Does Supplementing Outcome Feedback with Performance Feedback Improve Probability Judgments?

Ning Du
Associate Professor, School of Accountancy and Management Information Systems
DePaul University, DePaul Center, One East Jackson Boulevard, Chicago, IL 60604, USA
Tel: 312-362-8308 E-mail: ndu1@depaul.edu

Sandra Shelton
Professor, School of Accountancy and Management Information Systems
DePaul University, DePaul Center, One East Jackson Boulevard, Chicago, IL 60604, USA
Tel: 312-362-8308 E-mail: sshelton@depaul.edu

Ray Whittington
Professor, School of Accountancy and Management Information Systems
DePaul Universit, DePaul Center, One East Jackson Boulevard, Chicago, IL 60604, USA
Tel: 312-362-8308 E-mail: rwhittin@depaul.edu

Received: June 19, 2012 Accepted: July 3, 2012 Online Published: October 10, 2012
doi:10.5430/ijfr.v3n4p19 URL: http://dx.doi.org/10.5430/ijfr.v3n4p19

Abstract

This study examines the effect of feedback on calibration of auditors' probability judgment. Results from our experiment indicate that auditors' probability judgments are indeed poorly calibrated and auditors are in general overconfident. In addition, we investigate whether we can use feedback as a means to improve judgment quality. Our evidence indicates that outcome feedback, in spite of its simplicity, is effective in reducing overconfidence. Moreover, we find supplementing outcome feedback with performance feedback does not have any added benefits. Sometimes, it actually decreases judgment performance.

Keywords: Feedback, Overconfidence, Probability judgment, Confidence interval, Calibration

1. Introduction

An extensive body of research in auditing has examined the quality of auditors' judgments and found that auditors are poor judges when they assess the accuracy of their own knowledge and judgment (Smith and Kida 1991). Prior studies use calibration to measurethe appropriateness of confidencein subjective judgments anddocument a general pattern of overconfidence because auditors' subjective probability judgments are frequently more extreme than corresponding accuracy rates (Kennedy and Peecher 1997, Solomom et al. 1985, Tomassini et al. 1982). Poorly calibrated judgments in audit planning can demonstrate in either overconfidence or under-confidence. Overconfidence may lead to a failure to detect unexpected fluctuation in account balances and other financial relationships, compromising audit effectiveness.Under-confidence may lead to unnecessary investigation through excessive evidence gathering and substantive testing, compromising audit efficiency (Koonce 1993, Glover et al. 2005). Prior research has documented both overconfidence and under-confidence. Kennedy and Peecher (1997) investigate direct probability judgments (i.e. the probability that the predicted event will occur) and find auditors are overwhelmingly overconfident. However, when Tomassini et al (1982) ask auditors to provide confidence interval judgments (i.e. an interval offering a minimum and a maximum bound within which the actual value is expected to lie with a specified probability), they demonstrate that auditors are sometimes under-confident. Because these studies employ different decision contexts, their results are not directly comparable. In addition, the likelihood judgment expressed as a discrete subjective probability may differ froman expression of confidence interval. In this study we

ask auditors to perform the same audit task, but we use different methods to elicit confidence judgments so we can compare the discrete probability judgment and the confidence interval judgment to better understand the calibration of auditors' judgments and pinpoint the cause of mis-calibration.

Moreover, this study investigates whether feedback maybe effective in reducingmis-calibration and improving auditors' judgment quality. Auditors receive feedback on their judgments from reviewers and the environment (Solomon and Shields1995). For simple tasks like estimations or predictions, outcome feedback regarding the correct answer (i.e., from events that occur during or after the audit) is readily available (Bonner and Pennington1991). Outcome feedback provides a good learning opportunity for auditors and thus, is quite effective in improving auditors' judgment performance (Hirst and Luckett 1992, Bonner and Walker1994). However, many researchers are skeptical of the effectiveness of outcome feedback because of its high random noise, and instead, argue the superiority of performance feedback, which provides information about the ability to assign appropriate probabilities to outcomes and is effective in reducing overconfidence. However, performance feedback without accompanying outcome feedback is unlikely to improve calibration, because it provides no information to help determine whether or not a particular event will occur (Benson and Onkal 1992, Yates 1990). Thus, another goal of this study is to understand whether performance feedback provides any incremental effect that is above and beyond outcome feedback.

Results from our experiment suggest that auditors' confidence judgments are indeed poorly calibrated. Auditors are overconfident when they provide either direct probability judgments or interval judgments at the 90% confidence level. In general, we find that providing outcome feedback, a simple form of intervention, significantly improves the quality of auditors' judgments. We also find that performance feedback does not provide any added benefits. In addition, we find the extent of improvement in judgment calibration depends on the specific type of probability judgment. For direct probability judgments, our results show that the additional performance feedback actually reduces judgment quality as the amount of overconfidence reduction is less for performance plus outcome feedback auditors than the outcome alone auditors. For confidence interval judgments, we find outcome feedback and performance plus outcome feedback conditions are equally effective in reducing overconfidence.

2. Literature Review and Hypothesis Development

A general finding in psychological studies is that people are systematically overconfident about the accuracy of their knowledge and judgment, as their subjective probabilities are frequently more extreme than corresponding accuracy rates (Budescu et al. 1997, Yates1990). For example, when people express 95% confidence, they may be correct only about 80% of the time. These studies also find that the amount of overconfidence depends on the elicitation method (i.e., direct probability judgmentsor confidence interval judgments).

