Received: 27 December 2023 | Revised: 9 July 2024 | Accepted: 16 July 2024

DOI: 10.1002/acp.4236

RESEARCH ARTICLE

WILEY

The effect of calibration training on the calibration of intelligence analysts' judgments

Megan O. Kelly[1] | David R. Mandel[1,2]

[1]Toronto Research Centre, Defence Research and Development Canada, Toronto, Ontario, Canada

[2]Department of Psychology, York University, Toronto, Ontario, Canada

Correspondence
David R. Mandel, Toronto Research Centre, Defence Research and Development Canada, Toronto, ON, Canada.
Email: david.mandel@forces.gc.ca

Present address
Megan O. Kelly, Department of Psychology, University of Waterloo, Waterloo, Canada.

Funding information
Canadian Safety and Security Program, Grant/Award Number: #2018-TI-2394; Accelerate Command, Control, and Intelligence Project Activity, Grant/Award Number: #AC2I-019

Abstract

Experts are expected to make well-calibrated judgments within their field, yet a voluminous literature demonstrates miscalibration in human judgment. Calibration training aimed at improving subsequent calibration performance offers a potential solution. We tested the effect of commercial calibration training on a group of 70 intelligence analysts by comparing the miscalibration and bias of their judgments before and after a commercial training course meant to improve calibration across interval estimation and binary choice tasks. Training significantly improved calibration and bias overall, but this effect was contingent on the task. For interval estimation, analysts were overconfident before training and became better calibrated after training. For the binary choice task, however, analysts were initially underconfident and bias increased in this same direction post-training. Improvement on the two tasks was also uncorrelated. Taken together, results indicate that the training shifted analyst bias toward less confidence rather than having improved metacognitive monitoring ability.

KEYWORDS
bias, calibration, intelligence analysts, overconfidence, training, underconfidence

1 | INTRODUCTION

Across many domains, experts are expected to make judgments to support the decision-making of their organizations. Such judgments may regard the probability that a given claim is or will become true (e.g., "There is an *x*% chance that China will invade Taiwan by 2025") or as the subjective confidence interval for some continuous quantity (e.g., "We assess with 90% confidence that between *x* and *y* US companies will report ransomware attacks in 2025").[1] *Calibration* is the degree to which confidence coincides with judgment accuracy (Keren, 1991). An assessor is perfectly calibrated if the proportion of judgments assigned a given probability equals the proportion that is true (Lichtenstein & Fischhoff, 1980). Across a series of judgments assigned 80% probability, 80% should be true, and across a series of interval judgments made with 90% confidence, 90% of true values should fall within the relevant intervals. Provided the various implications of judgments, (e.g., financial, medical, safety, etc.), the degree to which an assessor is calibrated (or not) is a vital aspect in understanding the potential validity or utility of a judgment.

1.1 | Miscalibration

In contrast to perfect calibration, an assessor may demonstrate *miscalibration* if the proportion of judgments assigned a given probability *exceeds* the proportion of true judgments, indicating an *overconfidence* bias, or *falls below* the proportion of true judgments, indicating an *underconfidence* bias (Lichtenstein & Fischhoff, 1980). In the case of interval judgments, overconfidence manifests as *overprecision* with assessors providing excessively narrow confidence intervals, whereas

underconfidence is expressed as *underprecision* or excessively wide intervals (Du & Budescu, 2007; Moore & Healy, 2008).

Miscalibration can undermine accurate probability assessment (Mandel & Barnes, 2014), information processing (Meyer & Singh, 2017; Zacharakis & Shepherd, 2001) and decision-making quality (Berner & Graber, 2008; Biais et al., 2005). Unfortunately, miscalibration is widespread among experts and nonexperts alike (Arkes, 2001; Bier, 2004; Lichtenstein & Fischhoff, 1980; McKenzie et al., 2008), and is common across a variety of domains including healthcare and medical science (Benjamin et al., 2022; Berner & Graber, 2008; Brinkman et al., 2015; Meyer et al., 2013), geopolitical forecasting (Chang & Tetlock, 2016; Mandel & Barnes, 2018; Tetlock, 2005), the military (Kelly et al., 1975; Phelps et al., 1980), finance (Ben-David et al., 2013; Biais et al., 2005; Du & Budescu, 2007), meteorology (Charba & Klein, 1980; Murphy & Winkler, 1984), sports betting (Bum et al., 2018; Erceg & Galić, 2014; Johnson & Bruce, 2001) and education (Callender et al., 2016; Foster et al., 2017; Huff & Nietfeld, 2009).

