Performance Calibration in Sport: Implications for self-confidence and metacognitive biases

Gerard J. Fogarty and David Else

University of Southern Queensland

Keywords: METACOGNITON, CALIBRATION, SELF-CONFIDENCE

Contact Details:
Dr Gerard Fogarty
Faculty of Sciences, University of Southern Queensland
Toowoomba, QLD, 4350; Australia
fogarty@usq.edu.au

Acknowledgements: We would like to thank Chris Graham who assisted with data collection for this research and the golfers at Toowoomba Golf Club who gave up their time to participate in the research. We would also like to thank the reviewers and editors who provided helpful input.

Full reference: Fogarty, G., & Else, D. (2005). Performance calibration in sport: Implications for self-confidence and metacognitive biases. *International Journal of Sport and Exercise Psychology*, 3(1), 41-57.

Abstract

When people are asked to make judgments about their own performance, either retrospectively or prospectively, they typically overestimate their level of performance, leading some researchers to claim that overconfidence is a pervasive metacognitive bias. Evidence for such a trait in sport has implications for the way we assess confidence and for our understanding of athletes' perceptions of their own abilities and their reactions to performance feedback. To gain a better understanding of this issue, we used the calibration paradigm to measure metacognitive bias in a sample of 54 male golfers varying widely in age (13 to 75 years) and ability level (1 to 27 handicap). Golfers were required to complete a putting task and a chipping task (20 trials each) after first estimating how well they would perform on each of the tasks. The exercise was repeated once. Results indicated that golfers tended to be reasonably well-calibrated on the putting tasks but slightly overconfident on the chipping tasks used in this study. They were also overconfident on a test of knowledge of golf rules. There was no effect for level of expertise. Golfers differed in their ability to use feedback on the first set of trials to achieve better calibration on a second set of trials. Discussion centres on the potential benefits of using the calibration paradigm in a range of sports as an adjunct to assessments and interventions by sports psychologists.

Performance Calibration in Sport: Implications for Self-Confidence and Metacognitive Biases

The measurement and study of self-confidence is a central theme in sport psychology where it is considered to be a key mediating variable linking ability with performance (Feltz, 1988). Interest in self-confidence in sport psychology is largely due to the perceived dramatic effect it has on performance. This impact is both positive, as when successful athletes report greater self-confidence than less successful athletes (George, 1994), and negative, as evidenced by the sometimes dramatic effect that a loss of self-confidence has on performance in an experience commonly referred to as 'choking' (Vealey, Hayashi, Garner-Holman, & Giacobbi, 1998). However, there is an important difference between states of justified confidence and feelings of overconfidence or underconfidence. Techniques that help to assess the accuracy of athletes' confidence judgments are therefore of major interest in sport psychology, from both a theoretical and an applied point of view. The present study represents one of the first attempts in the field of sport psychology to apply what is known as the "calibration paradigm" to discover some of the factors that contribute to accurate self-knowledge in the sport of golf. More specifically, this study aims to extend our current understanding of self-reported estimates of performance capabilities, or efficacy beliefs, and factors such as feedback, age, and level of expertise. Possible implications in terms of training will be discussed. We begin with a description of the calibration paradigm itself.

The Calibration Paradigm

Calibration studies stemmed from research investigating human decision making under conditions of uncertainty. These studies endeavoured to assess the appropriateness of people's subjective probability estimates, or confidence, in their judgments and predictions (Keren, 1991). More simply, calibration refers to the accuracy with which one can rate or predict one's own performance. For example, in the testing domain where much of this research has been conducted, participants are typically asked a question and then asked to state how confident they are that the answer they gave is correct. This confidence rating is expressed in terms of a percentage. If this procedure is repeated over a number of trials, three measures can be obtained: the average number of items correct; the average confidence rating for the set of items; and the difference between these two, which is usually called the bias score. A person with a large positive difference when mean performance score is subtracted from mean confidence score is overconfident, a person with a large negative difference is underconfident. A person whose average confidence rating corresponds with the actual percentage correct is said to be well calibrated. The further a bias score deviates from zero, the poorer the realism of the confidence judgments. The bias score has been found to be highly reliable and can be taken as an estimate of where a respondent lies on the underconfidence-overconfidence dimension (Pallier et al., 2002).

Empirical Findings from Calibration Studies

Festinger (1954) postulated a universal motive to acquire self-knowledge. Trope (1979) argued that people have a need for accuracy and found evidence that people prefer to engage in tasks that provide them with valid information about their abilities. Despite these assurances, empirical evidence continues to mount indicating that inaccurate self-monitoring is the norm with inaccuracy usually taking the form of overconfidence. Overconfidence has been demonstrated across numerous domains and tasks including predicting the outcome of past events (Lichtenstein, Fischhoff, & Phillips, 1982), assessment of reading skills (Glenberg & Epstein, 1987), marketing management predictions (Mahajan, 1992), categorical judgment tasks (Schneider, 1995), motor task performance (West & Stanovich, 1997), eye witness memory (Bornstein & Zickafoose, 1999), economic forecasts (Braun & Yaniv, 1992), and general knowledge tasks (Kleitman & Stankov, 2001).

