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

Running head: OVERCONFIDENCE

# Overconfidence: Feedback and Item Difficulty Effects

**Briony D. Pulford and Andrew M. Colman**

***University of Leicester***

*Author Notes*

Requests for reprints should be sent to Andrew M. Colman, Department of Psychology, University of Leicester, Leicester LE1 7RH, U.K. E-mail: amc@leicester.ac.uk.

Preparation of this article was supported by a research studentship awarded by the Economic and Social Research Council (ESRC) to the first author and by Grant No. L122251002 to the second author as part of the ESRC research programme on Economic Beliefs and Behaviour.

## Summary

Overconfident subjects were given immediate feedback of results in a general knowledge test in an attempt to de-bias them. In a 2 x 3 x 4 mixed factorial design (Feedback × Question Difficulty × Trial Blocks), the accuracy, confidence, and overconfidence of judgements of 150 subjects (48 male and 102 female) were measured. Hard questions produced significantly higher levels of overconfidence than medium-difficulty and easy questions, which in turn resulted in underconfidence. Combining all levels of difficulty, females were significantly less overconfident than males. No significant effect of external feedback was found, although better calibration in latter trial blocks for hard-level questions suggests that intrinsic feedback through self-monitoring occurred but was effective in reducing the bias only for hard questions.

## Overconfidence: Feedback and Item Difficulty Effects

There are marked individual differences in people's degrees of confidence in relation to their corresponding levels of accuracy in everyday judgements. For people at the extremes of this dimension, confidence is poorly calibrated inasmuch as it is markedly lower or higher than is warranted by the accuracy of the corresponding judgements, and such a mismatch between confidence and accuracy is a common and well-established finding (Lichtenstein, Fischhoff, & Phillips, 1982). In particular a person is said to manifest the trait of overconfidence when "confidence judgments are larger than the relative frequencies of the correct answers" (Gigerenzer, Hoffrage, & Kleinbölting, 1991, p. 506; see also Gigerenzer, 1991).

Several questions arise from this. Under what circumstances does poor calibration occur? How does such lack of calibration develop in the first place? Can people become better calibrated? Are there gender differences in overconfidence? In theory, confidence is related to experience: if feedback is positive and shows that accuracy is being achieved, then confidence should increase, whereas if the feedback is negative, then confidence should decrease (or remain stable if the feedback is disregarded, for example to protect self-esteem). Poor calibration may arise from a lack of feedback during the relevant time span when the level of confidence in a task is forming, or alternatively feedback may be available but may be misinterpreted or not used to alter confidence sufficiently.

To test the robustness of judgemental biases, studies have been carried out that try to de-bias subjects by various means and to specify the conditions under which the biases operate. Fischhoff (1982) pointed out that the source of a bias may lie in either the task performed, the people making the judgement, or a mismatch between the two. Leaving aside the latter two possibilities, the task may be inappropriate inasmuch as subjects may not understand the instructions, may perceive the task in a different way from how the experimenter sees it, may not be able to explain their responses clearly, may not be motivated highly, or may use

stereotypic answers to avoid thinking about each question.

Fischhoff (1982) proposed several ways in which people may be de-biased: warning subjects about the bias, describing the bias and its magnitude, providing subjects with feedback, and training subjects. As Fischhoff pointed out, however, even if a de-biasing technique is found to work, the underlying mechanisms still need to be understood, and if de-biasing can prove effective, then why does it not occur in everyday life? There have been many experimental attempts to reduce overconfidence by using these de-biasing techniques.

**Attempts to De-bias Via Training and Feedback**

Comprehensive training in calibration and probability assessment was used by Lichtenstein and Fischhoff (1980), after testing sessions involving general knowledge questions, and the training did result in improved calibration and reduced overconfidence, mainly after the first feedback session. This study showed that training can produce improvements in calibration for some people but not for others -- not, for example, for those who were initially well calibrated. The improved calibration was shown to generalize to some other tasks but not to all, indicating that learning had occurred but with limited generalization. Some criticisms of this research seems justified, on the grounds, for example, that the range of subjects used was limited -- only 12 acquaintances of the experimenters. Also, the feedback employed at the end of each session of 200 general knowledge questions was very artificial and statistical in nature, which means that the findings are of debatable relevance to understanding how people learn to be better calibrated in everyday life where this type of training does not occur.