Direct probability judgments require estimates of the likelihood that the population value exceeds (or falls short of) some specified value (Keren 1991, Simith and Kida 1991). In direct probability judgments, people are considered accurate (or well calibrated) if the relative frequencies of true statements match the stated probabilities (e.g., 90% of all events assigned a probability of .9 should be correct). The judgment accuracy is often depicted by acalibration curve which plots the proportion of true (correct) items as a function of the judges' subjective probabilities. The 45 degree line represents perfect calibration, and points below (above) this line reflect over - (under -) confidence (Lichtenstein and Fischhoff 1977). The Brier score and its two components ---calibration (or reliability) and resolution ---- provides quantitative measures of the quality of these judgments (Brier 1950, Murphy 1973, Yates 1982, 1990). Over - /under- confidence is measured by subtracting the percentage of correct answers from the average probability judgment.

The confidence interval method requires estimates for specific percentiles of a population value or intervals that correspond to pre-stated probabilities (Keren 1991, Juslin et al. 1999). Accuracy is measured by the decision makers' hit, and surprise, rates. For example, when auditors are asked to provide interval predictions of account balances at the X% (e.g., X=90) confidence level, X% of the values should be within the interval, if they are perfectly calibrated. If the hit rate (or the proportion of values within the interval) is lower than X%, the decision makers are considered overconfident. Conversely, under-confidence is inferred when the hit rate is higher than the pre-stated probability.

An auditor is accurate, or well-calibrated, if for all predicted outcomes assigned a given probability, the proportion of those outcomes that occur (or the proportion correct) is equal to the probability (Tomassini et al.1982). When performing an audit task, auditors may express their confidence about their estimates of an outcomeby directly estimating the probability of the "true" value, or by providing an interval within which the true value would fall with a given probability. The two methods are formally equivalent and should yield identical conclusions, but empirical evidence has shown that the judgments based on these two methods are sometimes different (Budescu and Du 2007).

Therefore, we examine the quality of confidence judgments elicited by both methods. Prior auditing research employs different elicitation methods to examine the calibration of confidence judgments, and demonstrate a mixed pattern of over- and under- confidence. Twocalibration studies are particularly relevant to this study. Kennedy and Peecher (1997) rely on direct probability judgments toinvestigate auditors' assessments of their own and their subordinates' technical knowledge. They find that auditors overestimate the accuracy of their knowledge assessments, and are overwhelmingly overconfident. Tomassini et al. (1982) focus on the calibration of confidence interval judgments and ask auditors to specify the minimum and maximum values for 7 fractiles for accounts receivable balances (50, .25, .75, .10 .90, 0.01 and .99). They observe under-confidence at the 50% and 80% confidence intervals. These two studies differ in many ways, including the sample of auditors, the task, and the elicitation method, and thus, it is difficult to conclude the direction of calibration of auditors' judgments. In addition, in the same study conducted by Tomassini et al. (1982), overconfidence is found to be associated with the 98% confidence interval. It is likely that in confidence interval judgments, overconfidence dominates relatively high probability levels but under-confidence permeates at the lower probability levels. Therefore, in this study we focus on confidence interval at a relatively high probability of 90%. We expect direct probability judgments and 90% confidence interval judgments to be overconfident.

Another goal of this study is to identify apotential intervention to reduce any observed overconfidence.Auditors receive feedback on their judgments from reviewers and the environment (Solomon and Shields 1995). Outcome feedback is information about the realization of a previously predicted event. In many simple audit tasks, outcome feedback regarding the correct answer is available for tasks like simple predictions because the auditor can receive feedback about the quality of task performance from events that occur during or after the audit (Bonner and Pennington 1991). Prior research suggests that outcome feedback can aid judgment performance because it enables people to detect inappropriate judgment strategies and to learn appropriate strategies; through a process of intelligent strategy revision, outcome feedback enables better understanding of task knowledge, and consequently improves judgment performance (Bonner and Walker 1994, Hirst and Luckett 1992). Thus, we expect outcome feedback to improve the accuracy of auditors' judgments.

Other researchers suggest that compared to outcome feedback, performance feedback will be more effective in improving the calibration of probability judgments (Benson and Onkal 1992, Fischer 1982) and argue that knowledge of only the actual outcome – outcome feedback – is not sufficient for people to either understand the forces that have caused the outcome or know how to improve the prediction of future outcome (Benson and Onkal1992). In contrast, performance feedback provides judges with information about their ability to assign appropriate probabilities to outcomes, as well as information that people would find useful for calibrating their probability judgments to better reflect the relative frequency of occurrence of the predicted events (Benson and Onkal1992).

Performance feedback differs for different elicitation methods. For adirect probability estimate taskit includes the calibration curve, a component of Brier score and the over - /under - confidence index. The effects of performance feedback on improving judgment accuracy in discrete probability judgment are well documented. For example, Lichtenstein and Fischhoff (1980) and Stone and Opel (2000) document that performance feedback improves both the subjects' calibration and overconfidence performances in general-knowledge tasks. Similar results are obtained in Benson and Onkal's study (1992) where students make direct probability estimates for the outcomes of major college football games. Interestingly, virtually all of the improvement in calibration occurs in one step (Benson and Onkal 1992).

However, the evidence regarding performance feedback in the confidence interval task is quite limited. Performance feedback in this setting normally includes the hit or surprise rate. In a stock price forecasting task, Bolger and Onkal (2004) require their participants to specify ranges within which the true values would fall with a given probability, and find that the forecasts are initially overconfident but improve significantly after receiving performance feedback in the form of the hit rate.