The extent and apparent generalisabilty of this overconfidence effect across many different domains and tasks has led to the claim that the phenomenon of overconfidence is a pervasive cognitive bias (Baron, 1994). Calibration studies, however, have also reported underconfidence in some domains and participant groups. Studies suggest that people tend to be underconfident when responding to sensory discrimination

tasks (Bjorkman, Juslin, & Winman, 1993), when answering questions about future events (Vreugdenhil & Koele, 1988), and with visual perceptual tasks such as the line length task (Stankov & Crawford, 1996). Good calibration has been found for tasks such as short-term memory (Fogarty, Burton, & Baker, 2000) and Raven's Progressive Matrices (Stankov & Crawford, 1996).

Calibration in sport.

While there has been some interest in the generalisability of the overconfidence-underconfidence phenomenon to motor tasks (West & Stanovich, 1997), to date there have been few studies of calibration in sport. In one of these studies Vertinsky, Kanetkar, Vertinsky, and Wilson (1986) studied the ability of female field hockey players to assess win/loss probabilities in an actual competition situation. Specifically, they had to forecast the results of games still to be completed in the regular season. The forecasts were made prior to each round with various types of information made available to the competitors. They found that the players were able to provide accurate probability assessments of game results. In other words, these elite hockey players were well calibrated as far as predicting results was concerned. In a study of competition bridge players, Keren (1987) reported that expert players had a much better appreciation of the likelihood of final bidding contracts being made. Horgan (1992) found that chess players with a high Elo rating (international chess rating system) made more realistic estimates of their likely performance against hypothetical opponents.

These studies are of interest because they concern competition and games. However, they are of the same genre as the research reported earlier in that they are still concerned with cognitive skills, primarily decision making and forecasting. It is possible that there are crucial differences between the cognitive activities examined in most calibration research and the types of motor skills one sees in sport, a point made by Bandura (1977) when he observed that one must distinguish between knowledge and performance skills. Kruger and Dunning (1999) took this further when they noted that the tendency for individuals to overestimate competence is likely to depend on the domain under consideration. In some domains, such as sport, competence is not wholly dependent on knowledge or wisdom but also depends on actual physical skill. Kruger and Dunning commented that in a sport such as golf, it should be obvious to less skilled individuals that they do not hit the ball as far or as accurately as some of their fellow competitors. Whether this translates into better calibration, we do not know because there have been very few attempts to apply techniques introduced by Lichtenstein and Fischhoff (1977) to the calibration of athletic skills. The present study addressed this shortcoming by assessing calibration in the sport of golf. The hypotheses were based on findings from the literature which were expected to generalise to sport settings. The rationale for these hypotheses is set out in the following paragraphs.

One of the most consistent findings in the calibration literature is that people tend to be overconfident on tests of general knowledge (Kleitman & Stankov, 2001). There are many types of knowledge that one can acquire in golf but perhaps the most inescapable form of learning involves mastery of the rules of the game. Some players know most of the rules, some know only a few that cover commonly encountered situations. What is certain is that few players at club level would ever have found themselves in a situation where they were tested on the rules of the game. These conditions should elicit the usual overconfidence effect observed with tests of general knowledge. It was therefore hypothesised (H1) that overconfidence would be exhibited on a test of the rules of the game.

Horgan (1992) argued that the task context is of vital importance in determining when people can learn from experience. In the case of performance skills and motor tasks where immediate feedback is provided in the form of success or failure, calibration becomes an integral part of learning the task and environmental cues are always available to ensure the accuracy of calibration. Golf requires continual predictions about club and shot selection, swing rate, and environmental conditions, all of which have a direct impact on performance. Thus, despite the general tendency towards overconfidence on cognitive tasks, it was felt that the domain of sport provides sufficient environmental cues and feedback mechanisms for people to develop accurate probabilistic mental models of their performance capabilities. It was therefore hypothesised (H2) that good calibration would be exhibited on tests of putting and chipping skills, both

fundamental and well-practised skills in the sport of golf.

Kruger and Dunning (1999) observed that in the cognitive domain, less competent individuals accounted for a large proportion of the overconfidence effect. Furthermore, studies reported earlier in this paper on the ability of players to predict the outcome of sporting events (Vertinsky et al., 1986; Keren, 1987) showed an effect for expertise. Taken together, these findings led to our prediction that better golfers would be more realistic in their expectations of what they could achieve. It was therefore hypothesised (H3) that low handicap golfers would be better calibrated than high handicap golfers.