A general knowledge test was used by Sharp, Cutler, and Penrod (1988) who, over a period of four weeks, gave subjects either a detailed statistical breakdown of their individual performance on a similar test a week before or no feedback at all. No significant improvement in calibration or reduction in overconfidence resulted from this type of statistical feedback. In an earlier investigation, however, Adams and Adams (1961) had

reported an improvement in subjects' calibration when information about calibration was given as feedback, and this improvement was found to be transferable to some related tasks. An unsuccessful attempt to reduce overconfidence was made by Fischhoff and Slovic (1980) by warning subjects that the task that they were attempting -- to discriminate between the drawings of Asian and European children -- may be impossible. In this case considerable overconfidence was reported and only reduced by about 5 per cent when the warning was given.

In the studies just mentioned, the experimenters tried to train subjects to be better calibrated by giving them information about their overall calibration of confidence or warning them of the bias. Alternative de-biasing techniques have also been used, such as where the correct answer has been given to the subjects after every trial or group of trials or where the subjects have been required to supply justification for answers given.

Arkes, Dawes, and Christensen (1986) suggested that high overconfidence in a task may result in corrective feedback being ignored when people believe that their accuracy is already high. Their later research, however, showed that corrective feedback, in the form of the correct answers given to a highly overconfident group, did significantly improve the subjects' calibration (Arkes, Christensen, Lai, & Blumer, 1987). Arkes et al. manipulated the apparent difficulty of the practice session questions and found that the greatest reduction in confidence results from feedback that contradicted subjects' feelings the most -- subjects overcompensated and became underconfident when they believed they were doing well but were told that they were performing poorly. Telling subjects that they are doing badly when they were already aware of it did not have such a marked effect on confidence.

Other methods of de-biasing subjects have also shown some success. Koriat, Lichtenstein, and Fischhoff (1980) proposed that overconfidence may result from biases in information processing, such that recall from memory, either during or after the decision making process, tends to be biased towards evidence supporting a tentative answer rather

than contradicting it. To test this hypothesis, attempts were made to de-bias subjects (that is, to reduce their overconfidence) by asking them to write down all the reasons they could think of why each of the multiple choice answers was right or wrong. This method resulted in subjects becoming very well calibrated for all levels of confidence except the very highest, and a significant reduction in overconfidence resulted. The generation of pro/con reasons had no significant influence on subjects' accuracy levels, but it did lower confidence, resulting in reduced overconfidence. However, when the subjects were required to generate only one piece of supporting evidence or one supporting and one contradicting, this had no effect on calibration.

According to Koriat, Lichtenstein, and Fischhoff (1980), people search their memories to come up with an answer and then review the evidence and assess their confidence in that answer. The results show that the important factor was the production of reasons that contradicted the subjects' chosen answers. This finding suggests that it is the act of generating reasons why the answer may be wrong that reduces overconfidence. These researchers concluded from their findings that de-biasing techniques are likely to be most effective if they encourage the generation of contradicting evidence and suppress the generation of supporting evidence.

Some experimenters have found a natural improvement in calibration without the need for feedback from an external source. In tasks in which subjects perform the same judgement task over and over again, or in which the accuracy of their judgements is obvious to the subjects, self-feedback occurs almost inevitably. Understanding how self-feedback operates is obviously important because it occurs so frequently in everyday situations where feedback from other sources is unavailable.

In one of the earliest studies in this field, Adams and Adams (1961) found that subjects who were intentional learners showed less overconfidence than incidental learners (who were exposed to the same material but did not deliberately attempt to learn it), and improved their

calibration with practice, over 16 trials, which incidental learners did not. This improved calibration occurred in the absence of specific external feedback about performance or calibration; it was a natural improvement. In this experiment the same task was repeated up to 16 times, so subjects presumably saw the test materials again and again and thus indirectly had feedback about their performance in the previous session. It seems that intentional learners were motivated to monitor their performance with the goal of improving it, and thus they gave themselves feedback which improved their calibration. Although the authors did not comment on this, a possible reason why the incidental learners did not show corresponding improvement may have been that they were not motivated enough either to monitor their performance or to modify their confidence. Motivation may play a crucial role in improving calibration of confidence ratings by providing internal feedback through the increased monitoring of performance.

