time					
7. Digit span forward accuracy					
8. Digit span forward time					
9. Digit span backward accuracy					
10. Digit span backward time					
11. Tonal memory accuracy					
12. Tonal memory confidence					
13. Tonal memory time					
14. Symbol memory accuracy					
15. Symbol memory confidence	—				
16. Symbol memory time	–.11	—			
17. Odor memory accuracy	.19	.01	—		
18. Odor memory confidence	.49	–.04	.28	—	
19. Multiple-choice smell accuracy	.06	.01	.29	.00	—
20. Multiple-choice smell confidence	.43	–.15	.15	.72	.19
21. Open-ended smell accuracy	–.01	–.13	.25	.15	.02
22. Open-ended smell confidence	.37	–.09	.03	.65	–.18
23. Line length accuracy	.20	–.08	.33	.05	.06
24. Line length confidence	.53	–.11	.01	.50	–.10
25. Line length time	–.08	.57	.10	.05	–.06
26. Square gaps accuracy	.05	–.02	.23	.05	.10
27. Square gaps confidence	.36	–.01	–.06	.47	–.03
28. Square gaps time	–.02	.39	.15	.01	.00
29. Pitch discrimination accuracy	.11	–.03	.10	.06	–.09
30. Pitch discrimination confidence	.23	–.02	–.09	.34	–.20
31. Pitch discrimination time	.14	.28	.20	.11	.03
32. Odor discrimination accuracy	.25	–.01	.12	.23	.14
33. Odor discrimination confidence	.45	.09	.05	.68	.00

Note. RPM = Raven's Progressive Matrices.

reported by Stankov and Pallier (2002) and were thus not contentious. The factor intercorrelations allowed for an assessment of our two secondary aims in Experiment 2. First, because there was virtually zero correlation (r = .05) between confidence (Factor 1) and mental speed (Factor 5), it appears reasonable to conclude that there is no relation between mental speed and the confidence trait. On the other hand, there was a moderate correlation (r = .33) between the confidence and the SAR factors; thus, there remains the possibili-

20	21	22	23	24	25	26	27	28
—								
.21	—							
.58	.29	—						
.07	.05	.05	—					
.55	.01	.46	.14	—				
–.12	.01	–.01	.01	–.24	—			
.09	.01	.02	.28	.05	–.07	—		
.48	–.06	.41	.05	.59	–.07	.03	—	
–.04	.10	.00	.10	–.15	.57	.01	–.19	—
.16	.01	.00	.23	.08	.05	.13	.06	.00
.29	.02	.22	.18	.34	.03	.10	.40	.00
.09	.02	–.01	.11	.11	.26	.09	.02	.26
.16	–.10	.13	.17	.10	.00	.03	.23	.11
.61	.09	.61	.06	.54	.04	–.07	.56	.09

(table continues)

ty that short-term memory ability might have an effect on the accuracy of confidence ratings. It is important to reiterate, however, that there were no salient loadings from any of the SAR measures on the Confidence factor (Factor 1), and any relationship between SAR and the bias score was likely to be rather weak

Nonetheless, in the correlational analyses that follow, we included the digit-span tests along with the personality measures in an attempt to further clarify any possible interaction between memory span and confidence.

TABLE 10 (Continued)

Measure	29	30	31	32	33
1. General knowledge accuracy					
2. General knowledge confidence					
3. General knowledge time					
4 RPM accuracy					
5. RPM confidence					
6. RPM time					
7. Digit span forward accuracy					
8. Digit span forward time					
9. Digit span backward accuracy					
10. Digit span backward time					
11. Tonal memory accuracy					
12. Tonal memory confidence					
13. Tonal memory time					
14. Symbol memory accuracy					
15. Symbol memory confidence					
16. Symbol memory time					
17. Odor memory accuracy					
18. Odor memory confidence					
19. Multiple-choice smell accuracy					
20. Multiple-choice smell confidence					
21. Open-ended smell accuracy					
22. Open-ended smell confidence					
23. Line length accuracy					
24. Line length confidence					
25. Line length time					
26. Square gaps accuracy					
27. Square gaps confidence					
28. Square gaps time					
29. Pitch discrimination accuracy	—				
30. Pitch discrimination confidence	.78	—			
31. Pitch discrimination time	.23	.32	—		
32. Odor discrimination accuracy	–.05	.01	–.03	—	
33. Odor discrimination confidence	.02	.27	.07	.27	—

Note. RPM = Raven's Progressive Matrices.