Given the importance attached to good calibration in real-world decision making and accumulating evidence that people tend to be overconfident, it is important to examine the modifiability of this tendency. Self-correction in the face of performance feedback is the most obvious mechanism for bringing about change in self-perceptions. Lichtenstein and Fischhoff (1980) used this technique over a series of 11 training trials to improve calibration on a general knowledge task. Specifically, they gave personalised feedback after every trial, including calibration graphs showing deviations from perfect calibration, and discussed the feedback with participants. They found a marked improvement in calibration performance with most of the improvement occurring after just one session. These results led Lichtenstein and Fischhoff to conclude that one intensive feedback session is sufficient to achieve good calibration. Similar findings have been reported by other researchers (e.g., Bornstein & Zickafoose, 1999; Kruger & Dunning, 1999), but there are also instances of contrary findings, where overconfidence persisted despite feedback (Pulford & Colman, 1997; Stankov & Crawford, 1997).

Keren (1987) argued that improvement in calibration is most likely to occur in situations where the components of a task are highly related and where feedback is immediate and accurate. He noted that these conditions were met in the game of bridge where the components are the bidding and ensuing play. Feedback in golf is also immediate and accurate and sufficiently obvious to obviate the need for the type of intensive mediation used by Lichtenstein and Fischhoff (1980). One would expect that players introduced to a skills-testing situation would modify their performance estimates on the basis of knowledge of their performance on previous trials. It was therefore hypothesised (H4) that golfers would make use of feedback on early trials to improve (calibrate) estimates of performance on later trials.

Method

Participants

A total of 54 male golfers were recruited from golf clubs located in the Queensland cities of Toowoomba and Brisbane. The average age was 39 years (SD = 16.83) with the youngest participant being 13 and the eldest 75. Participants were recruited through personal contact and represented a range of handicap levels (1 to 27) with the mean being 14 (SD = 7.25).

Materials

Materials consisted of two short questionnaires assessing self-confidence in putting skill and chipping short distances (not used in this study and not described here), a putting task and a chipping task that formed the basis of the calibration measures, a four-item self-efficacy measure that was used in conjunction with the calibration tasks (not used in this study and not reported), a golf knowledge test, and club golf handicap. For readers not familiar with the sport of golf, a golfer's handicap is the traditional measure of skill in this sport, with low handicaps indicating a high level of competence.

The calibration tasks were adapted from those used by Thomas and Fogarty (1997) in their learning preferences intervention study.

Putting Task 1. Participants were required to hit 20 putts on a carpeted floor through an 11.43 cm (4.5 inches) target set 2.5 metres away. Participants completed 10 practice putts and were then asked to estimate how many putts out of 20 they could hit through the target. Instructions emphasised that we were seeking a realistic estimate of their score and not what they would 'like' to score. This procedure was similar to that employed by Ryckman, Robbins, Thornton, & Cantrell (1982) in their attempt to operationalise self-

confidence. Participants then completed the 20 putts. Three scores were obtained from this task: putting estimate 1 (converted to a percentage), putting score 1 (converted to a percentage), and putting bias 1; the bias score being the difference between obtained and estimated scores, where positive scores suggest overconfidence and negative scores underconfidence.

Chipping Task 1. Participants were required to hit 20 chip shots so that they landed in a circle with a diameter of 3 metres with its centre 17.5 metres from the teeing spot. The same procedure was followed as for Putting Task 1, with three outcome measures: chipping estimate 1 (converted to a percentage), chipping score 1 (converted to a percentage), and chipping bias 1. In the same procedure as the putting task, participants answered four items that assessed their level of efficacy that they could achieve this target immediately prior to hitting the 20 chip shots.

Putting Task 2. The putting task was repeated immediately after the completion of Chipping Task 1, giving a further three measures: putting estimate 2, putting score 2, and putting bias 2. Only five practice trials were allowed for both of the repeated tasks because it was assumed participants were sufficiently familiar with the task by this stage.

Chipping Task 2. The chipping task was also completed a second time, giving chipping estimate 2, chipping score 2, and chipping bias 2.

Putting Estimate 3 and Chipping Estimate 3 At the completion of the two 'rounds', participants were asked to estimate what they would score on the putting and chipping tasks if they were given the opportunity of one more trial. These estimates were converted to percentages to allow easier comparison with other estimates.

Golf knowledge test. A further set of questions was designed by the first author to assess calibration in knowledge of golf rules. Participants were asked to indicate, by ticking the appropriate box, the penalty that applies to each of a series of 10 possible situations that could arise in stroke play. Some items were easy, others were quite difficult. For example: "You putt out with the wrong ball and then tee off on the next hole." The response format consisted of four options wherein 1 indicated 'No penalty'; 2 indicated a 1-stroke penalty; 3 a 2-stroke penalty; and 4 indicated Disqualification. Following their response to each of these scenarios, participants were asked to rate (out of 10) how confident they were that their response was correct. These ratings were later converted to percentages. Total scores were computed for golf knowledge and also converted to percentages. This measure was treated as an index rather than a scale (Diamantopoulos & Winklhofer, 2001), so Cronbach's alpha was not calculated.

Handicap. Club golf handicaps as measured by the Australian Golf Union were recorded at the commencement of the test sessions. Handicaps represented a range from 1 through to 27 with a mean of 14 (SD = 7.21).