Despite the claim of its superiority, performance feedback has not been examined in the auditing setting. It is not clear whether the different type of feedback may improve the quality of auditors' judgments differently. We state the hypothesis as follows:

H1: Auditors are overconfident when they provide 90% confidence interval judgments as well as direct probability judgments.

H2: Supplementing outcomefeedback with performance feedback is more effective than outcome feedback alone in reducing overconfidence.

3. Experimental Method

We test our hypotheses in an experiment. We manipulated the feedback type (outcome, performance plus outcome) as a between-subjects variable. Participants' primary task was to predict account balances based on past information. To assess the judgment quality, we measured overconfidence and calibration. A group of auditors in an international accounting firm participated in our study which relied on web-based survey technology. Half of them provided direct probability judgment and the other half confidence interval judgment.

3.1 Participants

The survey administrator announced the study via email to the firm's auditors. Participating auditors could click on a hyperlink in the email to be connected with the survey website, and then were randomly assigned to each of the four conditions in the study as they logged on. A total of 45 auditors started the survey, but only 38 auditors completed the instrument online during a one-month response period in June 2006. The other seven auditors accessed but did not start the survey, so they are not included in the analysis.

3.2 Task and Procedures

We selected eight financial statement items for twelve different companies and asked auditors to determine their future balances. We randomly selected twelve public companies from the Thomson database. For each company, we provided a brief description of the client's background information, industry and economic factors and a table of Year 1, Year 2 and Year 3 outcomes for eight financial statement items --- sales, operating income, net income, total assets and total liabilities, accounts receivable, inventory and accounts payable (see Appendix A for information about Company 1). After reading the background and past information, we asked participants to predict Year 4 balance for each item. In total, each participant made ninety six judgments (i.e., eight items x twelve companies).

The experiment consists of three stages. At stage one, participants predicted balances for a set of six companies. At stage two, participants received feedback. At stage three, they predicted balances for a different set of six companies. Participants could move back and forth between screens, but could not make change once they had entered their answers. Finally, participants finished by answering a short series of manipulation check questions.

3.3 Dependent and Independent Variables

3.3.1 Elicitation Method

We instructed half of the participants to make direct probability judgments, and the other half confidence interval judgments. For direct probability judgments, participants determined how probable it is that the balance for these ninety-six items in Year 4 will be higher than in Year 3. Specifically, we asked the participants to select a probability judgment from 0–100% for each item, given in increments of 10%, from the drop down menu. For confidence interval judgments, participants predicted account balances in Year 4. Instead of a single point estimate, we asked them to provide a range of possible numbers (including minimum and maximum) that would make them 90% sure.

3.3.2 Feedback Types

Within each elicitation method, half of the participants received outcome feedback and the other half received performance plus outcome feedback. For the outcome feedback group, we obtained the actual (audited) account balances in Year 4 from the database, and presented them to the participants in a tabulated form.

For the performance plus outcome feedback group, auditors received real-time analysis about the accuracy of their judgments in addition to the actual (audited) results (see Appendix B for an example of performance feedback in each elicitation condition). The performance feedback varies depending on the specific elicitation method. For direct probability judgments, auditors received three measures – calibration score, overconfidence score and calibration graph, to determine the degree of accuracy. We included discussions about the concept of calibration of probability judgments and explained how the calibration and overconfidence scores are computed and how the calibration curve is drawn. For confidence interval judgments, auditors received the hit rate, which is the percentage of actual (audited) results that fall within the ranges they had provided. Similarly, we provided a brief explanation about how the hit rate is calculated and how over- and under- confidence is determined.

4. Results

We first report the results for the probability judgment task, followed by the confidence interval task. Within each task, we first discuss the results as a group and then the analysis of individual auditors.

4.1 Direct Probability Judgments

In this task, participants provided a best probability estimate that the outcomes in Year 4 will be higher than in Year 3. They selected a probability judgment from 0–100% in increments of 10%. We grouped all probability judgments (a total of 2,112) in 11 categories (ranging from 0% to 100%), and simply counted the relative frequencies of items with outcomes in Year 4 that are higher than Year 3 in each category. Table 1 presents the distribution for the probability judgments. The row labeled "judgments" is the average judgment provided by the participants across all items in that category. The row labeled "% True" is the frequency of true events. Panel A includes 2,112 probability judgments for all groups. The modal category was 0.5 (535 responses), and the distribution slightly skews towards the right, indicating an overall tendency of overconfidence given the base rate of 50%. Probability judgments for the outcome and performance plus outcome feedback groups are presented in Panels B and C of Table 1.

<Insert Table 1 Here>

<Insert Figure 1 Here>

We constructed bivariate calibration plots of the relative frequencies as a function of the mean subjective probabilities for each of the 11 categories (see Figure 1). The abscissa includes the participants' probability judgment categories and the ordinate indicates the percentage of times the target event occurred for each judgment category. The diagonal line represents perfect calibration, as points on that line indicate that the judge is perfectly calibrated for each of his or her probability judgment categories (judgment of .30 occurs 30% of the time, etc.). Points far from the base rate (in this case, 50%) on the ordinate indicate good discrimination, as they discriminate between situations where the target event occurs from when it does not occur. The calibration analysis was performed by assigning items to the 11 categories (see Table 1). Figure 1, Panel A, presents the calibration plot based on all 2,112 judgments, along with the calibration score and the confidence score.

