Calibration in tennis: The role of feedback and expertise

Gerard J. Fogarty (fogarty@usq.edu.au)
Department of Psychology
University of Southern Queensland, Toowoomba QLD 4350 Australia

Anthony Ross (a_ross4@hotmail.com)
Department of Psychology
University of Southern Queensland, Toowoomba QLD 4350 Australia

Abstract

People are usually overconfident in their expectations of what they can do. The term used to describe the extent which people are accurate in these self-assessments is "calibration". The present study focused on the effects of feedback and expertise on calibration in the physical task of serving in tennis, and three cognitive tests related to tennis. Sixty-four male and female tennis players, ranging in ability level from social to professional, took part in the study. Participants completed a tennis rules test, a tennis general knowledge test, and a tennis technique test, along with confidence ratings regarding their answers. They then completed two trials of two tennis serving tasks, which also involved estimating their expected performance on each trial. The results indicated that participants were overconfident on the general knowledge test, the rules test, and the more difficult serving task, but well calibrated on the technique test, and the easier serving task. Expertise was not strongly related to calibration and feedback was not beneficial for the more difficult serving task. The results have implications for decision making in tennis players, especially in relation to the tendency towards overconfidence on difficult tasks.

Introduction

The ability to monitor past performance and predict future performance is an important part of day-to-day life. People use beliefs about their own abilities to help judge performance. Because these beliefs often do not match objective performance, they can lead to performance judgments that do not relate to real accomplishment (Dunning, Johnson, & Ehrlinger, 2003). The ability to be realistic when rating previous performance and making future probability judgments, often referred to as being "well calibrated", has been shown to have benefits in areas such as motivation and goal setting (Horgan, 1992).

In the cognitive domain, where much of the calibration research has been conducted, simple techniques are used to assess calibration. Participants usually answer knowledge-related questions and then indicate how confident they are that their answer was correct in percentage terms (Lichtenstein & Fischoff, 1977). When all questions are completed, the bias score can be obtained by subtracting the percentage of correct responses from the average confidence rating. If the proportion of correct responses corresponds with the average confidence rating, the subject is well calibrated. A positive bias score indicates overconfidence, while a negative bias score indicates underconfidence. Research using the calibration paradigm to judge metacognitive bias in the sporting domain normally varies from calibration testing in the cognitive domain, in that subjects are immediately aware of their result in physical tasks. Therefore, performance predictions are made before a block of attempts, rather than after.

Studies in calibration have found that people are generally overconfident when predicting their own performance. In the domain of motor skills, Cohen, Dearnaley, and Hansel (1956) found that when bus drivers were asked to judge whether they could drive through a narrow gap, they were generally overconfident and more experienced drivers were not any better calibrated than less experienced drivers. West and Stanovich (1997) also found overconfidence when participants completed a penny slide task on a table top, and although calibration improved on a second trial, significant overconfidence remained.

Calibration studies of actual physical performance in sport are hard to find. Jagacinski, Isaac, and Burke (1977) tested the ability of college-level and professional basketball players to take uncontested shots from different positions in the court. Before they took the shot, both the player and a passive observer predicted if the shot would be made. No evidence was found that the players were more accurate than the observers, and both were overconfident in their predictions even when there were penalties for poor predictions. McGraw, Mellers, and Ritov (2004) measured the confidence that recreational basketball players felt while making shots and the pleasure they felt with the results of those shots. They also found that most players were overconfident, and those who were more overconfident experienced less enjoyment.

Fogarty and Else (2005) used the calibration paradigm to measure metacognitive bias in 54 male golfers ranging in age from 13 to 75. Golfers were required to complete a putting task and a chipping task after first estimating how well they would perform on each of the tasks. Each exercise was repeated once. Results indicated that golfers tended to be reasonably well calibrated on the putting task but

slightly overconfident on the chipping task. Participants were also overconfident on the golf rules test, which is consistent with other cognitive calibration research.

This research was extended by Graham (2006) when he studied 137 junior golfers who gave estimates of their ability on putting, chipping, and pitching tasks before completing the physical tests. Two experiments were conducted. Experiment 1, which required players to putt and chip to the shortest target, revealed good calibration whereas experiment 2, which required players to chip and pitch at a more difficult target, revealed overconfidence.

Summarising these findings, it appears that overconfidence generally exists in the cognitive domain, while in physical tasks and sport, the limited research suggests that people vary from good calibration to overconfidence. In relation to the present study of metacognition in the sport of tennis, it was therefore hypothesised (H1) that overconfidence would be displayed on a test of tennis rules, a test of tennis technique, a test of tennis general knowledge, and on two tennis serving tasks.