Paese and Sniezek (1991) found that subjects' confidence in a task (of judging the future performance of professional baseball players) increased with practice without any feedback being supplied. Although no improvement in performance had in fact occurred, subjects seemed to believe that they must have improved with practice and raised their confidence levels. As a result overconfidence increased with practice because the subjects' internal feedback was incorrect. If subjects who do not receive feedback are unable to improve their performance or to lower their confidence through self-feedback, then differences between the feedback and no-feedback groups are generally found in experiments. Other factors, however, may account for differences between these groups.

According to Lichtenstein, Fischhoff, and Phillips (1982), there may be times when people are not motivated to be honest in their assessments of confidence, because situations may reward or punish honesty differentially. Thus subtle pressures to conform, impress, or deny may be strong reasons to be mis-calibrated in one's judgements. Arkes et al. (1987) found that anticipating a discussion of their responses with peers lengthened the amount of

time that subjects spent making decisions and significantly reduced their overconfidence, which suggests that motivational factors regarding self-presentation were in operation.

Social pressures may influence actual and portrayed confidence, and such pressures may have different effects on different people. For example, the level of confidence expressed by women compared with men appears to be situationally dependent, and the type of task being undertaken is also relevant (Lenney, 1977). According to Lenney, women are less self-confident and devalue their performance as compared with men when feedback about their performance is either ambiguous or absent.

Lenney (1977) asserts that when people are informed of norms or a pass grade prior to a task, then females' expectations tend to be lowered. She cites evidence for the pressure of social cues on women's self-confidence and concludes that situational factors, including the individuals or groups with whom self-comparisons are made, influence self-confidence in women much more than men. Another important factor that affects expressed confidence in predicting how well a task will be performed is whether the task is perceived as being congruent with one's sex role. Tasks that are presented as sex-role inappropriate result in lower expectations of success and lower confidence (Stein, Pohly, & Mueller, 1971).

**Rationale for the Experiment**

The experiment described below was designed to investigate how the type of feedback available in everyday life can alter confidence and whether it can make the level of confidence more appropriate to the level of accuracy achieved. In previous experiments the effect of feedback may have been confounded with the level of difficulty of the questions, thus this experiment manipulates the level of difficulty to see whether feedback differentially affects overconfidence when the question difficulty varies. When the level of difficulty is constantly either very hard or very easy, subjects should learn quite quickly to adjust their confidence to an appropriate level. However, if the task consists of a mixture of hard, easy, and middling questions, then feedback will be of little use in determining confidence for the next question of unknown difficulty. Many judgements in everyday life are of course of the latter type, and this is probably why learning to be accurately calibrated in confidence is difficult.

Accurate calibration is said to exist when people have a realistic understanding of how right they are or how well they are performing. Feedback should therefore reduce overconfidence only if it changes people's perceptions or beliefs about their performance. The greatest improvements in calibration should occur when feedback about performance is consistent, and this in turn should occur when the questions are consistently hard or easy, but when there is a mixture of difficulty levels, the feedback about the difficulty of the task is necessarily ambiguous and calibration should not improve as much or at all.

Detailed statistical information on calibration, which has sometimes been given as feedback to train subjects to be better calibrated, seems very artificial and is not the type of information that people receive in everyday life. Thus the experiment reported below aims to be more true to life by using immediate right/wrong feedback about the accuracy of the subjects' responses to each question. With this form of feedback, subjects should realize either that they are answering many questions correctly, which should raise their confidence,

or that they are getting many answers wrong, which should lower their confidence.

Two treatment conditions (feedback and no-feedback) are used. In the feedback condition the correct answer is revealed to the subjects after each of their answers. In this condition the accuracy level should remain stable, and the subjects' confidence levels should either increase or decrease depending on how well or badly they are performing, thereby producing better calibration in the feedback group.

The effect of the presentation position of the questions (trial block position) is also examined to see whether subjects learn to reduce their confidence over time. It is expected that subjects will lower or raise their confidence estimates for the later questions after having received feedback on the accuracy of their answers in earlier trials. In the first trial block no difference in overconfidence should be found between the feedback and no-feedback conditions, because the subjects will not yet have received enough feedback. If feedback has an effect, then the differences between the two conditions should increase over trial blocks. The level of over/underconfidence should remain stable across trial blocks for the no-feedback condition. Thus an interaction between feedback/no-feedback and trial block is predicted for overconfidence.

## Method

### Subjects

For the purpose of selecting items for the main study, a pilot study was conducted with 57 subjects (13 males and 44 females) drawn from the same pool of subjects as for the main study. The subjects in the pilot study were undergraduate and postgraduate students studying a variety of subjects. In the main study, the subjects were 150 undergraduate students (48 males and 102 females) with an average age of 20.69 years (range 18 to 48 years).