Procedure

All data were collected at the University's Centre for the Assessment of Human Performance and in its environs, or at the practice facilities of a local golf club. Immediately prior to the experimental procedure, each participant was provided with a brief verbal description about each of the tasks, told approximately how long it would take to complete the tasks (1-2 hours), and the importance of providing realistic estimates was stressed. Testing was conducted on an individual basis with the tasks presented in the order indicated in the method section of this paper. The project was approved by the University of Southern Queensland Ethics Committee.

Results

Preliminary Analyses

The data were screened through the Statistical Package for the Social Sciences (SPSS) Version 11.0.1, for accuracy of data entry, missing values and fit between the distributions and assumptions required for inferential statistics. There were no missing values as all testing was conducted in a one-on-one format. Descriptive statistics are shown in Table 1.

Table 1

Descriptive Statistics for all Variables

Variable	M	SD	Items
Age	39.54	16.84	
Handicap	13.91	7.25	
Putting Estimate 1	64.63	15.48	1
Putting Task 1	66.76	16.97	20
Chipping Estimate 1	48.52	15.25	1
Chipping Task 1	41.3	18.51	20
Putting Estimate 2	67.78	15.38	1
Putting Task 2	70.65	15.78	20
Chipping Estimate 2	49.44	16.56	1
Chipping Task 2	41.76	17.86	20
Putting Estimate 3	76.85	14.87	1
Chipping Estimate 3	54.26	18.57	1
Golf Knowledge Correct	66.85	15.64	10
Golf Knowledge Confidence	80.76	13.02	10

Examination of the descriptive data suggests that the tasks were of an appropriate level of difficulty for the purposes of this study. Relatively high standard deviations for many of the variables indicate considerable individual variation, an important requirement for calibration research. Examination of the skewness and kurtosis statistics indicated that most variables were normally distributed.

Hypothesis one was based on frequent reports in the literature of overconfidence on tests of general knowledge. In the present case, the Golf Knowledge Test was used to assess participants' understanding of the rules of golf. The data show that there was almost a 14% difference between mean accuracy score and mean confidence rating: t (53) = 5.34, $p < .01$. In other words, as a group, these golfers were overconfident regarding knowledge of the rules of the game. Hypothesis one was therefore supported.

The remaining hypotheses were tested using repeated measures MANOVA with three within-subject factors and with one between-subjects factor. The first within-subjects factor was Task and although there were no hypotheses relating to differences between putting and chipping, the inclusion of this factor allowed for the exploration of possible differential effects of feedback and calibration across tasks. The second within-subjects factor was Feedback wherein the first and second set of trials for putting and chipping constituted no-feedback and feedback conditions respectively. The third within-subjects factor was labelled Calibration and was designed to capture the differences between predicted and actual performance. To test the hypothesis relating to skill level within this MANOVA design, the sample was divided into three groups corresponding to the three grades used for club championships: A Grade (handicaps 0-11, n = 20); B Grade (handicaps 12-18; n = 19); and C Grade (handicaps above 18; n = 15). The combination of these different factors yielded a 2 x 2 x 2 x 3 design (Task: putting/chipping x Feedback: absent/present x Calibration: predicted/actual x Group: low/medium/high handicap).

Hypothesis two stated that good calibration would be observed on these two sporting tasks. Inspection of the MANOVA output showed that there were no higher order interactions but that the Calibration x Task interaction was significant: F (1, 51) = 20.82, p < .01. A test of simple effects showed that there were no differences between estimates of performance and actual performance for putting on either the first or second trial but that participants were overconfident on both trials for the chipping task: t (53) = 2.88, p < .01 for trial one and t (53) = 3.15, p < .01 for trial two. In other words, the familiar overconfidence effect was observed for the chipping tasks.

The third hypothesis stated that low handicap golfers would be better calibrated than high handicap golfers. There was no support for this hypothesis.

The fourth hypothesis stated that calibration would improve on the basis of feedback from earlier trials. Although there was no support for this hypothesis, it was noted that the highest-order interaction term (Task x Calibration x Feedback x Group) just failed to reach significance: F (2, 51) = 2.89, p = .06. The plot of this higher-order interaction captures the main trends in the present study and we therefore report it in Figure 1.