Confidence, Bias, Personality, and SAR: Evidence From a Correlational Analysis

To assess the relationship of short-term memory and personality measures with the confidence and bias scores, we subjected the latter two indices to a principal components analysis. Only the first principal component was retained in each case. We then correlated those components, using Pearson product–moment correlations, to the indices derived from the three personality measures and to a composite score from the two digit-span tasks. Recall that confidence ratings

TABLE 11
Exploratory Factor Analysis (Principal Axis Factoring Extraction With Promax Rotation) of the Variables Measured in Experiment 2

Measure	Factor 1	2	3	4	5	6	7	h^2
1. General knowledge accuracy			.54					.28
2. General knowledge confidence	.45		.55					.57
3. General knowledge time			–.30		.46			.29
4. RPM accuracy		.71				.32		.69
5. RPM confidence	.39	.46						.72
6. RPM time		.44						.38
7. Digit span forward accuracy				.68				.43
8. Digit span forward time					.63			.38
9. Digit span backward accuracy				.40				.29
10. Digit span backward time					.68			.45
11. Tonal memory accuracy				.55				.55
12. Tonal memory confidence	.56			.30		–.31		.74
13. Tonal memory time					.51			.36
14. Symbol memory accuracy				.71		.35		.62
15. Symbol memory confidence	.47			.61				.73
16. Symbol memory time					.72			.53
17. Odor memory accuracy						.60		.47
18. Odor memory confidence	.87							.75
19. Multiple-choice smell accuracy						.44		.29
20. Multiple-choice smell confidence	.79							.64
21. Open-ended smell accuracy						.35		.28
22. Open-ended smell confidence	.72							.57

(table continues)

TABLE 11 (Continued)

Measure	Factor 1	2	3	4	5	6	7	h^2
23. Line length accuracy		.46						.30
24. Line length confidence	.69							.58
25. Line length time					.82			.66
26. Square gaps accuracy		.60						.28
27. Square gaps confidence	.63							.51
28. Square gaps time					.53			.37
29. Pitch discrimination accuracy							.84	.72
30. Pitch discrimination confidence							.89	.94
31. Pitch discrimination time					.35			.25
32. Odor discrimination accuracy								.11
33. Odor discrimination confidence	.84							.62

Note. RPM = Raven's Progressive Matrices. Only salient loadings (above 0.30) are reported.

were not obtained for the digit-span tasks. Therefore, those measures were independent of the overall confidence and bias scores, and thus could not introduce a statistical artifact into the analysis. The results are presented in Table 13.

First, inspection of Table 13 indicates that there was virtually zero correlation between SAR and both confidence ($r = -.03$) and bias scores ($r = -.09$). Thus, if there was a relationship between short-term memory capacity and the accuracy of confidence judgments, it was very weak.

TABLE 12
Factor Intercorrelations Between the Factors Identified in Table 11

Factor	1	2	3	4	5	6	7
1. Confidence	—						
2. Visualization/fluid intelligence	.27	—					
3. General knowledge	.19	.05	—				
4. Short-term memory	.33	.42	–.02	—			
5. Mental speed	.05	.32	.11	–.05	—		
6. Olfactory memory	–.06	–.11	–.04	–.01	–.03	—	
7. Auditory ability	.24	.15	–.15	.19	–.02	–.01	—

TABLE 13
Pearson Product–Moment Correlations Between the Measures of Personality and the Confidence and Bias Scores

	First principal component obtained from the:	
	Confidence ratings	Bias scores
Personality measure		
Proactiveness	**.35**	**.29**
Self-monitoring	.18	.11
Extraversion	.08	.15
Warmth	–.03	.06
Gregariousness	–.02	.11
Assertiveness	.10	.11
Activity	**.26**	**.27**
Excitement-seeking	.04	.05
Positive emotions	.01	.08
Short-term memory		
Average digit span	–.03	–.09

Note. Figures in bold indicate significance at $p < .01$.

Second, and more importantly, there was a significant correlation between some of the personality measures and both confidence and bias scores. The personality construct, or constructs, underlying both Irvine's proactiveness measure and the Activity facet of the NEO PI-R appeared to contribute to the accuracy of confidence judgments (see Table 13). One might therefore need to consider some aspect of personality when explanations of the over- or underconfidence phenomenon are presented. There was no significant correlation between Lennox and Wolfe's (1984) Self-Monitoring Scale and either confidence or bias scores. It therefore seems reasonable to conclude that self-monitoring (as measured by that scale) is not related to confidence ratings obtained from a mixture of perceptual and cognitive ability tasks.