Miscalibration often takes the form of overconfidence (Lichtenstein et al., 1982; Russo & Schoemaker, 1992) including among corporate executives predicting market performance (Ben-David et al., 2013), clinicians making medical diagnoses (Berner & Graber, 2008), oncologists forecasting the efficacy of clinical trials (Benjamin et al., 2022), and geopolitical forecasters (Tetlock, 2005). Oft-cited exceptions include meteorologists (Charba & Klein, 1980; Murphy & Winkler, 1984), racetrack bettors (Johnson & Bruce, 2001), and experienced bridge players (Keren, 1987), all of whom receive timely feedback on their judgments and exhibit excellent calibration.

In the intelligence domain, there is evidence that analysts exhibit both forms of miscalibration. One early study found that US intelligence analysts exhibited a "small but consistent bias towards overconfidence" when forecasting military-relevant events (Kelly et al., 1975 as cited in Phelps et al., 1980, p. 11). An analysis of over 2000 forecasts found miscalibration in the form of underconfidence (Mandel & Barnes, 2014, 2018) and related work found underconfidence in analysts was most pronounced when accountability pressures were likely to be especially high (Mandel et al., 2014; Mandel & Barnes, 2014). While overconfident assessments may undermine decision-making by encouraging excessive risk-taking or a misperception of risk, underconfidence, too, can undermine decision-making by reducing the informativeness of intelligence (Mandel & Barnes, 2014). Estimates couched in more uncertainty than warranted can dilute the signal value of those estimates.

1.2 | Improving calibration

Given the potential for calibration to signal judgment accuracy, many have aimed to improve calibration, including through intervention (Gruetzemacher et al., 2023; Lawrence et al., 2006; Lichtenstein et al., 1982). For example, one common approach to improving calibration involves providing *performance* or *outcome* feedback to participants with the intent to help correct biases. Performance feedback compares actual and expected results, whereas outcome provides the correct answers and a total score. Several studies suggest that performance feedback can improve calibration in a variety of judgment tasks (Adams & Adams, 1958, 1961 Benson & Onkal, 1992; Bolger & Onkal-Atay, 2004; Lichtenstein & Fischhoff, 1980; Onkal & Muradoglu, 1995; Stone & Opel, 2000; although, see Sharp et al., 1988 and Doussau et al., 2023 for exceptions), including exam performance in classroom settings (Saenz et al., 2019) and geopolitical forecasting (Mellers et al., 2014; Moore et al., 2017; Tetlock & Gardner, 2015). Moore et al. (2017) found that a training intervention that included providing performance feedback was associated with improved calibration driven by reduced overconfidence.

There is some evidence that outcome feedback can improve calibration (Callender et al., 2016; Du et al., 2012; Niu & Harvey, 2022; O'Connor & Lawrence, 1989; although see Foster et al., 2017). Du et al. (2012) found that combining outcome and performance feedback was no more effective at improving calibration than outcome feedback alone. Other studies suggest that providing outcome feedback can precipitate overcorrection from overconfidence toward underconfidence (Arkes et al., 1987) or contribute to overconfidence in the form of hindsight bias (Hoch & Loewenstein, 1989), thus, outcome feedback might be less effective than performance feedback (Lawrence et al., 2006).

The effects of feedback-centered interventions on calibration demonstrate that feedback can improve calibration, although not necessarily consistently across contexts. Successful calibration training may require additional approaches which we aim to address currently. A third approach to improving calibration is *metacognitive* training, which broadly encompasses interventions designed to facilitate reflective thinking, self-questioning, and awareness of biases relating to calibration (Busby et al., 2018; Jaspan et al., 2022). While some studies show little evidence of improved calibration (Alpert & Raiffa, 1982; Emory & Luo, 2022), others demonstrate improvement (Guiterrez & Schraw, 2015; Huff & Nietfeld, 2009; Kruger & Dunning, 1999; Nietfeld & Schraw, 2002), Nietfeld et al. (2006) found that calibration could be improved by combining metacognitive training with performance feedback. Metacognitive training can also be augmented with instruction on specific debiasing strategies, such as techniques that force subjects to consider alternative hypotheses or generate reasons why their judgments may be wrong (Busby et al., 2018; Hoch, 1985; Jaspan et al., 2022; Koriat et al., 1980; Veinott et al., 2010). However, elicitation techniques based on similar principles, such as encouraging individuals to consider alternatives to their "best guess", have not received much support (Mandel et al., 2020).