An inspection of Panel A of Figure 1 indicates that participants are overconfident, as the line connecting the points on the calibration plot is too horizontal. In the plot, the line crosses the diagonal at the 50% point, with the subjective probabilities lower than the observed relative frequencies from 0.0 to 0.5, but higher from 0.6 to 1.0by including referencepoints on both lines. This pattern indicates an overall overconfidence tendency, because participants overestimate the most likely hypothesis and assign probabilities that are consistently too *extreme* (i.e., too close to either 0 or 1) (Griffin and Brenner 2004). The analysis also shows a calibration score of 0.02. We also calculated the overconfidence score by subtracting the true frequency from probability judgment. A score of 0.04 indicates that the overall judgments by auditors are slightly overconfident.

In addition, we constructed the calibration plot for each feedback group (see Panels B and C of Figure 1). The performance plus outcome feedback group has a slightly lower overconfidence score than the outcome feedback group (0.02 vs. 0.04). Because the group overconfidence score is a combination of pre-feedback and post-feedback scores, it is impossible to detect any effect of feedback based on this score. To examine whether the specific type of feedback reduces overconfidence, we conducted calibration analysis for pre-feedback and post-feedback judgments within each feedback group. Table 2 compares the judgment quality before and after feedback within each feedback group. Indeed, judgments are better calibrated after feedback than before feedback as we observed a decrease in the mean overconfidence score after feedback in both feedback groups.

<Insert Table 2 Here>

Calibration refers to the match between the judgment categories and the percentage of times that the target event actually occurs. There are many ways in which a judge could exhibit poor accuracy, but the most commonly discussed one is overconfidence, where the judge overestimates his or her likelihood of being correct (Stone and Opel 2000). Therefore, we focus on the overconfidence score to test our hypotheses. We ran the calibration analysis for each auditor across all his or her judgments, and calculated the mean overconfidence score for each auditor. This generates a distribution of 23 individual overconfidence scores. Panel A of Table 3 presents the mean pre- and post-feedback individual overconfidence score by the type of feedback.

<Insert Table 3 Here>

There are 11 auditors in the outcome feedback group (with an average overconfidence score of 0.02), and 12 in the outcome plus performance feedback group (with an average overconfidence score of 0.06). After collapsing across all cells in Table 3, we have an average mean overconfidence score of 0.04, indicating an overall tendency of overconfidence. A perfectly accurate auditor should have an overconfidence score of 0. To test H1, we compared the mean individual overconfidence score of 0.04 to the test value of 0 (no overconfidence). Panel B of Table 3 presents

the one-sample t test, which indicate that auditors (n=23) are indeed overconfident (t=2.17, p<0.05). The result supports H1.

Next, we investigate whether feedback reduces overconfidence. We calculate the change index for each auditor by taking the difference between post-feedback and pre-feedback overconfidence scores. This gives us a distribution of 23 individual change indices, with 11 in the outcome feedback group and 12 in the outcome plus performance feedback group. If feedback has no effect on judgment quality, we should expect a change index of 0, i.e., no difference between pre-feedback and post-feedback scores. We compare the change index in each feedback group to the test value of 0 (i.e., no change). Panel C of Table 3 presents results of one-sample t-tests. In the outcome feedback condition, the change index is marginally different from 0 (mean = -0.08, t = -1.59, one-sided p = 0.07), implying the outcome feedback is marginally effective in reducing overconfidence. However, in the performance plus outcome feedback condition, the change index is not statistically different from 0 (mean = 0.01, t = 0.53, one-sided p = 0.31), suggesting the performance plus outcome feedback is not effective in improving judgment quality at all. In addition, we conducted a one-way ANOVA on the change index with the type of feedback as the independent variable (see Panel D of Table 3). Surprisingly, we find that outcome feedback is more effective than outcome plus performance feedback in terms of improving judgment accuracy and reducing overconfidence ($F_{(1, 21)}$= 2.86, one-sided p=0.05).

The above analysis suggests that auditors are overconfident when providing direct probability judgments. However, results fail to support H2, because judgment performance decreases when auditors receive performance feedback along with outcome feedback for direct probability judgments.

4.2 Confidence Intervals

The participants provided interval estimates with a pre-stated probability at 90%. We calculated the proportion of judgments that were in fact correct. Participants are considered accurate if this fraction matched the stated confidence of 90%. If confidence is higher (lower) than the actual hit rate, the participants are said to be overconfident (under-confident). For all confidence interval judgments (1,272 in total), the hit rate is only 52%, much lower than 90% for perfect calibration, indicating a pattern of overconfidence. Table 4 shows the pre- and post- feedback hit rate for each feedback group. The outcome feedback group is more overconfident (49% hit rate) than the performance plus outcome feedback group (55% hit rate). An examination of the change within each group (the difference between pre-and post-feedback) indicates that the outcome feedback increases the hit rate by 11 %, and the performance plus outcome feedback increases the hit rate by 14%.

<Insert Table 4 Here>

<Insert Table 5 Here>

Moreover, we analyze the data at the individual level to detect any significant effects. Table 5 present descriptive statistics and analysis for hypothesis testing. We calculated the average overconfidence score (90% minus hit rate) for each auditor across all his/her confidence interval judgments, and generated a distribution of 15 individual overconfidence scores. Table 5, Panel A, presents the mean overconfidence scores. There are 8 participants in the outcome feedback group with an average overconfidence score of 35% and 7 auditors in the performance feedback group with an average overconfidence score of 39%. Apparently, auditors are overconfident.