Discussion

The Accuracy of Confidence Judgments and Personality

Only a small range of personality measures were used in Experiment 2. It would therefore be imprudent to draw any firm conclusions regarding the potential impact of personality on bias scores. The measures of personality used were somewhat speculatively chosen subcomponents of instruments that appeared likely to capture some of the variance common to the confidence factor and the bias score. In terms of a more comprehensive assessment of the personality domain, however, those measures might be inadequate. Notwithstanding, there was an indication that there was a small relationship between the confidence factor and the personality constructs of proactiveness and activity.

The Robustness of the Confidence Trait Across Perceptual and Cognitive Abilities

The replication of the study by Stankov and Pallier was evidenced in the empirical results. The confidence trait remained an independent factor in a battery of tests that included a variety of tasks involving both cognitive and perceptual ability. These results are difficult to interpret in any way except as strong evidence that the accuracy of confidence judgments is moderated by a metacognitive trait that lies on the boundary between personality and intelligence.

The Role of Short-Term Memory and Mental Speed in the Accuracy of Confidence Judgments

The results of Experiment 2 indicated that superior mental speed is unlikely to affect the magnitude of bias scores. Similarly, the evidence presented in Table 13 suggested that SAR does not play a major role in determining confidence levels. On the other hand, the factor intercorrelations suggested that those with better short-term memory scores are likely to express lower levels of confidence bias. A possible explanation for that seeming contradiction follows. There are claims

in the literature that fluid intelligence is the core component of intellectual ability (see, e.g., Gustafsson, 1992) and that fluid abilities are dependent, at least to some extent, on short-term memory (Kyllonen & Christal, 1990); the moderate correlation (r = .42) between Factors 2 and 4 in the present results provided support for that claim. The small correlation between intelligence and confidence bias identified in Experiment 1 might thus have been represented in Experiment 2 by the correlation between the confidence and short-term memory factors. That outcome could therefore be an indication of the slight influence of intelligence on bias scores, rather than the effect of SAR per se.

GENERAL DISCUSSION

Individual Differences in the Accuracy of Confidence Judgments

The tasks presented to participants in the two present experiments covered a wide range of cognitive and perceptual abilities. Questions were presented in different formats, had a number of response choices (e.g., three, four, or five alternative answers), did or did not have time restraints, required the use of different sensory modalities, and had both paper-and-pencil and computer formats. The application of both confirmatory and exploratory factor analytic techniques to the data obtained from this extensive testing procedure appeared decisive and unequivocal: Humans have a trait that mediates their ability to evaluate the accuracy of their responses. It is important to reiterate here that the present findings replicate those of several studies conducted by Schraw, Stankov, and their respective colleagues. Therefore, there can now be little reason to question empirical evidence that human beings maintain a consistent, and at times substantial, relationship in the expression of confidence in the accuracy of their responses across a wide array of capabilities.

The Nature of the Confidence Trait

The confidence trait identified in these studies has been shown to rely, to a relatively small extent, on an individual's cognitive ability. That finding provides an answer to Lichtenstein and Fischhoff's (1977) question: Those who know more, do know (slightly) more about how much they know. The trait is seemingly associated, again to a small extent, with some aspects of personality constructs. It thus appears to lie in what Stankov (1999) has called the no-man's-land between personality and intelligence. Although moderated to a small extent by those constructs, the confidence trait itself is a major determinant of the accuracy of self-assessment in a wide variety of tasks. The trait also appears to be relatively stable across a range of difficulty levels and is thus not a state-dependent variable. Evidence supporting that claim was presented in Tables 4 and 10; the correlations between confidence ratings in those tables are consistently higher than are those among the accuracy levels for the various tests (cf. Schraw, 1994, 1997).

Implications for Current Theories in Calibration Research

The results reported in this article are in agreement with a number of aspects of theories put forward by proponents of both the ecological and the heuristics and biases approaches. The support (or otherwise) seemingly depends on the type and the nature of the cognitive tasks involved. For instance, question format seemed to affect the accuracy of confidence judgments, as suggested by some heuristics and biases theorists. On the other hand, the lack of any meaningful difference between bias scores, as obtained from Raven's matrices task in Experiment 1, could be taken as support for the ecological model. Furthermore, if, as some of the results presented herein indicated, there is a small effect of short-term memory capacity on bias scores, other concepts, such as Juslin and Olsson's (1997) model of the accuracy of confidence judgments in sensory tasks, were also partially supported by the current results.