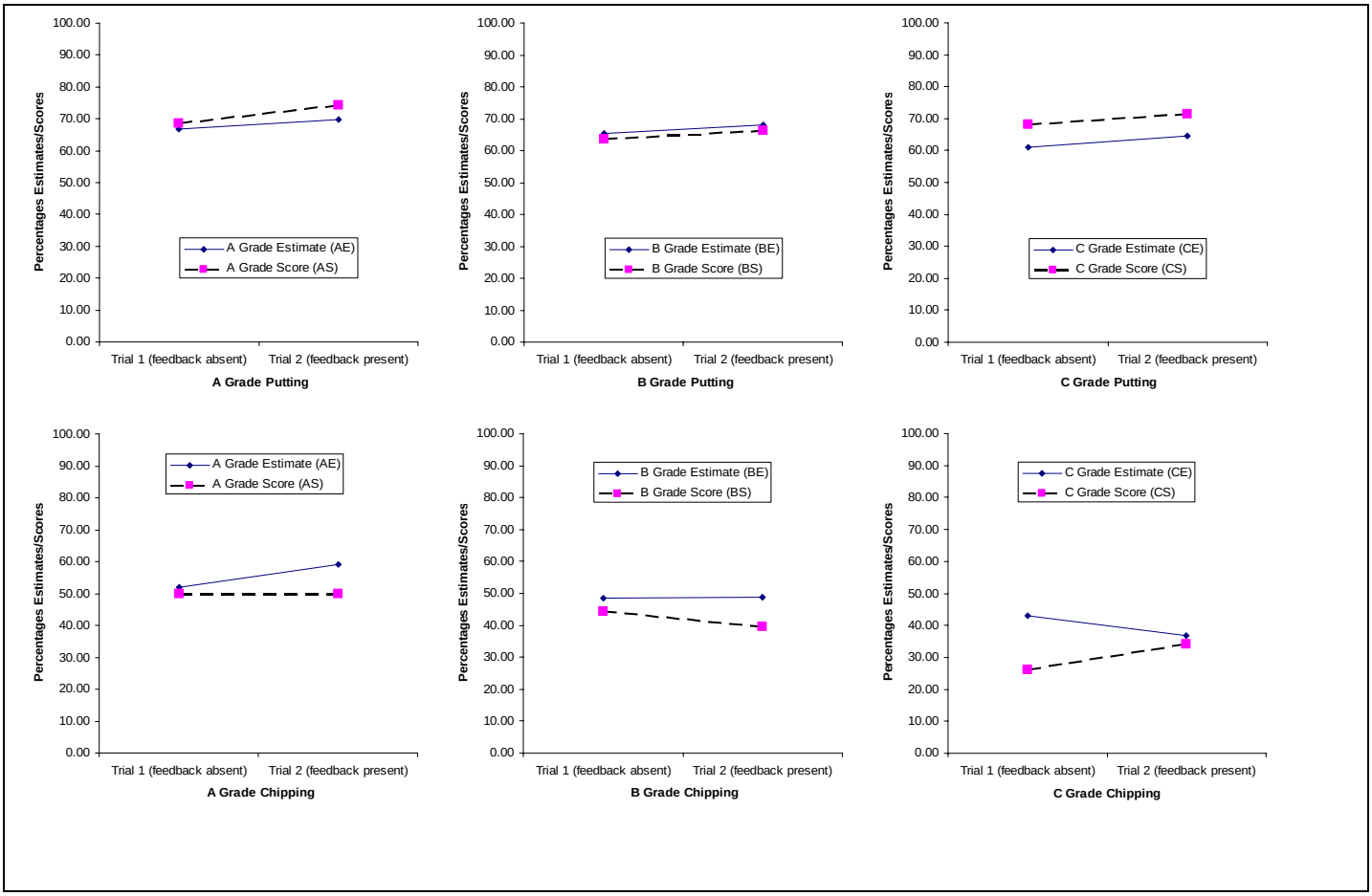

Figure 1. Estimated and actual mean performance scores across putting and chipping tasks for different levels of golfing expertise

Looking at the three plots in the top panel of Figure 1, it is obvious that very little happened in the putting task: there were no interactions, no effects across levels of expertise, and no feedback effects. The third panel suggests that C Grade golfers underestimated their performance but this effect failed to reach significance. The plots on the lower panel for the chipping task are more interesting. The A Grade golfers were well calibrated on trial one and then raised their estimates significantly for Trial 2 but did not perform any better on the second trial, leading to significant miscalibration. There were no significant effects for the

B Grade golfers. The C Grade golfers were significantly miscalibrated on Trial 1 but well calibrated on Trial 2.

Before concluding this section of the results, we draw the reader's attention to one further variable that we have not yet analysed. After the completion of the two trials for putting and chipping, participants were asked to set new targets for both tasks for an imaginary third trial. Because the participants knew that the third trial would not take place, we cannot be certain of the validity of these estimates. We note here that each of the three groups set targets that were approximately 6% higher than the targets set for trial 2. The A-Grade group would have needed a 9% performance improvement to meet this target, the B-Grade group a 14% improvement, and the C-Grade group a mere 4% improvement.

Ad Hoc Findings

A negative relationship was observed between age and the setting of targets for each of the tasks. That is, as the age of the participants increased, the target set for each of the tasks became more conservative. These correlations are shown in Table 2.

Table 2

Product Moment Correlations Between Age and Performance Estimates Across Trials

	Estimates for Putting Trials 1-3			Estimates for Chipping Trials 1-3		
Age	-.36	-.31	-.36	-.37	-.23	-.37

Note. All correlations significant at .05 level.

It should also be pointed out that whilst performance estimates became more conservative, there was no relationship between age and calibration. A similar finding was reported by Crawford and Stankov (1996) in their research on cognitive tasks.