If there is no overconfidence, the overconfidence score should be 0. Thus, we use 0 as the test score for assessing overconfidence. Panel B of Table 5 indicates that the mean average overconfidence score of 37% is significantly greater than 0 (t = 6.64, p<0.05). This supports Hypothesis 1 that auditors are overconfident. In addition, if feedback has no effect on overconfidence, we should observe no difference between the pre- and post- feedback overconfidence scores. Thus, we calculated the change index by taking the difference of the pre- and post-overconfidence scores for each of the 15 auditors. This gives us a distribution of 15 individual change indices, 8 in the outcome feedback group and 7 in the performance plus outcome feedback group. For each group, we compare the mean of the change index to the test value of 0 (no change). The one-sample t test indicates that the mean change index is significantly different from the test-value of 0 for both groups. For the outcome feedback group, the mean change is -0.12 (t = -2.74, p< 0.05). For the performance plus outcome feedback group, the mean change is -0.10 (t = -2.64, p< 0.05). Apparently, both types of feedback are effective in reducing overconfidence. In addition, we perform a one-way ANOVA for the change index across the two feedback groups (n=15) to investigate which type of feedback is more effective in overconfidence reduction. The mean change index of the outcome feedback group is not significantly different from the mean of the performance plus outcome feedback group ($F_{(1,13)}$ = 0.14, p = 0.72). Apparently, both groups are equally effective in overconfidence reduction. The results fail to support H2.

In summary, our results strongly support H1 as auditors are indeed overconfident when performing a simple analytical procedure task. However, results fail to support H2. Apparently, performance feedback does not provide any benefits beyond and above the outcome feedback.

5. Discussion and Conclusions

This study examines the accuracy of auditors' probability judgments and the impact of feedback on judgment in an experiment. We asked participants to provide probability judgments elicited by two different methods, i.e., direct probability and confidence interval. Outcome feedback and performance plus outcome feedback were manipulated as a between-subject variable. Auditors from an international accounting firm participated in our web-based study, where they predicted financial outcomes. We measured the accuracy of probability judgments by the correspondence between an auditor's predictions of account balances and outcome realizations. Twelve case scenarios were developed based on real companies and an audited (or true) account balance was known in each case. Thus, the probability judgment assessed could be compared to these known balances to determine its calibration. The use of real companies as case scenarios provides a readily observable external criterion of realized outcomes (audited values) to measure judgment calibration, ex post, in a variety of ways including calibration and overconfidence.

Our evidence shows that, in general, auditors are not very accurate in assessing their own judgments. They are overconfident when they provide 90% confidence interval judgments as well as direct probability judgments. Our results are consistent with prior findings, and demonstrate that overconfidence, as a general phenomenon, persists regardless of the task context. In addition, we investigate whether we can use feedback as a means of improving judgment quality. Our evidence indicates that outcome feedback, in spite of its simplicity, is effective in reducing overconfidence. In addition, we find providing performance does not have any added benefits. Sometimes, it actually decreases judgment performance. Our evidence suggests that performance feedbackmay not be effective for simple tasks such as prediction outcomes. Our results have practical implications because outcome feedback is common and readily available for auditors in the audit setting. To the extent that auditors can rely on outcome feedback as a means for improving judgment quality, auditors, despite their initial mis-calibration, will be able to correct their own judgments and decisions over time when they gain experience and learn from the audit environment. Our evidence also suggests that auditing firms should provide timely feedback on auditors' performance.

Nevertheless, our study has multiple limitations. First, prior studies suggest that inaccuracy in probability judgments can be attributed to systematic cognitive biases and / or the "noise" (the stochastic component) associated with the judgment process (Budescu et al. 1997). Our experiment only allows us to document the deficiency in auditors' confidence judgment, but it does not provide evidence on "why" this deficiency occurs and does not provide information about different types and amounts of bias or errors related to different elicitation methods. Future studies should investigate the underlying judgment process to suggest possible explanations for the occurrence of overconfidence in the first place. Secondly, the decreased performance in the performance plus outcome condition may be driven by the complexity of performance feedback. Compared to the simple and clear structure of outcome feedback, calibration feedback is complicated and includes multiple pieces of information, e.g., calibration curve, calibration score, overconfidence index. The complexity of calibration feedback may result in cognitive overload for auditors, and potentially deter its effectiveness. Future studies should either reduce the amount of information included in calibration feedback or increase the number of feedback sessions to give auditors sufficient time to understand the concepts of calibration. Finally, we only elicited interval judgments at the 90% confidence level. Prior studies find that probability judgments for different confidence intervals are not equally mis-calibrated (Budescu and Du 2007, Tomassini et al. 1982), so we cannot generalize our overconfidence finding to other confidence levels. Future studies should investigate whether overconfidence in analytical procedures persists in various confidence levels to obtain a complete picture of the quality of auditors' confidence judgments.

References

Benson, P. G., & Onkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting, 8,* 559-573. http://dx.doi.org/10.1016/0169-2070(92)90066-I

Bolger, F., & Onkal, D. (2004). The effects of feedback on judgmental interval predictions. *International Journal of Forecasting, 20,* 29-39. http://dx.doi.org/10.1016/S0169-2070(03)00009-8

Bonner, S. E., & Penningtonm, N. (1991). Cognitive processes and knowledge as determinants of auditor expertise. *Journal of Accounting Literature, 10,* 1-50.