### Design

A 2 × 3 × 4 mixed factorial design (Feedback × Question Difficulty × Trial Blocks) was used. The first factor was manipulated by giving some of the subjects feedback in the form

of the correct answer to the each of the questions immediately after they had answered, and others received no feedback. The second factor was varied across three levels of question difficulty: hard, medium, and easy. The last (within subjects) factor was varied across four levels: Trial Block 1 (the first 5 questions), Trial Block 2 (Questions 6-10), Trial Block 3 (Questions 11-15), and Trial Block 4 (Questions 16-20). The dependent variables were the levels of confidence, accuracy, and overconfidence (where overconfidence = confidence - accuracy) of the subjects' judgements. Subjects were randomly assigned to each of the six treatment conditions (Feedback x Question Difficulty) with 15 females and 8 males in each treatment condition.

**Materials**

For the pilot study, 60 questions were selected from the board game Trivial Pursuit. The criteria used to select the questions were that they should not be multiple choice, should have short one-word or two-word answers, and should not be subject to rapidly becoming out of date. A typical example of the questions used is: "What flavour is Grand Marnier?" (answer: orange).

Subjects completed the pilot study questionnaire at their own pace within a half-hour session. They were requested to give an answer for each of the 60 questions in turn, and to state their level of confidence in their answer, using any number between 0% ("no confidence") and 100% ("total confidence"). Subjects were told that if they made a wild guess they should indicate that they had little confidence that their answer was right, and thus choose a number close to zero per cent, but that if they were confident that the answer was correct, they should choose a number nearer 100 per cent. The subjects' responses were anonymous and confidential, and they were informed of this before they began the test.

The pilot study was used to determine the level of difficulty of each of the questions, i.e., the base rate proportion of subjects who answered each question correctly. The questions were ranked according to level of difficulty and then divided into three categories: hard (0-

33% accuracy), medium (34-66%), and easy (67-100%), with 20 questions in each of these categories.

**Procedure**

In the main study, subjects were tested in small groups. Each of the 20 questions was read out, and subjects were given 30 seconds to write down their answer and their level of confidence that the answer they had given was correct, using the scale and instructions described above. In the feedback condition the correct answer was read out after the 30 seconds had elapsed, and the subjects marked their answer as either correct or incorrect before proceeding to the next question. In the no-feedback condition the correct answers were not given until after the subjects had responded to all of the questions, at which point all the correct answers were read out and subjects marked their own answer papers. At the end of the testing session subjects were debriefed minimally to avoid knowledge of the purpose of the experiment leaking into the student population. The subjects were told that a full description of the experiment and its findings would be displayed on a notice board after all the subjects had been tested and any questions would be answered at that time.

## Results

**Manipulation Check**

The subjects' mean accuracy levels (see Table 1) were 21.00% for hard questions, 57.53% for medium questions, and 79.00% for easy questions, and these accuracy levels were all significantly different from each other, $F(2, 138) = 207.07$, $p < .001$, Tukey-HSD = 10.50, $p < .05$, effect size $\eta^2 = .75$, which provides a manipulation check confirming the categorization of the questions into three levels of difficulty respectively.

---

Insert Table 1 about here

---

**Question Difficulty**

The obtained levels of confidence for the three levels of difficulty were: hard ($M$ = 28.50), medium ($M$ = 48.03), and easy ($M$ = 67.58), which all differed significantly from one another, $F(2, 138) = 88.30$, $p = .001$, Tukey-HSD = 11.41, $p < .05$, effect size $\eta^2 = .56$. This indicates that 56 per cent of the variance in confidence ratings was due to the difficulty of the questions.

The level of difficulty of the questions also had a predictable effect on overconfidence (see Table 1), $F(2, 138) = 38.04$, $p < .001$, effect size $\eta^2 = .35$. A Tukey-HSD test showed that hard questions ($M$ = 7.50) resulted in significantly higher overconfidence than medium ($M$ = –9.50) and easy questions ($M$ = –11.42), Tukey-HSD = 8.49, $p < .05$. This is in line with previous research and is explained in terms of the *hard-easy* effect. Medium and easy questions did not differ significantly in overconfidence. Overall, hard questions produced the most appropriate levels of confidence, being closer to perfect calibration, although on the overconfident side, than medium-difficulty and easy questions, which in turn produced underconfidence.