Discussion

The main purpose of this study was to examine the concept of self-confidence in the sport of golf by applying the calibration paradigm. With this in mind, there are three main findings that warrant discussion; firstly, the finding that golfers displayed general levels of overconfidence in their judgments about their knowledge of golf rules and their chipping capabilities, but were well calibrated when judging putting performance. Secondly, the finding that there was no improvement in calibration across the two trials used in this study. The third finding of interest is that better golfers did not display better calibration. That is, level of expertise, as measured by official club handicap, was not related to ability to monitor performance on the present tasks. We discuss each of these findings under the general headings of calibration effects and expertise.

Calibration Effects

Regarding calibration, we observed good calibration on the putting tasks but a tendency towards overconfidence on chipping. One possible explanation for the difference observed between putting and chipping concerns the difficulty of the task. Apart from a few individuals, participants found the chipping task more difficult (see Table 1). Research on cognitive tasks has shown that individuals tend to exhibit greater levels of overconfidence for more difficult tasks, a pattern that can be observed right up to the range

of 80 percent accuracy, when individuals begin to display underconfidence (Lichtenstein & Fischhoff, 1977). This relationship between task difficulty and underconfidence or overconfidence is called the calibration difficulty effect and has been demonstrated across numerous domains (Keren, 1991).

The finding that golfers were overconfident on a test of golf rules adds to the long list of reports of overconfidence on tests of general knowledge (see Kleitman & Stankov, 2001). It has been suggested that this effect is task-related, induced by researchers selecting general knowledge items that are tricky and unrepresentative of the real world (Gigerenzer, Hoffrage, & Kleitenbolting, 1991). The present findings are of theoretical interest because they demonstrate that the overconfidence effect is not confined to abstract knowledge tasks used in psychological experiments but also applies to knowledge bases that are relevant and of intrinsic interest to participants.

The effects of feedback were investigated in the present study by using repeated measures MANOVA to check for interactions between Calibration and Feedback across three groups defined on the basis of ability. Evidence for the efficacy of feedback would take the form of converging estimate (E) and score (S) lines in Figure 1. It has been noted in the Results section that the C grade golfers were the only group to show improved calibration following feedback. At first glance, this finding appears to be at odds with the claim by Lichenstein and Fischhoff (1980) that performance feedback can bring about good calibration within two trials. One likely reason for the discrepant findings is that the feedback provided by Lichenstein and Fischhoff was more extensive than that provided in the current study. In addition to giving information about performance, after each trial they conducted personalized feedback sessions lasting 5-10 minutes in which they explained the principles of calibration using graphical techniques. In the present study, there was no discussion of discrepancies between targets and actual performance, no description of calibration techniques, and no attempt to administer anything other than standard experimental instructions to each participant. Under these conditions, it may take more trials to achieve good calibration.

In the somewhat novel situation encountered by participants in this study, we also have to consider the possibility that where miscalibration occurred, performance estimates may not have been at fault. In other words, miscalibration in sport can occur because of variability in performance as well as unrealistic expectations. In support of this interpretation, inspection of the individual scores revealed two instances where the chipping performance scores were well below estimates and also lower than one might expect on the basis of handicaps. It is likely that these participants with what were presumably uncharacteristically low scores continued to rely on cumulative ‘real life’ experience to set their estimates rather than using the feedback gleaned from the first chipping task.

Expertise

Various researchers have reported an effect for expertise (e.g., Kruger & Dunning, 1999; Schraw, 1997; Glenberg & Epstein, 1987; Vertinsky et al., 1986) on calibration tasks. Results from the present study suggest that skill level is unrelated to self-monitoring for motor tasks required in the sport of golf. That is, no evidence was found to suggest better golfers were better calibrated. This finding appears to be at odds with what has been found in the cognitive domain and with games such as chess and bridge (e.g., Horgan, 1992; Keren, 1987). However, one possible explanation is that good calibration depends on experience as much as expertise (C. Janelle, personal communication, July 9, 2003). Golf is a sport that can be played into old age. It is not uncommon to find older players with high handicaps who were once very good players. Thus, they have experience but no longer have the skill. In the present study, there was no measure of experience, so it was not possible to disentangle the effects of experience and expertise.

A further possible reason for the lack of a relationship between expertise and calibration is the unfamiliarity of the experimental situation itself. Whilst the chipping and putting skills are well-learned, it is likely that golfers would have considered the experimental situation in which they were asked to demonstrate these skills to be novel. A longer sequence of trials or more extended practice trials would ensure greater familiarity and help to overcome problems caused by the novelty of the situation and the inherent variability in sporting performance.

A final possibility is that despite their higher skill level, expert golfers are subject to the same

metacognitive biases as their less skilled colleagues. Diaz reported that when professional golfers were asked to guess the probability of sinking a six-foot putt on the Professional Golfers' Association (PGA) tour, they predicted the probability to be over 70%, a figure well above the tour average of 55% (as cited in Ericsson, 2001). Such findings suggest that professional golfers overestimate their putting capabilities. However, some caution is required here because Diaz did not ask the golfers to estimate their own performance in a particular putting situation, as would happen in most calibration studies. At this stage we don't know a lot about calibration among elite golfers. On the basis of the present data we tentatively conclude that they share with golfers of all levels a tendency towards overconfidence.