Bonner, S.E., & Walker, P. L. (1994, January). The effects of instruction and experience on the acquisition of auditing knowledge. *The Accounting Review, 69,*157-178.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1-3. http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO; 2

Budescu, D. V., & Du, N. (2007). The Coherence and consistency of investors' probability judgments. *Management Science*, forthcoming. http://dx.doi.org/10.1287/mnsc.1070.0727

Budescu, D. V., Erev, I., Wallsten, T. S., & Yates, J. F. (1997). Introduction to special issue: Stochastic and cognitive models of confidence. *Journal of Behavioral Decision Makin,g 10*, 153-285. http://dx.doi.org/10.1002/(SICI)1099-0771(199709)10:3<153::AID-BDM279>3.0.CO;2-K

Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavioral and Human Performance, 14*, 352-369. http://dx.doi.org/10.1016/0030-5073(82)90250-1

Glover, S. M., Prawitt, D. F., & Wilks, T. J. (2005). Why do auditors over-rely on weak analytical procedures? The role of outcome and precision. *Auditing:A Journal of Practice and Theory, 24*, 197-220.

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. In Nigel Harvey, & Derek Koehler (Eds.), *Blackwell Handbook of Judgment and Decision Making.* Oxford, Blackwell. http://dx.doi.org/10.1002/9780470752937.ch9

Hirst, D. E., & Luckett, P. F. (1992). The relative effectiveness of different types of feedback in performance evaluation. *Behavioral Research in Accounting, 4*, 1-22.

Kennedy, J., & Peecher, M. E. (1997). Judging auditors' technical knowledge. *Journal of Accounting Research, 35*(Autumn), 279-293. http://dx.doi.org/10.2307/2491366

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77, 217–273. http://dx.doi.org/10.1016/0001-6918(91)90036-Y

Koonce, L. (1993). A Cognitive Characterization of Audit Analytical Review. *Auditing:A Journal of Practice and Theory, 12*, 57-76.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about howmuch they know?: The calibration of probability judgments. *Organizational Behavior and Human Performance, 20*, 159–183. http://dx.doi.org/10.1016/0030-5073(77)90001-0

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, 26, 149-171. http://dx.doi.org/10.1016/0030-5073(80)90052-5

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595- 600. http://dx.doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

Smith, J. F., & Kida, T. (1991). Heuristics and biases: Expertise and task realism in auditing. *Psychological Bulletin*, *109*(3), 472-489. http://dx.doi.org/10.1037/0033-2909.109.3.472

Solomon, I., Ariyo, A., & Tomassini, L. A. (1985). Contextual effects on the calibration of probabilistic judgments. *Journal of Applied Psychology*, *70*(3), 528-532. http://dx.doi.org/10.1037/0021-9010.70.3.528

Solomon, I., & Shields, M. D. (1995). Judgment and decision-making research in auditing. In R. H. Ashton, & A. H. Ashton (Eds.), *Judgment and Decision-Making Research in Accounting and Auditing*. Cambridge, MA: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511720420.008

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: the effects of performance and environmental feedback. *Organizational Behaviorand Human Decision Processes, 83*, 282-309. http://dx.doi.org/10.1006/obhd.2000.2910

Tomassini, L. A., Solomon, I., Romney, M. B., & Krogstad, J. L. (1982). Calibration of auditors' probabilistic judgments: Some empirical evidence. *Organizational Behavior and Human Performance, 30*, 391-406. http://dx.doi.org/10.1016/0030-5073(82)90227-6

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30*, 132-156. http://dx.doi.org/10.1016/0030-5073(82)90237-9

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

Notes

Note 1. For example, 1) the calibration graph plots the frequencies of true items as a function of the probability judgments, and the 45 degree line represents perfect calibration, and points below or above this line reflect imperfect

calibration. 2) A calibration score of 0 indicates perfect calibration, and a number other than 0 indicates participants are not well calibrated or their judgments are not accurate. 3) A positive confidence score (subtract frequency from probability judgment) indicates that participants are over-confident, and a negative confidence score under-confident.

Note 2. Please note that the analysis was conducted on aggregated, not independent, judgments of a particular category, and these judgments provide a useful means for evaluating judgment performance.

Note 3. N is the number of items judged; pi is the probability assigned to item i; and di =0 (if the item is false) or 1 (if it is true). Calibration is calculated based on the partition of [0, 1] interval into J distinct intervals:

$$N^{-1} \sum_{j=1}^{J} N_j d_j (f_j - d_j)^2$$

where fj is the jth probability value (or in our case the mean judgments of each category), and dj is the fraction of items assigned to the jth interval that are true (di =1). The calibration score is referred to by Murphy (1973) as reliability and by Lichtenstein and Fischhoff (1977) as calibration, and measures the goodness of fit between the probability assessments and the corresponding proportions of correct responses (or the deviation of the calibration curve from the 45 degree line) (Keren 1991).

Table 1. Distribution of direct probability judgments for all participants

Panel A: All Groups (n = 2,112)											
Judgments	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Proportion (of 2,112)	29	47	110	170	274	535	394	254	179	73	47
% True (in each judgment category)	0.31	0.32	0.4	0.39	0.41	0.52	0.59	0.54	0.56	0.59	0.47
Panel B: Outcome Feedback Group (n =1,152)											
Judgments	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Proportion (of 2,112)	4	21	60	95	183	217	269	139	118	40	6
% True (in each judgment category)	0.25	0.24	0.43	0.41	0.46	0.52	0.59	0.47	0.53	0.50	0.67
Panel C: Performance plus Outcome Feedback Group (n = 960)											
Judgments	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Proportion (of 2,112)	25	26	50	75	91	318	125	115	61	33	41
% True (in each judgment category)	0.32	0.38	0.36	0.36	0.31	0.52	0.58	0.63	0.64	0.70	0.44

(The row labeled “judgments” is the average judgment provided by the participants across all items in that category. The row labeled “% True” is the frequency of items with outcomes in Year 4 that are indeed higher than in Year 3).