Recent years have seen the proliferation of commercial training courses that draw on the research outlined above as well as from the companies' samples (e.g., the *Calibrated Probability Assessments* (CPA) course from Hubbard Decision Research, 2019 and the Superforecasting Fundamentals course by Good Judgment Incorporated, 2022). The increased availability of these courses coincides with calls for the intelligence community to explore training aimed at improving judgment accuracy including calibration skill and better enabling the systematic monitoring of judgment accuracy and calibration (Chang & Tetlock, 2016; Dhami & Mandel, 2021; Dhami et al., 2015;

Friedman, 2019; Mandel, 2015; Mandel, 2019; Mandel & Irwin, 2021; Rieber, 2004). Intelligence analysts have long received instruction on analytic techniques meant to improve calibration, but few of these methods have undergone rigorous empirical testing (Chang et al., 2018). Moreover, analysts do not receive timely or comprehensive performance feedback, if they receive it at all (Moore, 2011; Rieber, 2004).

To our knowledge, no study has examined the effect of commercial training on calibration skills among professional intelligence analysts despite them being experts who routinely produce probabilistic judgments and who may respond differently to calibration training than convenience samples (Hubbard, 2014). Moreover, experimental results from experts in other domains may not generalize to the national security and intelligence domain. To the extent that commercial training improves calibration skill among professional analysts, these findings can help inform intelligence organizations (and other expert communities that routinely issue probabilistic assessments) considering whether to provide calibration training to their analysts.

1.3 | The present investigation

Using a pre-post design, we assessed the effect of a commercial training course on calibration skill in a sample of Canadian intelligence analysts. The online self-paced course, *CPA*, typically takes 3–4 h to complete and comprises six video modules during which participants receive metacognitive training and techniques to improve their calibration (Hubbard Decision Research, 2019). Throughout, participants complete calibration tests and receive outcome *and* performance feedback. Performance is also visualized alongside the average performance of past participants, enabling social comparison. Each calibration test comprises two sets of general knowledge questions corresponding to the two general types of subjective estimates routinely produced by experts (Hubbard Decision Research, 2019): One set of binary statements (e.g., "Mars is farther away from Earth than Venus.") wherein participants judge whether the statement is true/false along with a confidence rating of 50%–100% and one set of interval estimation questions wherein participants must provide lower- and upper-bound estimates (e.g., "In what year was William Shakespeare born?") at 90% confidence.

During the course, participants build toward a four-step "calibration process" which participants are instructed to apply to each judgment. The general process involves (1) choosing an initial estimate (for interval questions, starting with wide ranges to avoid overprecision), (2) applying the equivalent bet test (meant to quantify uncertainty, Hubbard, 2014, pp. 102–106), (3) assuming the judgment is wrong and giving plausible reasons as to why (i.e., Klein's, 2008 premortem), and finally, (4) applying any recommended adjustments based on feedback after each calibration test.

Given extant calibration training research and that HDR's CPA course combines several evidence-based techniques over a relatively long timeframe, we hypothesized that training would improve both (1) the calibration of intelligence analysts' point estimates and (2) the calibration of their range estimates. Day-to-day, intelligence analysts may be incentivized to dilute the certainty of their assessments as a blame avoidance strategy (Gentry, 2017; Mandel & Irwin, 2021). However, these accountability pressures were largely absent during our experiment. Thus, we further hypothesized that (3) in contrast to results from Mandel and Barnes (2014, 2018), analysts would tend to exhibit overconfidence during the baseline calibration test.

2 | METHOD

Data S1 including the data, analyses, and other supporting files are available from the Open Science Foundation (OSF) project page: osf.io/yqxpw/.

2.1 | Participants

The experiment was completed by 70 Canadian intelligence analysts (34% female) aged 22 to 73 ($M = 35.6$, $SD = 11.33$) and years of experience in the intelligence community ranged from 0 to 24 years ($M = 4.98$, $SD = 5.60$). Seventy-three percent of participants reported prior statistics or probability theory experience, but no participant reported experience with HDR's CPA course before the study.[2] Eighty-six percent of participants were of civilian rank with the remaining participants reporting as senior officer (7%), junior officer (3%), junior non-commissioned member (1%), or unspecified (3%). The highest education attained by participants included doctoral degree (6%), master's degree (54%), undergraduate degree (36%), trade school (non-military; 3%), and high school or equivalent (1%).