The Complexity of the Self-Assessment Paradigm: Toward a Consensual Model

The results of the experimental manipulations presented herein suggest that one needs a complex explanation to adequately account for the present findings. Probabilistic accounts, stemming from both the Brunswikian and Thurstonian traditions (see Juslin & Olsson, 1997) and from cognitive biases approaches seem to be only partially able to explain the under- and overconfidence phenomenon. An enhanced theoretical position, allowing a more flexible account, must encompass a number of causes affecting the accuracy of confidence judgments. Included in that account must be due consideration of the role of individual differences variables—cognitive ability, personality, and metacognitive processes—in determining the accuracy of confidence judgments. Accordingly, in a more convincing psychological account of the confidence paradigm all the aforementioned propositions should be considered as potential determinants.

The Utility of the Confidence Paradigm

It appears reasonable to ask, as Stankov and Dolph (2000) remarked, "what real-life behaviors might be predicted by self-confidence scores" (p. 224)? Those authors reported two instances of the predictive validity of confidence bias. First, "people with higher self-monitoring scores tend to be judged as better than their peers in the ability to perform tasks that require dealing with the public" (p. 224). Second, and somewhat disturbingly, they noted a significant correlation between high self-confidence scores and bad driving practices such as speeding and running red lights. Further investigation of those (and probably other) useful aspects of behavioral predictors from the confidence paradigm would seem a worthwhile endeavor.

Conclusion

The outcome of the studies reported in this article suggest that there might be a small relationship between cognitive ability, certain personality traits, and the accuracy of confidence judgments. The relationship is in need of further research. In fact, the authors are presently analyzing data that include confidence and accuracy scores from an expanded test battery (which includes the Armed Services Vocational Aptitude Battery), in addition to a measure of all factors comprising the Big Five factor model of personality and motivational constructs such as the need for cognition.

Given the evidence from these two experiments, it appears that the confidence trait is generalizable across many domains of behavior. However, there are a number of issues that require further investigation in that regard. For example, do predictions of future events subscribe to these findings? Does confidence in one's ability to perform successfully in sporting or related events follow the individual differences approach? Would a person alter his or her confidence ratings if betting on the outcome or other financial risks are involved, or is the confidence trait consistent across an expanded range of circumstances? Is the confidence trait stable over time? Are there gender and age differences within the confidence paradigm? Those are some of the questions that remain to be addressed, but the evidence accumulated so far indicates that the confidence trait is likely to be robust and to affect the accuracy of confidence judgments across a range of domains. All else aside, it appears that investigators would be remiss if, in their considerations within the confidence paradigm, they do not recognize the presence of individual differences in the outcome of their experimental designs.

NOTES

1. Investigators often determine the accuracy of confidence ratings by using a decomposition of the mean probability score (commonly known as the *Brier Score;* Brier, 1950). That procedure yields a calibration and resolution score, via the so-called Murphy partition (Murphy, 1973). Calibration, resolution, and Brier scores were all computed in the current study, but low reliabilities (e.g., for the Raven's Progressive Matrices resolution measure, Cronbach's $\alpha = .07$; similarly, for the vocabulary calibration measure, $\alpha = .12$) precluded their use in further analyses. Stankov and Crawford (1996a) reported similar results.

2. Soll (1996; cf. Erev, Wallsten, & Budescu, 1994;) has suggested "a random error extension of PMM" (p. 120). Random error in that model is characterized by "sampling variability in learning" rather than "cognitive inconsistency" (Soll, p. 122). According to that position, only two alternative forced-choice questions, where confidence is bounded by 50% and 100%, are currently considered. Moreover, it is maintained that even in a best-case scenario, where a person might provide responses that, on average, correspond to the cue probability, random error will sometimes result in the correct response, contrary to probability and vice versa. In reality, however, when because of random error the nonnormative answer is chosen, that choice will result in a greater number of incorrect responses (in the ratio of the cue probability) and thus in overconfidence. Furthermore, the confi-

dence scale, unlike the probability level, is restricted to a minimum of 50%, and therefore, Soll argued, reported confidence will, on average, be greater than ecological validity.