Table 2. Quality of direct probability judgments (Means) (n=2,112)

Means	**Outcome Feedback**			**Performance plus Outcome Feedback**	
	Before	After		Before	After
Calibration	0.05	0.01		0.04	0.03
Overconfidence	0.09	-0.01		0.06	0.01

(Calibration is calculated based on the partition of [0, 1] interval into J distinct intervals. N is the number of items judged; p_i is the probability assigned to item *i;* and d_i =0 (if the item is false) or 1 (if it is true).

$$N^{-1}\sum_{j=1}^{J} N_j d_j (f_j - d_j)^2$$

where f_j is the jth probability value (or in our case the mean judgments of each category), and d_j is the fraction of items assigned to the jth interval that are true (d_i =1).

Overconfidence is calculated by subtracting the true frequency from the probability judgment in each category. A positive sign means overconfidence and a negative sign indicates under-confidence).

Table 3. Tests of hypotheses for probability judgments

Panel A: Mean Overconfidence Score for Individual Auditor

	Outcome Feedback (n=11)	Performance plus Outcome Feedback (n=12)	Mean Across Pre- and Post- Feedback
Pre-Feedback	0.06	0.05	0.055
Post-Feedback	-0.02	0.06	0.02
Mean across Feedback Type	0.02	0.055	0.04

Panel B: Test of H1

Compare Overconfidence Score with the test value of 0 (one sample t-test)

t	*df*	*Mean Difference*	*P (2-sided)*
2.17	22	0.04	0.04

Panel C: Test of H2

Compare Change Index with the test value of 0 (one sample t-test) by Feedback Type

	t	*df*	*Mean Difference*	*P (1-sided)*
Outcome Feedback	-1.59	10	-0.08	0.07
Performance plus outcome Feedback	0.53	11	0.01	0.31

One-Way ANOVA with Change Index as the Dependent Variable

Source	*SS*	*df*	*MS*	*F*	*P (1-sided)*
Feedback Type	0.04	1	0.04	2.86	0.05
Within Cells (Error)	0.33	21	0.15		

(The overconfidence score is calculated by subtracting the true frequency from the probability judgment).

Table 4. Calibration of all confidence interval judgments (n=1,272)

		Hits	Surprises	Total
Outcome Feedback	Pre-feedback	170 **(44%)**	182	352
	Post-feedback	208 **(55%)**	128	336
Total		378 **(49%)**	310	688
Performance plus outcome Feedback	Pre-feedback	129 **(48%)**	167	296
	Post-feedback	157 **(62%)**	131	288
Total		286 **(55%)**	298	584

(Hit (or surprise) rate is the percentage of true values falling within (or outside) the lower and upper bounds provided by auditors).

Table 5. Tests of hypotheses for confidence interval judgments

Panel A: Mean Individual Overconfidence Score

	Outcome Feedback (n=8)	Performance plus Outcome Feedback (n=7)	Mean Across Pre- and Post- Feedback
Before	41%	44%	42.5%
After	29%	34%	31.5%
Mean across Feedback Type	35%	39%	37%

Panel B: Test of H1

Compare Overconfidence Score with the test value of 0 (one sample t-test)

t	*df*	*Mean Difference*	*P (2-sided)*
6.64	14	0.37	0.00

Panel C: Test of H2

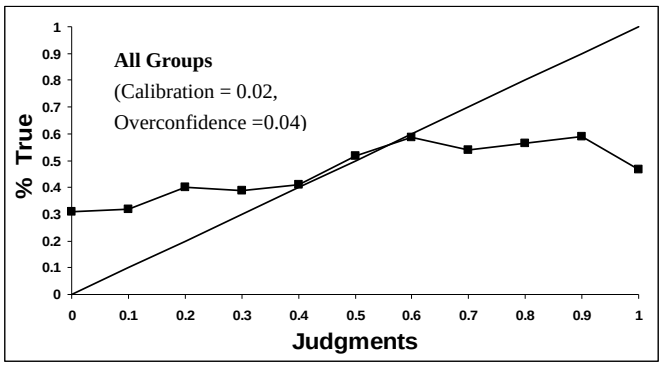
Compare Change Index with the test value of 0 (one sample t-test) by Feedback Type

	t	*df*	*Mean Difference*	*P (2-sided)*
Outcome Feedback	-2.74	7	-0.12	0.03
Performance plus Outcome Feedback	-2.64	8	-0.10	0.04

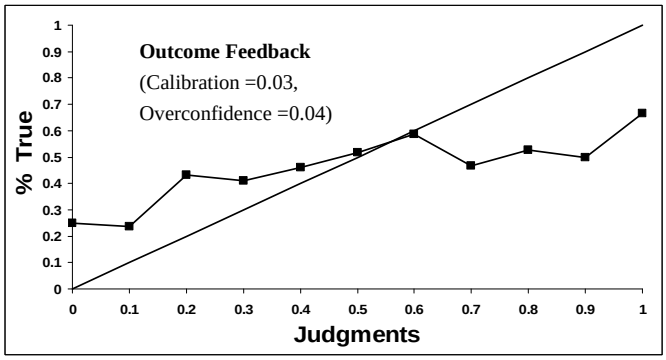
One-way ANOVA with Change Index as the Dependent Variable