2.2 | Design

Participants completed both pre-training and post-training experimental tasks and within each, engaged in both a binary judgment task and an interval judgment task. The experiment was a fully within-subjects 2 (Training: pre-training, post-training) × 2 (Task: binary choice, interval estimation) design.

2.3 | Procedure and materials

Defence Research and Development Canada Human Research Ethics Committee approved the study before its commencement. Participants were permitted to complete the training and experiments either during or outside of working hours. They were informed that their participation was voluntary and that there was no remuneration, but that if they successfully completed the CPA course, they would receive a certificate of completion (which was, indeed, the case). Pre- and post-training surveys were administered via anonymous Qualtrics links distributed by the researchers and post-training was only administered upon the completion of the CPA course. The mean number of days between pre-training and training was 7.96 ($SD = 2.57$; median = 7) and the mean days between training and post-training was 2.20 ($SD = 2.78$;

TABLE 1

Phase	Major components
Pre-training	1. 20 binary choice questions and 20 interval estimation questions 2. Order of question type (binary vs. interval) randomized
Training: CPA course	***Module 1*** 1. Overview and objectives of course 2. Benchmark test (10 binary +10 interval estimation questions) ***Module 2*** 1. Definitions of calibration and overconfidence covered 2. Summary of research on overconfidence 3. The effects of the course on calibration skill in previous cohorts 4. Outcome and performance feedback on benchmark test ***Module 3*** 1. Taught that consistent, unambiguous, and immediate feedback is vital for improving calibration skill 2. Introduced to the equivalent bet test for improving skill at subjectively assessing probability 3. Participants practice equivalent bet test 4. Complete a second calibration test (10 binary +10 interval estimation questions) 5. Outcome and performance feedback on second calibration test ***Module 4*** 1. Participants are told that the majority of people applying the equivalent bet test to interval estimation questions choose Game B and need to widen their ranges but fail to do so sufficiently 2. Participants are instructed to start with extremely wide ranges so that they narrow their estimations as a result of choosing Game A 3. Introduction to Klein's premortem 4. Review of the updated calibration process: 5. A third calibration test, this time comprising 20 binary choice +20 interval estimation questions 6. Outcome and performance feedback on this third calibration test ***Module 5*** 1. Reminder about wide intervals for interval estimation questions 2. Shown examples of interval ranges provided by calibrated vs. miscalibrated individuals 3. Told if their expected performance was not within 2 points of their actual performance on the 20 binary questions, they were not yet calibrated 4. Told if their performance on the interval task was not at least 14, they were likely overconfident as most individuals who are calibrated typically score between 17 and 19 on the interval questions 5. Review of the updated calibration process (including do's and don'ts) 6. Completed a series of three calibration tests, each comprising 20 binary choice and 20 interval estimation questions each (outcome and performance feedback provided after each test) ***Module 6*** 1. Review of the calibration process 2. Misconceptions about applying the calibration process to real-world estimates refuted 3. Encouragement of the tracking and scoring of real-world estimates to remain calibrated
Post-training	1. 20 binary choice questions and 20 interval estimation questions 2. Order of question type (binary vs. interval) randomized 3. Rating of 71 unique urban agglomerations (UAs) encountered during both pre-training and post-training phases

median = 1). All but one participant completed the pre-training survey at least 5 days before undergoing calibration training and 87% of participants completed the post-training survey within 5 days of their self-reported CPA course completion.[3] During the completion of the surveys, participants were unable to view or modify responses entered on previous screens. Table 1 presents a summary of the procedure including each major component of the training modules. Further description of the CPA course training is in Data S1 available at osf.io/yqxpw/.

2.3.1 | Pre-training

Participants were informed that they would complete two tests designed to measure their calibration before and after receiving training. After providing consent, they responded to sets of 20 binary choice and 20 interval estimation questions (set order was randomized). The question formats reflected those in the CPA course (i.e., the training phase) and questions regarded the populations of geographic regions (a common topic from previous work, e.g., Juslin et al., 1999; Klayman et al., 1999; Subbotin, 1996; Teigen & Jørgensen, 2005)—specifically, urban agglomerations (UA).[4] The full list of questions is available at osf.io/yqxpw/.