3. The broad framework for the selection of the cognitive tests was provided by the theory of fluid and crystallized intelligence (Gf/Gc; see Horn, 1998; Horn & Noll, 1994). For that purpose, we administered two tests of fluid (Gf) and four tests of crystallized (Gc) ability. In addition, two recognized markers of the broad visualization factor (Gv) were used. Note that the two Gv tests used demarcate different primary mental abilities but are expected to load on a common visualization factor at the second-order. Interestingly, previous findings of underconfidence in tests of Gv have all emanated from rather simple perceptual tasks. It remains unclear whether the inclusion of more cognitively demanding visualization tests will perhaps temper the typical underconfidence reported in visual-perceptual tasks to produce a better-calibrated result.

4. Because responses to the open-ended questions might be correct but not exactly the same as a standard response, participants' answers to those questions were vetted. A consensual scoring technique was applied to responses that fell in that category. For example, in the vocabulary test, *conventional* was considered the standard answer for *orthodox*. However, responses such as *conformist* and *customary* were also recorded as correct.

5. Note that, on the basis of pilot data, we allowed an extra minute (relative to Test 3) to accommodate the possibility of lengthy, handwritten responses from the participants.

REFERENCES

Annett, J. M. (1996). Olfactory memory: A case study in cognitive psychology. *Journal of Psychology, 130,* 309–319.

Arbuckle, J. L., & Wothke, W. (1999). *AMOS 4.0 user's guide.* Chicago, IL: SPSS Inc.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics, 55,* 412–428.

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception and Psychophysics, 61,* 1369–1383.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception and Psychophysics, 54,* 75–81.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 97,* 404–431.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, UK: Cambridge University Press.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40,* 153–193.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 161–179.

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R manual.* Odessa, FL: Psychological Assessment Resources.

Crawford, J., & Stankov, L. (1996a). Age differences in the realism of confidence judgments: A calibration study using tests of fluid and crystallized intelligence. *Learning and Individual Differences, 6,* 84–103.

Crawford, J., & Stankov, L. (1996b). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 6,* 971–986.

Cutler, B., & Wolfe, R. (1989). Self-monitoring and the association between confidence and accuracy. *Journal of Research in Personality, 23,* 410–420.

Danthiir, V., Roberts, R. D., Pallier, G., & Stankov, L. (2001). What the nose knows: Olfaction and cognitive abilities. *Intelligence, 29,* 337–361.

Davies, M., Stankov, L., & Roberts, R. D. (1998). Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology, 75,* 989–1015.

Doty, R. L. (1995). *The Smell Identification Test administration manual.* Haddon Heights, NJ: Sensonics, Inc.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 83,* 37–64.

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Manual for reference tests for cognitive factors.* Princeton, NJ: Educational Testing Service.

Fullerton, G. S., & Cattell, J. M. (1892). *On the perception of small differences.* Philadelphia, PA: University of Pennsylvania Philosophy Series, No. 2.

Gigerenzer, G. (1991). How to make cognitive illusion disappear: Beyond heuristics and biases. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 2, pp. 83–115). Chichester, UK: Wiley.

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

Gregory, R. J. (1996). *Psychological testing: History, principles, and applications.* Boston MA: Allyn and Bacon.

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

Griffing, J. H. (1895). On sensations from pressure and impact. *Psychological Review Monographs, 1,* 1–88.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education.* New York: McGraw-Hill.

Gustafsson, J-E. (1992). The relevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27,* 239–248.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Science, 1,* 78–82.

Herz, R. S., & Engen, T. (1996). Odor memory: Review and analysis. *Psychonomic Bulletin and Review, 3,* 300–313.

Horn, J. L. (1998). A basis for research on age differences in cognitive capabilities. In J. R. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice.* (pp. 57–91). London: Erlbaum.

Horn, J. L., & Cattell, R. B. (1966). Refinement of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57,* 253–270.

Horn, J. L., & Noll, J. (1994). A system for understanding cognitive capabilities: A theory and the evidence on which it is based. In D. K. Detterman (Ed.), *Current topics in human intelligence: Vol. IV: Theories of intelligence* (pp. 151–204). Norwood NJ: Ablex.

Horn, J. L., & Stankov, L. (1982). Auditory and visual factors of intelligence. *Intelligence, 62*(6), 165–185.

Irvine, S. H. (1999, May 27). *The True Self-Report Inventory.* Workshop presented at The University of Sydney, Australia.

Jensen, A. R. (1993). Why is reaction time correlated with psychometric *g*? *Current Directions in Psychological Science, 2,* 53–56.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations.* Fort Worth,TX: Harcourt Brace Jovanovich.

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., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104,* 344–366.