**Effects of Feedback**

The mean overconfidence was –5.62% for the feedback group and –3.33% for the no-feedback group. This difference, though in the predicted direction, was non-significant. Irrespective of the level of difficulty, the experiment resulted in a significant overall level of *under*confidence ($M$ = –4.47), (accuracy = 52.51% vs. confidence = 48.04%), due to the large number of easy questions, $t(149) = 3.74$, $p < .001$.

No significant differences were found between the feedback and the no-feedback groups for either accuracy ($M$ = 53.22 vs. $M$ = 51.80 respectively) or confidence ($M$ = 47.60 vs. $M$ = 48.47). Thus, feedback had no significant effect on confidence, accuracy, or overconfidence across all sixty questions combined. There were no significant interactions involving the factors feedback and question difficulty for accuracy, confidence, or overconfidence.

**Trial Blocks**

Considering both the feedback and no-feedback groups together (bearing in mind that no significant differences were found between them), a small though significant effect of trial block on overconfidence was found: $F(3, 414) = 7.27$, $p < .001$, effect size $\eta^2 = .05$ (see Table 2). A posteriori comparison, using the Tukey-HSD test, revealed that Trial Block 3 ($M = –8.78$) yielded significantly lower overconfidence than Trial Blocks 1 and 2 ($M = –1.58$ and $M = –1.60$ respectively), Tukey-HSD = 5.20, $p < .01$. Trial Block 4 ($M = –6.00$) was not significantly different from the other three trial blocks.

---

Insert Table 2 about here

---

The predicted difference between the feedback and no-feedback groups in confidence, accuracy, and overconfidence did not occur, and there were no significant interactions with trial blocks. The two groups showed very similar patterns of overconfidence across trial blocks, with the feedback group having only slightly (though non-significantly) lower overconfidence. This may be because subjects intuitively knew whether they were getting the questions correct or not, and thus feedback from the experimenter was not necessary.

Paradoxically, the subjects were better calibrated in the first two trial blocks, that is Questions 1 to 10, and became more underconfident on later trials. This may be less anomalous than it appears, however, because the mean overconfidence scores conceal different patterns of overconfidence which are revealed when differing levels of question difficulty are taken into account. This becomes clear when overconfidence is broken down across trial blocks for each level of question difficulty.

**Interaction of Question Difficulty and Trial Blocks**

A significant interaction was found between question difficulty and trial blocks with respect to overconfidence: $F(6, 414) = 7.12$, $p = .001$, effect size $\eta^2 = .09$ (see Table 2). This appears to be due to the high levels of overconfidence for hard questions, which declined

sharply in later trial blocks. Almost perfect calibration was achieved for hard questions in Trial Blocks 3 and 4, which elicited significantly lower levels of overconfidence than Trial Blocks 1 and 2, $F(3, 196) = 16.57$, $p = .001$, Tukey-HSD = 11.81, $p < .01$. There was no improvement in overconfidence over the course of the trial blocks for either medium-difficulty or easy questions, contrary to our prediction that there would be an improvement for easy questions (see Table 2 for all means).

**Gender Differences**

Combining all levels of difficulty, male subjects were found to be slightly more overconfident than female subjects, $M = –.74$ vs. $M = –6.23$, $F(1, 138) = 6.90$, $p < .01$, effect size $\eta^2 = .04$, although overall both groups were underconfident. Male subjects also found the task easier than female subjects, at all levels of question difficulty, and achieved significantly higher accuracy scores, $M = 59.06\%$ vs. $M = 49.43\%$, $F(1, 138) = 14.59$, $p < .001$, effect size $\eta^2 = .09$. Male subjects correspondingly expressed significantly higher confidence in their judgements, $M = 58.32\%$ vs. $M = 43.20\%$, $F(1, 138) = 34.53$, $p < .001$, effect size $\eta^2 = .20$. The level of question difficulty did not differentially influence male and female subjects: both groups responded with similar patterns of overconfidence, but the female subjects were significantly less confident and accurate than male subjects for all levels of question difficulty.