In the binary choice task, participants responded to true/false statements about the relative populations of two UAs—for example, "New York-Newark, USA had a larger population than Tokyo, Japan in 2020." For each question, they were also asked to indicate "How confident are you that your answer is correct?" using a slider ranging from 50 ("Not confident at all") to 100 ("Absolutely confident").[5] UAs in each pair were drawn from a list of the most populous UAs in 2020 from the United Nations Department of Economic and Social Affairs (UN, 2018a). In the interval estimation task, participants provided

lower- and upper-bound estimates for the population of a given UA with 90% confidence—for example, "What was the population of Toronto, Canada in 2020?"

Pre-training and post-training questions were matched for difficulty (binary choice task) and UA familiarity (interval estimation task) based on an earlier unpublished experiment (Martin & Mandel, 2021). Among participants in the prior experiment, the mean accuracies of two binary question sets were 69.15% (SD = 7.56) and 68.60% (SD = 8.38) with no statistical difference [$t(37.60) = 0.22$, $p = .829$]. Likewise, among participants in the same prior experiment, mean familiarities of the two sets of UAs were 2.45 (SD = 0.71) and 2.46 (SD = 0.75) with no statistical difference [$t(37.85) = 0.04$, $p = .971$]. The procedures for equating difficulty and familiarity across pre/post question sets are available at osf.io/yqxpw/.

After completing pre-training, participants answered demographic questions (i.e., sex, age, education and years in the intelligence community) to characterize the sample before being debriefed.

2.3.2 | Training

The CPA course comprises six video modules and the questions of the course used a wide variety of general knowledge questions (whereas pre-training and post-training materials focused on the populations of major UAs).[6] At the beginning of the training, participants were presented with the course objectives and completed a benchmark test (10 binary and 10 interval estimation questions). They also learned about calibration and overconfidence research, and the effects of the course on the calibration skill in previous cohorts. After that, they received both outcome and performance feedback on their initial benchmark test at the question level. Throughout the course, outcome and performance feedback on test performance was in the form of general progress reports tracking calibration accuracy (actual compared to goal performance) across tests (6 in total) *and* question-level feedback for binary choice and interval estimation tasks wherein they were shown the question, the correct response, their response, and their response accuracy.

After the initial stages of training, participants were taught key strategies to improve their calibration and completed two calibration tests with both outcome and performance feedback. One strategy they learned about was the equivalent bet test where participants are asked to choose between two hypothetical games (Hubbard & Seiersen, 2016). In Game A, they win $1000 if their judgment is correct. In Game B, they spin a wheel with a chance to win $1000 equal to their stated confidence. If they prefer Game A, they increase their confidence level, whereas if they prefer Game B, they decrease their confidence level. The participant replays the equivalent bet until they are indifferent between Games A and B, indicating their true degree of uncertainty. The same process is applied to interval estimation questions by treating the lower- and upper-bound estimates as separate binary choice questions. Participants adjust the relevant bound while confidence remains fixed at 95%.

After learning the equivalent bet test, participants were introduced to Klein's premortem, which was described as a "prospective hindsight approach" to improving calibration. The premortem is applied in four steps: (1) participants make their initial estimate, (2) assume their answer is wrong, (3) explain why their answer is wrong, and (4) update their estimate accordingly. For example, a participant might estimate the population of Toronto, assume that their estimate is wrong, identify plausible reasons why (e.g., lack of familiarity with Toronto), and then widen their range.

Near the end of training, participants completed three calibration tests in series, receiving feedback after each one and were now familiar with each of the steps in the calibration training process. At the end of the training, participants reviewed a final calibration process: (1) choose initial estimates starting with extremely wide ranges for interval questions, (2) apply the equivalent bet test, (3) apply Klein's premortem, and (4) apply recommended confidence and interval adjustments. Participants were reminded to apply each step of the calibration process and to review previous modules if their performance did not improve over the next sequence of tests. They were also instructed to avoid backsliding if they initially performed well.

2.3.3 | Post-training

The post-training followed an identical procedure to the pre-training with one exception. After completing the judgment tasks, participants were asked to indicate their familiarity with the 71 unique UAs presented in the experimental task on a scale from 1 ("I have never heard of it prior to this study") to 5 ("I am very knowledgeable about it"). UAs were presented in random order per participant.