Implications for Sport

From an applied perspective, coaches, scientists and educators working in sport need to be wary of people's judgments of their own capabilities. As we have demonstrated, and as researchers have demonstrated repeatedly in the cognitive domain, people often lack insight into their motivations, emotions, and capabilities. Horgan (1992) suggested that good calibration has motivational benefits. Players with good calibration will tend to make appropriate attributions for both success and failure with fewer harmful effects when it comes to future participation, persistence and self-evaluation. Players with persistently poor calibration, on the other hand, especially those who are overconfident, are unlikely to learn from their mistakes, will probably suffer frustration, and eventually find that they lack motivation to continue to strive for the highest levels of achievement.

In a sense, there is nothing new in these conclusions. McClelland (1961) taught us years ago that one characteristic of high achieving individuals is that they set realistic goals. The calibration methodology is perhaps the best and easiest way of assessing the realism of these goals. It is a technique that can be used by researchers and coaches alike. In the words of Tommy Armour, one of golf's great teachers: "Every golfer scores better when he learns his capabilities" (Armour, 1965, p. 26). As a tool for skill development, calibration techniques are most likely to help those with inaccurate or biased self-perceptions.

Regarding interventions to improve calibration, we have to say that this is an area that requires a lot more research. Our present position is that personalised performance feedback, of the type used by Lichenstein and Fischhoff (1980), is probably the best method for improving calibration. In the context of sport, this would involve repeated use of the types of procedures employed in the current study coupled with graphic feedback techniques such as calibration curves (Keren, 1991).

Limitations and Suggestions for Further Research

As is typical of new approaches in a field, the present study was limited in scope. A larger and more varied sample of golfers would have permitted better tests of hypotheses relating to expertise, age, and experience. For example, there were no professional golfers in the present sample and few junior golfers. A more extended set of trials would have allowed a better test of the effect of feedback conditions and a more accurate assessment of calibration. We acknowledge that despite our attempts to simulate golf conditions, the situation was still different from actual playing conditions. It would be interesting to see whether all golfers adjusted to the conditions of the tasks over a longer sequence of trials. It would also be interesting to see a wider range of skills used, rather than just putting and chipping. The difficulty level of the skills could be manipulated to see whether calibration varies as a particular task becomes more or less difficult. For example, the putting task could introduce elements of break and speed, rather than being of uniform difficulty. These limitations lead to some suggestions for follow-up studies which are outlined in the next paragraph.

A rather obvious direction for future research is the inclusion of sufficient males and females to examine gender differences. Previous calibration research has demonstrated gender differences in the overconfidence–underconfidence dimension and it would be worthwhile to see if this extended to sport tasks. Fogarty et al. (2000), for example, found that in comparison with females, males were overconfident on tests of visual working memory. That is, females were actually better calibrated, although they were not better at the tasks. In the organisational psychology domain, Beyer (1990) found gender differences in the

accuracy of self-evaluations on masculine tasks but not on gender-neutral tasks. Gender differences of this kind would have implications for sport, perhaps leading to different norms for males and females on tests of self-confidence if it was known the males generally overstate their level of confidence.

A second avenue for future research is to explore different aspects of calibration in sport. The main analytical techniques used in this study were based on group means, which can mask individual differences relating to calibration and feedback. We attempted to overcome this shortcoming in the design by forming three groups of differing skill level and including this group factor in the repeated measures MANOVA. By using this technique, Kruger and Dunning (1999) were able to show that a large part of the variance in calibration on cognitive tasks was attributable to the "incompetent" group. We were not able to replicate that finding in our own study. However, the small sample size (N = 54) employed in the current study allowed us to inspect patterns of scores on an individual basis. At this level, we observed marked individual differences in calibration and response to feedback. Some individuals were well-calibrated on both tasks, some showed improved calibration on the second trial or in the targets they set for the hypothetical third trial, some were poorly calibrated throughout. All that we know from the current study is that expertise was not the basis for these individual differences. Future studies can explore other individual differences variables that promise to capture some of this variance. The research by Pallier and colleagues (Pallier et al., 2002) is a useful starting point for those interested in personality variables.

A third suggestion for follow-up studies involves applying the calibration paradigm to a wider range of sports. An interesting finding to emerge from the cognitive domain is that there are tasks where people typically display good calibration (e.g., short term memory tasks), tasks where overconfidence is the norm (e.g., tests of general information), and tasks where underconfidence is usually found (e.g., perceptual discrimination tasks). It may be that similar variation will be observed across the range of sporting skills and that this variation across different domains will in itself help to clarify the mechanisms underlying good calibration. Another area that requires further investigation is the disentangling of the effects of expertise and experience. One would expect these two constructs to be correlated in the cognitive domain but they may not be in sport. That is, one can be good at a sport without knowing a lot about it. Conversely, one can know a lot about a sport such as golf without being good at it. The failure to find an effect for expertise in the current study may well have been due to a confounding of expertise and experience.