## Discussion

No evidence was found for any significant effect of feedback on any of the dependent variables, but a small reduction in overconfidence was observed across trial blocks. This can probably be explained by the fact that it is relatively easy, at least in this type of general knowledge task, for subjects to self-monitor their own accuracy and thus to reduce their confidence appropriately. Subjects can do this even without direct feedback from the experimenter, because they know roughly whether they are answering questions correctly or not. Support for this self-monitoring interpretation comes from the observation that subjects

in the no-feedback, hard-level group showed much better calibration in the latter half of the trial blocks than earlier on. Thus subjects reduced their overconfidence for later questions, which suggests that they monitored their poor performance and in effect provided their own intrinsic feedback.

If internal monitoring accounted for the improved calibration of subjects in the no-feedback, hard-level group over trial blocks, then this process was apparently slightly less effective than external feedback at reducing overconfidence. In the absence of external feedback, subjects had to guess their accuracy rates and modify their confidence accordingly. This may have caused subjects to feel less certain about the validity of their own feedback in comparison to external feedback.

Subjects given medium-difficulty and easy questions showed no improvement in calibration of confidence ratings: their level of overconfidence remained stable across trial blocks and they were underconfident rather than overconfident. Several interpretations of this lack of improvement are possible. First, perhaps the subjects did not monitor their performance. However, the fact that subjects given hard questions and no feedback reduced their overconfidence levels over trial blocks suggests that self-monitoring occurred in that treatment condition, and it is reasonable to assume that subjects in other treatment conditions similarly engaged in self-monitoring. A second interpretation is that subjects may have attempted to self-monitor their performance but were unable to improve their calibration of confidence because the difficulty of the task was less clear-cut than in the hard-level treatment conditions. This is the most likely explanation for the lack of reduction in overconfidence over trial blocks in the medium-difficulty group. Third, perhaps the feedback, whether internal or external in origin, provided unambiguous information about the level of question difficulty, but subjects for some reason did not use it to modify their confidence ratings. This is probably the explanation for the lack of improvement of calibration among subjects in the easy-level group.

Why did subjects in the easy-level group apparently not use the feedback to increase their confidence and improve their calibration over trial blocks? The feedback groups did not differ from the no-feedback groups in confidence or overconfidence for questions of any level of difficulty. Thus feedback was obviously not the important factor, because it could be provided internally. Perhaps even though the subjects realized that the questions were very easy, through feedback or self-monitoring, there were social or psychological barriers preventing them from increasing their confidence. Subjects may have failed to increase their confidence for easy questions because they appeared too easy, and perhaps subjects suspected that there may be a trick to catch them out.

Perhaps people change their confidence levels only when past performance seems to be a reliable predictor of future performance, and this may be easier when questions are consistently hard rather than consistently easy. Subjects in the easy-level group did not seem to change their impression of the level of difficulty of the questions. They may have believed that luck rather than knowledge accounted for the apparent easiness of the questions, and luck is usually assumed to be something that can run out at any moment, and this may explain why confidence ratings were lower than they might have been. If people attribute higher than expected accuracy, when it occurs, to luck (external factors) rather than to internal factors such as skill, then perhaps they tend not to raise their confidence in judgements, because they feel that they have less control over those external factors.

The evidence from this experiment indicates that feedback, either internal or external, seems to improve calibration of confidence only when the questions are consistently hard. This could be because there is more social pressure to reduce confidence during hard tasks to save face in case of failure. This social pressure does not seem to raise confidence for easy questions where people are underconfident, because it may be more socially desirable to be underconfident and may even be a way of boosting self-esteem in cases of success when low confidence was previously expressed. Maintaining low confidence is also a protective

strategy, because if confidence is raised during an easy task, and the achieved accuracy is not as high as expected, then the task cannot easily be blamed for being too hard and failure must be internally attributed, which may threaten self-esteem.

Alternatively, a more cognitive explanation may account for the improvement in calibration for hard but not easy questions, in that feedback for hard questions may encourage subjects to generate evidence contradicting their chosen answers, which is known to reduce overconfidence (Koriat, Lichtenstein, & Fischhoff, 1980). Feedback to easy questions, however, does not encourage any generation of confirming evidence, which in any case is less effective in improving calibration.

In some situations feedback is disregarded when assessing future confidence levels. People may, for example, fail to alter their confidence levels because of motivational factors. However, it also seems logical to suppose that in the absence of external feedback subjects will not alter their confidence ratings if they do not realize that their confidence ratings are out of kilter with their true accuracy levels. If people cannot correctly assess their true accuracy levels, then calibration will not improve. In the absence of external feedback, improved calibration will occur only if subjects realize that their beliefs about their accuracy levels are out of line with their true accuracy levels. Thus to improve calibration, subjects must either be realistic in their assessments of their accuracy or have external feedback to give them an accurate measure of their true performance.