The findings relating to expertise are less clear, but researchers have generally found that expertise does lead to better calibration. Keren (1987) found that expert bridge players were well calibrated when predicting the chances of a final contract being reached whereas amateurs were overconfident. Horgan (1992) found that better chess players were well calibrated, whereas players with lower ratings were overconfident. Toward (1997) split 24 female undergraduate basketball players into expert and novice groups where classification was based on how many seasons of competitive basketball members had played. He tested the relationship between action and cognition in the basketball foul shot, and found that experts monitored and predicted outcomes better than novices. Against this trend, Fogarty and Else (2005) found no evidence to suggest that lower handicap golfers were better calibrated on chipping and putting than high handicap golfers. Despite this last finding, the weight of evidence suggests that calibration and expertise are associated. It was therefore hypothesised (H2) that expert tennis players would display better calibration on a serving task than non-experts.

The final variable examined in this study was feedback, where again the available evidence suggests an effect provided certain conditions are met. Keren (1987) suggested that the accuracy of calibration depends on the similarity of the mental processes necessary for repeated probability assessments. When task items are similar and sufficient practice has occurred, he argued that it is feasible to develop procedures that can lead to accurate predictions. Keren also suggested that immediate, relevant feedback is imperative for good calibration. In the domain of general knowledge, Pulford and Colman (1997) found that feedback is only effective in improving calibration for hard questions. Kruger and Dunning (1999) found that in comparison to their more competent peers, incompetent subjects were less able to use feedback to adjust calibration. Fischer and Budescu (2005) found that when testing categorical decision making, learning depends on the type of feedback given.

The ability of athletes to learn from feedback when completing physical tasks is critical for success in sport. Fogarty and Else (2005) found improvement in calibration of putting and chipping in golf when using only two trial blocks. Graham (2006), also working in the sport of golf, found that players who were initially poorly calibrated used feedback from earlier trials to become better calibrated. On the basis of these findings, it was therefore hypothesised (H3) that calibration would improve on a tennis serving task where feedback is immediate and complete.

Method

Participants

Sixty-four tennis players ranging in age from 14 to 48 years (M = 20.63, SD = 6.97) were recruited through personal contact in Cairns, the Sunshine Coast, Brisbane, and the Gold Coast to take part in calibration tests. Players were selected based on variation in expertise and gender. There were 41 male and 23 female players. Participants included current and former professional players, social adult players, and tournament standard junior players. Current and former professional players were defined as experts (n = 25), and juniors and social players were defined as non-experts (n = 39) for the expertise analyses. Participants were also ranked according to expertise by a representative from Tennis Queensland as a cross reference for these groupings.

Instruments

Test of tennis rules. Fifteen multiple-choice questions were designed to test calibration in knowledge of tennis rules (e.g., What is the ruling if during doubles a player receives out of turn?). Participants were asked to circle the correct answer. They were then asked to indicate how confident they were that their answer was correct by selecting a confidence rating for each question in percentage terms (25%, 50%, 75%, or 100%). Three scores were attained from this test: Tennis Rules Confidence Rating, Tennis Rules Correct Answers (converted to a percentage), and Tennis Rules Bias Score, the Bias score being the difference between predicted and obtained scores, where positive scores suggest overconfidence and negative scores suggest underconfidence.

Test of tennis general knowledge. Fifteen multiple-choice questions were designed to test calibration in knowledge of tennis general knowledge (e.g., Which year did Pat Cash win Wimbledon?). The same procedure was followed as for the test of tennis rules giving a further three measures: General Knowledge Confidence Rating, General Knowledge Correct Answers (converted to a percentage), and General Knowledge Bias Score.

Test of tennis technique. Fifteen multiple-choice questions were designed to test calibration in knowledge of tennis technique (e.g., Which grip would most advanced players use for a smash?). The same procedure was followed as for the previous tests giving a further three measures: Technique Confidence Rating, Technique Correct Answers (converted to a percentage), and Technique Bias Score.

Serving task 1. Participants were required to hit 10 first serves on a tennis court into a target area that measured one-quarter of the service box. Participants were allowed five warm up serves before the instructions were explained. They were then asked to estimate how many first serves out of the 10 they could hit into the target area. Instructions emphasized that the first serves were to be hit like they would in a real match, and the estimate was to be a realistic estimate of their actual score and not what they would 'like' to score. Participants then completed the 10 serves. Three scores were attained from this task: Serving Estimate 1 (converted to a percentage), Serving Score 1 (converted to a percentage), and Serving Bias 1, the Bias score being the difference between estimated and obtained scores, where positive scores suggest overconfidence and negative scores suggest underconfidence.

Serving task 2. Participants were required to hit 10 first serves on a tennis court into a target area that measured one-eighth of the service box. The same procedure was followed as for Serving Task 1 except that participants were not allowed warm up serves. Three outcome measures were attained: Serving Estimate 2 (converted to a percentage), Serving Score 2 (converted to a percentage), and Serving Bias 2.