Source	*SS*	*df*	*MS*	*F*	*P (2-sided)*
Feedback	0.00	1	0.00	0.14	0.72
Within Cells (Error)	0.18	13	0.01		

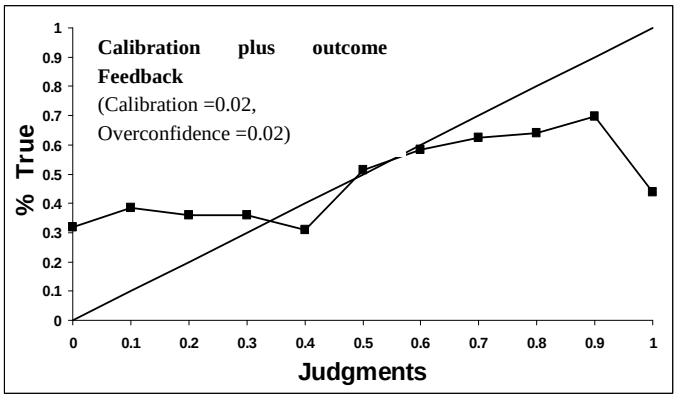
(The overconfidence score is calculated by subtracting the percentage correct (or the hit rate) from the target 90%).



Panel A. Calibration for all judgments (n=2,112)



Panel B. Calibration for the outcome feedback group (n=1,152)



Panel C. Calibration for the performance plus outcome feedback group (n=960)

Figure 1. Calibration plots for probability judgments

(% True is the frequency of items with outcomes in Year 4 that are higher than in Year 3).

Appendix A

Background Information for Company 1

Co.1 is an automotive parts manufacture, supplying a broad range of integrated systems, modules, and components for light vehicles, commercial trucks, trailer and specialty original equipment manufacturers globally. With a fluctuating economy, rising fuel prices and global uncertainty, it continues to be a demanding time for the automotive industry. Co. 1 experienced a stronger global commercial vehicle market recently. The company divested the Light Vehicle Aftermarket and coil coating business.

Direct Probability Estimate

Please indicate the probability that results in Year 4 will be higher than in Year 3.

($ in millions)	**Year 1**	**Year 2**	**Year 3**	**Year 4**
Sales	**$6,805**	**$6,882**	**$7,788**	**_____%**
Operating Income	**$279**	**$352**	**$309**	**_____%**
Net Income	**$35**	**$149**	**$140**	**_____%**
Total Assets	**$4,243**	**$4,464**	**$4,970**	**_____%**
Total Liabilities	**$3,466**	**$3,665**	**$4,007**	**_____%**
Accounts Receivable	**$965**	**$1,251**	**$1,340**	**_____%**
Inventory	**$457**	**$458**	**$543**	**_____%**
Accounts Payable	**$1,054**	**$1,150**	**$1,311**	**_____%**

Confidence Interval Estimate

Please provide an interval estimate (including lower and upper bounds) for Year 4.

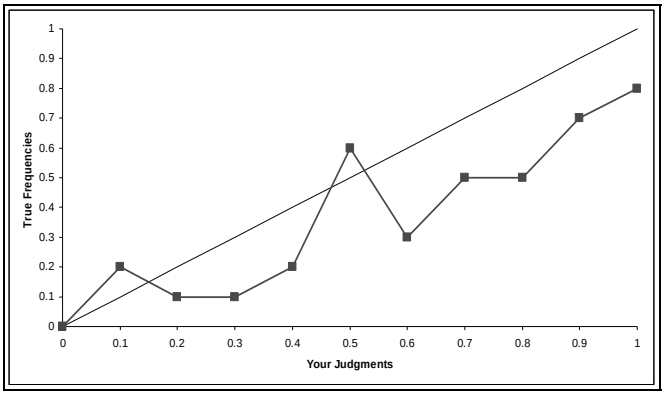
($ in millions)	**Year 1**	**Year 2**	**Year 3**	**Year 4**	
Sales	**$6,805**	**$6,882**	**$7,788**	**Lower Bound (from)**	**Upper Bound (to)**
Operating Income	**$279**	**$352**	**$309**	**$_____**	**$_____**
Net Income	**$35**	**$149**	**$140**	**$_____**	**$_____**
Total Assets	**$4,243**	**$4,464**	**$4,970**	**$_____**	**$_____**
Total Liabilities	**$3,466**	**$3,665**	**$4,007**	**$_____**	**$_____**
Accounts Receivable	**$965**	**$1,251**	**$1,340**	**$_____**	**$_____**
Inventory	**$457**	**$458**	**$543**	**$_____**	**$_____**
Accounts Payable	**$1,054**	**$1,150**	**$1,311**	**$_____**	**$_____**

Appendix B

Performance Feedback – Direct Probability Judgment

Here is the analysis of your judgments

The calibration score is 0.04, the confidence score is 0.14



The analysis indicates that you are not very well calibrated and are over-confident!

Performance Feedback – Confidence Interval Judgment

Here is the analysis regarding your estimates

72% of the actual results fall between the lower and upper bounds you provided, and 18% fall outside the given interval.

Your ranges are too narrow (narrow, wide, just right), and therefore, you are over-confident (Over-confident, under-confident, and well-calibrated)