The most inappropriate levels of overconfidence in this experiment occurred at the extremes of the difficulty continuum, with judgements that were either very hard or very easy, where subjects did not realize how difficult or easy the judgements were. Some judgement tasks in this experiment and indeed in everyday life are misleading in their level of difficulty inasmuch as they seem much harder or easier that they really are. This in turn leads to poor calibration, because the objective accuracy rate does not turn out to equal the subjective accuracy, which is what subjects are basing their confidence on.

Gender also appears to be associated with overconfidence: male subjects were significantly more overconfident than females. The female subjects found the task significantly harder than the males -- they exhibited lower accuracy overall. Harder tasks usually produce more overconfidence, but the female subjects in this experiment were more underconfident than the male subjects. The female subjects believed either that the general knowledge task was harder than it actually was or that their general knowledge was poorer that it actually was. Because they believed that they were doing less well than male subjects, their confidence was correspondingly lower. As much as 20 per cent of the variance in confidence ratings was explained by gender.

It is possible that there was more social pressure on the female than the male subjects in this experiment to show low confidence for this type of task. In most previous experiments, for example on social and self-predictions, gender differences in overconfidence have not been found. This raises the question: are female subjects actually less confident or are they under-reporting their confidence to conform to social pressure to be modest in a task resembling intelligence-test questions? Lenney (1977) argued that women are not universally low in confidence in all tasks, and thus the desire to appear modest cannot invariably account for lower self-confidence, but that "social factors do influence women's actual levels of self-confidence" (p. 9). This could have consequences for women in terms of success and achievement, and it merits further research.

To conclude, feedback seems to be effective in improving calibration of confidence and accuracy only when the questions are consistently difficult; it does not improve calibration very much for medium-difficulty questions or easy questions. Feedback, be it intrinsic or extrinsic, is important for regulating the level of overconfidence but is not always effective in improving calibration of confidence unless the task is consistently difficult. In everyday life, tasks are often a mixture of hard and easy components and thus overconfidence may not be reduced to such a large extent. Overconfidence or underconfidence in everyday life appears

therefore to be due to a combination of motivational factors associated with self-presentation and also uncertainty about the level of task difficulty leading to mis-calibration of confidence.

## References

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. Psychological Review, 68, 33-45.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. Organizational Behavior and Human Decision Processes, 39, 133-144.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. Organizational Behavior and Human Decision Processes, 37, 93-110.

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

Fischhoff, B., & Slovic, P. (1980). A little learning...: Confidence in multicue judgment tasks. In R. Nickerson (Ed.), Attention and performance VIII (pp. 779-800). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. Psychological Review, 98, 254-267.

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

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

Lenney, E. (1977). Women's self-confidence in achievement settings. Psychological Bulletin, 84, 1-13.

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

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds), Judgment under

uncertainty: Heuristics and biases (pp. 306-334). New York: Cambridge University Press.

Paese, P. W., & Sniezek, J. A. (1991). Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision making. Organizational Behavior and Human Decision Processes, 48, 100-130.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. Organizational Behavior and Human Decision Processes, 42, 271-283.

Stein, A. H., Pohly, S. R., & Mueller, E. (1971). The influence of masculine, feminine, and neutral tasks on children's achievement behavior, expectancies of success, and attainment values. Child Development, 42, 195-207.

Table 1. Mean confidence, accuracy, and overconfidence for the three levels of question difficulty

| | A: Hard | B: Medium | C: Easy | *F* |
|---|---|---|---|---|
| Confidence | 28.50 | 48.03 | 67.58 | 88.30* |
| Accuracy | 21.00 | 57.53 | 79.00 | 207.07* |
| Overconfidence | 7.50 | -9.50 | -11.42 | 38.04* |

*$p < .001$

Table 2. Overconfidence for different levels of question difficulty over trial blocks

| | TrialBlock | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | *F* |
| A:Hard | 13.30 | 18.00 | -0.18 | -1.14 | 16.57* |
| B:Medium | -9.27 | -12.77 | -12.25 | -3.82 | 2.37 |
| C:Easy | -8.80 | -10.02 | -13.90 | -13.03 | 1.11 |
| MeanOver-confidence | -1.59 | -1.60 | -8.78 | -6.00 | 7.27* |

*$p < .001$