Serving task 1 retest. Serving Task 1 was repeated immediately after the completion of Serving Task 2, giving a further three measures: Serving Estimate 3, Serving Score 3, and Serving Bias 3.

Serving task 2 retest. Serving Task 2 was also completed a second time, giving Serving Estimate 4, Serving Score 4, and Serving Bias 4.

Procedure

Ethics approval was attained from the University of Southern Queensland. Data were collected at regional tennis associations and various tennis centres across Brisbane. Prior to the experimental procedure, each participant was provided with an information sheet detailing the study. Parental consent was obtained for each participant under 18 years. Informed consent was held for those participants over 18 years. Participants were offered the chance to indicate their desire to receive a copy of the results of the study via the consent form. Immediately prior to each task, participants were given a verbal description of the task, told approximately how long it would take to complete the task, and that results would not be shared with anyone outside the experimental team. Participants completed the tests, the questionnaire, and the tasks in the order they appear above. Average testing time was 45 minutes.

Results

Six cases were identified as having missing values. These cases did not perform the serving tasks. This resulted in 64 cases completing the cognitive tasks and 58 cases completing the serving tasks. Examination of z-scores calculated from the skewness and kurtosis statistics indicated that the variables were normally distributed.

Hypothesis 1 stated that there would be general overconfidence displayed on the tennis rules test, the tennis general knowledge test, the tennis technique test, and the two serving tasks. The hypothesis was tested by running a repeated measures ANOVA for the cognitive tasks. This resulted in a 2 × 3 (calibration: confidence rating/percentage correct; task: rules/general knowledge/technique) within-subjects design. The analysis revealed a significant interaction between calibration and task [Wilks' Lambda $F(2,126) = 34.28$, $p < .05$], indicating that the amount of participant overconfidence depended on the type of cognitive task (See Figure 1).

Paired sample t-tests showed that on the rules test, participants were overconfident, $t(63) = 9.86$, $p < .05$. On the general knowledge test, participants were overconfident, $t(63) = 7.39$, $p < .05$. On the technique test, participants were well calibrated, $t(63) = 1.70$, $p > .05$.

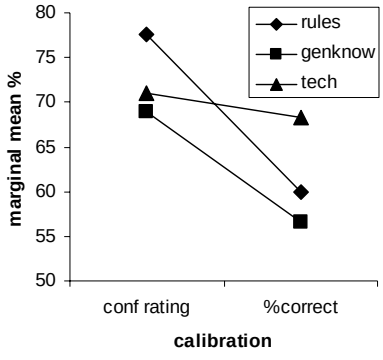

Figure 1. Overall calibration for cognitive tasks.

A repeated measures ANOVA was then run on the serving tasks. This resulted in a 2 × 2 × 2 (calibration: serving estimates/serving scores; task: 1, 2; trial: 1, 2) within-subjects design. There were no significant interactions. Inspection of the main effects revealed a significant effect for calibration [Wilks' Lambda $F(1,57) = 18.33$, $p < .05$] indicating that participants were overconfident across tasks.

Hypothesis 2 stated that experts would be better calibrated than non-experts. To test this hypothesis, participants were ranked in order of expertise then split into expert (n = 25) and non-expert (n = 39) groups. An expert was defined as any current or former professional player and a non-expert was defined as any junior or social player. Repeated measures ANOVAs were run on each of the serving tasks. Results indicated that there was no effect for expertise on the first serving task but that an effect was present in the more difficult second serving task [Wilks' Lambda $F(1,56) = 5.1$, $p < .05$]. Figure 2 shows the nature of this effect on Trial 2 of Serving Task 2.

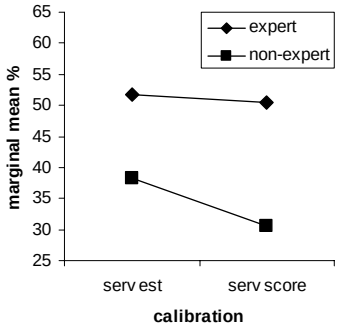

Figure 2. Experts versus non-experts.

Hypothesis 3 stated that feedback would improve calibration on the serving tasks. To test this hypothesis, an underconfident person was defined as someone whose serving estimate was lower than his/her serving score. An overconfident person was defined as someone whose serving estimate was higher than his/her serving score. Those who were underconfident on task 1 improved calibration (M = 10.95) on the retest, $t(20) = 4.26$, $p < .05$. Those who were overconfident on task 1 improved calibration (M = 18.97) on the retest, $t(28) = 4.56$, $p < .05$. Those who were underconfident on task 2 did not improve calibration on the retest, $t(13) = 1.24$, $p > .05$. Those who were overconfident on task 2 improved calibration (M = 13.44) on the retest, $t(31) = 3.70$, $p > .05$. To test this hypothesis further, a count was taken to determine whether individuals' calibration improved, declined, or stayed the same on the second trial. For task 1, 48.28% improved, 24.14% became worse, and 27.59% stayed the same between test and retest, indicating support for hypothesis 3. On task 2, little evidence of improved calibration was present, with 36.21% improving, 34.48% becoming worse, and 29.31% staying the same between test and retest.