In conclusion, calibration is a technique that is attracting some attention among researchers in the cognitive domain. It links up with topics of self-efficacy and self-confidence, both of which have a long research tradition in sports psychology. Its appeal lies in the fact that it yields measures that are based on both self-reports and actual samples of behaviour. Judging from findings in the cognitive literature, these measures are stable and not strongly associated with any known personality or ability construct. Instead, they are generally regarded as assessing a metacognitive skill that reflects the level of personal self-awareness (Kruger & Dunning, 1999). Given the concern in sport for accurate self-knowledge and performance measurement, we believe that calibration techniques have much to offer the development of theory and practice in sports psychology.

References

Armour, T. (1965). *How to play your best golf all the time* (3rd ed.). London: Hodder & Stoughton.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191-215.

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

Beyer, S. (1990). Gender differences in accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology, 59*, 960-970.

Bjorkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception and Psychophysics, 54* (1), 75-81.

Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it": The stability of the

confidence-accuracy relationship across domains. *Journal of Experimental Psychology: Applied, 5* (1), 76-88.

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

Crawford, J. D., & Stankov, L. (1996). Age differences in the realism of confidence judgment: A calibration study using tests of fluid and crystallized intelligence. *Learning and Individual Differences, 8* (2), 83-103.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*, 267-277.

Ericsson, K.A. (2001). The path to expert golf performance: Insights from the masters on how to improve performance by deliberate practice. In P.R. Thomas (Ed.), *Optimising performance in golf* (pp. 1-57). Brisbane, Australia: Australian Academic Press.

Feltz, D. L. (1988). Self-confidence and sports performance. *Exercise and Sport Science Reviews, 16*, 423-457.

Fogarty, G., Burton, L., & Baker, S. (2000). *Using Calibration Techniques to Improve Correlations Between Self-Report and Objective Measures of Visual Imagery.* Paper presented at the 27th Annual Conference of the Australasian Experimental Psychology Society, Queensland, 28-30 April.

George, T. R. (1994). Self-confidence and baseball performance: A causal examination of self-efficacy theory. *Journal of Sport and Exercise Psychology, 16*, 381-399.

Gigerenzer, G., Hoffrage, U., Kleitenbolting, H. (1991). Probabalistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory and Cognition, 15* (1), 84-93.

Horgan, D. D. (1992). Children and chess expertise: The role of calibration. *Psychological Research, 54*, 44-50.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 39*, 98-114.

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

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15* (3), 321-341.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-estimates. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

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

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*, 149-171.

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

McClelland, D.C. (1961). *The achieving society*. New York: Free Press.

Mahajan, J. (1992). The overconfidence effect in marketing management predictions. *Journal of Marketing Research, 29*, 329-342.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L, & Roberts, R. (2002). The role of individual differences in the accuracy of confidence judgements. *The Journal of General Psychology, 129*, 257-299.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences, 23*(1), 125-133.

Ryckman, R. M., Robbins, M. A., Thornton, B., & Cantrell, P. (1982). Development and validation of a

Physical Self-efficacy Scale. *Journal of Personality and Social Psychology, 42*, 891–900.
Schneider, S. L. (1995). Item difficulty, discrimination, and the confidence-frequency effect in a categorical judgment task. *Organisational Behavior and Human Decision Processes, 61*, 148-167.
Schraw, G. (1997). The effect of generalised metacognitive knowledge on test performance and confidence judgments. *The Journal of Experimental Education, 65* (2), 135-146.
Stankov, L., & Crawford, J. (1996). Confidence judgements in studies of individual differences. *Personality and Individual Differences, 21*, 971-986.
Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence, 25*(2), 93-109.
Thomas, P.R., & Fogarty, G. (1997). Psychological skills training in golf: The role of individual differences in cognitive preferences. *The Sport Psychologist, 11(1),* 86-106.
Trope, Y. (1979). Uncertainty-reducing properties of achievement tasks. *Journal of Personality and Social Psychology, 37,* 1505-1518.
Vealey, R.S., Hayashi, S. W., Garner-Holman, M., & Giacobbi, P. (1998). Sources of sport-confidence: Conceptualization and instrument development. *Journal of Sport & Exercise Psychology, 20,* 54–80.
Vertinsky, P., Kanetkar, V., Vertinsky, I., & Wilson, G. (1986). Prediction of wins and losses in a series of field hockey games: A study of probability assessment quality and cognitive information-processing models of players. *Organizational Behavior and Human Decision Processes 38,* 392-404.
Vreugdenhil, H., & Koele, P. (1988). Underconfidence in predicting future events. *Bulletin of the Psychonomic Society, 26* (3), 236-237.
West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin and Review, 4* (3), 387-392.