Juslin, P., Olsson, H., & Winman, A. (1998). The calibration issue: Theoretical comments on Suantak, Bolger, and Ferrell. *Organizational Behavior and Human Decision Processes, 73,* 3–26.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgments under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103,* 582–591.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Journal of Applied Cognitive Psychology, 15,* 321–341.

Koehler, D. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 461–469.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14,* 389–433.

Larsson, M. (1997). Semantic factors in episodic recognition of common odors in early and late adulthood. *Chemical Senses, 22,* 623–633.

Lennox, R. D., & Wolfe, R. N. (1984). Revision of the self-monitoring scale. *Journal of Personality and Social Psychology, 46,* 1349–1364.

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.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 11,* 595–600.

Newstead, S. E., & Collis, J. M. (1987). Context and the interpretation of quantifiers of frequency. *Ergonomics, 30,* 1447–1462.

Pallier, G., Roberts, R. D., & Stankov, L. (2000). Biological vs. psychometric intelligence: Halstead's (1947) distinction re-visited. *Archives of Clinical Neuropsychology, 15,* 205–226.

Phillips, L. D. (1973). *Bayesian statistics for social sciences.* London: Nelson.

Psychological Corporation. (1997). *WAIS-III technical manual.* San Antonio, TX: Harcourt Brace.

Raven, J. C., Court, J. H., & Raven, J. (1979). *Manual for Raven's Progressive Matrices and vocabulary scales.* London: H. K. Lewis & Co.

Roberts, R. D., Pallier, G., & Stankov, L. (1996). The basic information processing (BIP) unit, mental speed and human cognitive abilities: Should the BIP RIP? *Intelligence, 23,* 133–155.

Roberts, R. D., & Stankov, L. (1999). Individual differences in speed of mental processing and human cognitive abilities: Towards a taxonomic model. *Learning and Individual Differences, 11,* 1–120.

Roberts, R. D., Stankov, L., Pallier, G., & Dolph, B. (1997). Charting the cognitive sphere: Tactile-kinesthetic performance within the structure of intelligence. *Intelligence, 23,* 133–155.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19,* 143–154.

Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgments. *The Journal of Experimental Education, 65,* 135–146.

Schraw, G., & Dennison, R. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19,* 460–475.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory and Cognition, 22,* 63–69.

Seashore, C. B., Lewis, C., & Saetveit, J. G. (1960). *Seashore's measures of musical talent: Manual.* New York: Psychological Corporation.
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.
Stankov, L. (1997). *The Gf/Gc Quickie Test Battery.* Sydney, Australia: The University of Sydney, School of Psychology. Unpublished test battery.
Stankov, L. (1998). Calibration curves, scatterplots and the distinction between general knowledge and perceptual tests. *Learning and Individual Differences, 8,* 28–51.
Stankov, L. (1999). Mining on the "no man's land" between intelligence and personality. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 314–337). Washington, DC: American Psychological Association.
Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence, 28,* 121–143.
Stankov, L., & Crawford, J. (1996a). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 21,* 971–986.
Stankov, L., & Crawford, J. (1996b). Confidence judgments in studies of individual differences support the 'confidence/frequency effect.' In C. Latimer & J. Michell (Eds.), *At once scientific and philosophic: A festschrift in honour of J. P. Sutcliffe* (pp. 215–239). Sydney, Australia: The University of Sydney Press.
Stankov, L., & Crawford, J. (1997). Self-confidence and performance on cognitive tests. *Intelligence, 25,* 93–109.
Stankov, L., & Dolph, B. (2000). Metacognitive aspects of test-taking and intelligence. *Psychologische Beiträge, Band 42,* 213–227.
Stankov, L., & Horn, J. L. (1980). Human cognitive abilities revealed through auditory tests. *Journal of Educational Psychology, 72,* 19–42.
Stankov, L., & Pallier, G. (2002). *Accuracy, confidence, and speed in perceptual tasks: Calibration effects and individual differences.* Manuscript submitted for publication.
Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161–188.
Vickers, D. (1979). *Decision processes in visual perception.* New York: Academic Press.
Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised.* New York: The Psychological Corporation.
White, T. L. (1998). Olfactory memory: The long and the short of it. *Chemical Senses, 23,* 433–441.
Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgments: Two different accounts. *Scandinavian Journal of Psychology, 34,* 135–148.
Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice-Hall.
Zakay, D., & Glicksohn, J. (1992). Overconfidence in a multiple-choice test and its relationship to achievement. *Psychological Record, 42,* 519–524.

Manuscript received June 19, 2001
Revision accepted for publication March 18, 2002