Discussion

Hypothesis 1 was partially supported. Results of the rules test and the general knowledge test were in accord with previous research that has generally found overconfidence on cognitive tests (Lichtenstein & Fischhoff, 1977; West & Stanovich, 1997). However, participants were well calibrated on the technique test which was surprising. Participants scored approximately 10% more correct on the technique test than the other cognitive tests, but not only did participants know more on the technique test, they also knew more about how much they knew. It is possible that this was due to the emphasis that is placed on technical elements when learning tennis. An average player of tournament standard would have been bombarded with technical information for many years through coaching, and should generally have a good understanding of how much they know in this area. However, little emphasis is placed on the rules and the history of the game when learning to play, therefore making these subjects susceptible to the same biases that some researchers have argued cause overconfidence in people who have limited knowledge on cognitive tasks (Kruger & Dunning, 1999).

On the serving tasks, players were well calibrated on the easier task (task 1), but overconfident on the more difficult task (task 2 and the retest of task 2). Fogarty and Else (2005) and Graham (2006) also found that golfers were well calibrated on easier tasks (putting) and overconfident on more difficult tasks (chipping and pitching). This pattern is common in the cognitive field where it is known as the 'calibration difficulty-effect' (Keren, 1991). It appears that whatever causes this difficulty-effect is common to both physical tasks and cognitive tasks.

Findings relating to expertise were equivocal, with an effect emerging only for Serving Task 2. The fact that experts were better calibrated than non-experts on the more difficult task suggests that they may be more familiar with aiming at a smaller target area when serving than non-experts. Fogarty and Else (2005) also failed to find an effect for expertise in their study of golfers. Perhaps expertise actually interferes with calibration by making experts overly confident in their judgments, especially in the case of easier tasks,

The outcomes were also equivocal in relation to feedback. In the case of the first serving task, participants who were mis-calibrated on Trial 1 tended to improve on Trial 2, but this trend was not evident on the more difficult second serving task. Fogarty and Else (2005) and Graham (2006) both

found that poor calibration was more likely to occur on difficult performance tasks in golf, and that feedback did not have much effect on this mis-calibration. It appears that whatever causes the 'calibration difficulty-effect' in sporting tasks also contributes to making it more robust and resistant to change.

The findings of the current study should be treated with some caution. It is the first time calibration procedures have been applied to the sport of tennis, and to the best of our knowledge this is only the third time this technique has been used in any sport. The major limitation of our methodology is that conditions under which tennis players serve in competitive matches were not replicated in this study. Although the physical task of serving was the same as would occur in a match, there was no one returning the serve and participants were not required to finish the point. Calibration is important in all walks of life, including sport, and future research should be directed at methodological as well as theoretical issues.

References

Cohen, J., Dearnaley, E. J., & Hansel, C. E. M. (1956). Risk and Hazard: Influence of training on the performance of bus drivers. *Operational Research Quarterly, 7*, 67-82.

Dunning, D., Johnson, K., & Ehrlinger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83-87.

Ehrlinger, J., & Dunning, D. (2003). How Chronic Self-Views Influence (and Potentially Mislead) Estimates of Performance. *Journal of Personality and Social Psychology, 84*(1), 5-17.

Fischer, I., & Budescu, D. V. (2005). When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks. *Organisational Behavior and Human Decision Processes, 98*(1), 39-53.

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.

Graham, C. J. (2006). *Calibration in sport: The role of feedback, age and gender*. Unpublished manuscript, University of Southern Queensland, Toowoomba, Australia.

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

Jagacinski, R. J., Isaac, P. D., & Burke, M. W. (1977). Application of signal detection theory to perceptual-motor skills: Decision processes in basketball shooting. *Journal of Motor Behavior, 9*(3), 225-234.

Keren, G. (1987). Facing Uncertainty in the Game of Bridge: A Calibration Study. *Organisational Behavior and Human Decision Processes, 39*, 98-114.

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

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

McGraw, A. P., Mellers, B. A., & Ritov, I. (2004). The affective cost of overconfidence. *Journal of Behavioral Decision Making, 17*(4), 281-295.

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

Toward, J. I. (1997). Metacognitive knowledge and skilled sport performance. *Dissertation Abstracts International, 58*(6), 2136A.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

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.

NB. The correct APA citation for this paper is:

Fogarty, G., & Ross, A. (2007). Calibration in tennis: The role of feedback and expertise. In K Moore (Ed.), *Proceedings of the 2007 Conference of the Australian Psychological Society*, pp.148-152. Brisbane, Australia, 25-29 September, 2007.