2.3.4 | Metrics

The current primary dependent measures are miscalibration and bias. In the binary choice task, miscalibration across the set of items is the absolute value of the difference between the participant's mean confidence and the proportion of correct responses the participant had (i.e., the participant's accuracy rate). Bias is defined as the arithmetic difference between the participant's mean confidence and the proportion of correct responses the participant had. In the interval estimation task, miscalibration is defined as the absolute value of the difference between .90 (i.e., 90% confidence) and the proportion of correct responses (i.e., when the true value falls in the estimated interval), whereas bias is the arithmetic difference between .90 and the proportion of correct responses. Unsurprisingly, the correlation between accuracy and calibration is strong for both the binary task ($r[68] = .62$, $p < .001$) and the interval estimation task ($r[68] = -.98$, $p < .001$) and likewise for the correlation between accuracy and bias on both the binary ($r[68] = -.70$, $p < .001$) and interval estimation tasks ($r[68] = 1.00$, $p < .001$).

3 | RESULTS

Analyses were not preregistered but are available along with the data at osf.io/yqxpw/ (conducted in R). As noted in Footnote 2, one participant failed to report: (i) the date of their CPA training, (ii) whether

they had been familiar with such training, and (iii) their familiarity ratings of the 71 unique UAs at the end of the post-training phase (hence, could not be included in analyses with familiarity). Another participant consistently reported values in the interval task well beyond 10 billion (10,000,000,000), which would seem misguided given the current global estimated population. Despite the atypical responses of the aforementioned participants, results are qualitatively the same with and without these individuals. The current results include these individuals where possible in the main text (analyses without them can be found in Data S1 at osf.io/yqxpw/). Effect sizes from ANOVAs are generalized eta squared, η^2_G, which is more comparable across within- and between-participant designs and are expected to be smaller than partial eta squared in repeated-measures designs with multiple factors (Bakeman, 2005; Olejnik & Algina, 2003). The current sample size of $n = 70$ is sufficient to detect a Cohen's f of .14 (small-medium sized effect) with .80 power based on computations using G*Power 3.1 and ANOVA_exact (Lakens & Caldwell, 2021).

3.1 | Preliminary analyses

The CPA course is designed to improve calibration (and bias). However, one would not expect it to improve accuracy on binary answers to general knowledge questions as in the present binary choice task. In line with this expectation, the mean accuracy in the pre-training version of the binary choice task ($M = .80$, $SD = .12$) was not significantly lower than the mean accuracy in the post-training version of the same task ($M = .76$, $SD = .11$). If anything, accuracy was numerically higher before training, though not significantly ($t[69] = 1.98$, $p = .052$, $d = 0.24$).

3.1.1 | Familiarity

We examined the relation between participants' familiarity ratings with UAs presented in the post-training phase and their post-training accuracy within each task. Provided that participants rated the familiarity of the 71 UAs after the post-training task, analyses examining the influence of familiarity with accuracy are only appropriate for post-training data to rule out the possibility of history effects for assessments of UAs presented during pre-training. We computed the correlation coefficient between post-training UA familiarity and post-training accuracy within each task (for the binary choice test wherein two UAs were presented simultaneously, the mean UA familiarity was used) for each participant. The mean correlation between familiarity and accuracy was .08, 95% CI [.03, .13] in the binary choice task, and − .14, 95% CI [−.19, −.09] in the interval estimation task (20 participants were excluded due to 0% or 100% accuracy on the interval task). The 95% CIs were bias-corrected accelerated bootstrap confidence intervals using 10,000 samples and they indicate a small positive relation between familiarity and accuracy for the binary choice task, but they suggest a small negative relation between familiarity and accuracy for the interval task. One possible explanation for the latter finding is that as familiarity increases, interval width may decrease. Hence, if participants are giving more narrow intervals for UA regions they are familiar with, they may be incorrect and the correct population could be less likely contained by a narrower interval. Providing potential support for this idea, the mean correlation between post-training familiarity and post-training interval width was negative but with a small coefficient of −.11, 95% CI [−.18, −.04].

3.1.2 | Demographics

Figure 1 presents the correlations among the demographic variables (i.e., sex, age, education, and years in the intelligence community) and miscalibration and bias separately for the two tasks. The only significant correlation between a demographic variable and a dependent measure was between age and miscalibration and for the interval estimation task in particular. Older participants were better calibrated (i.e., had lower miscalibration), but note that the effect is small and the significance is uncorrected for family-wise error (would be not significant with a Bonferroni correction; Dunn, 1961).

3.2 | Miscalibration

We examined the influences of training and task on participants' miscalibration using repeated measures analysis of variance (ANOVA). There were main effects of training ($F[1, 69] = 59.48$, $p < .001$, $\eta^2_G = .13$) and task ($F[1, 69] = 73.61$, $p < .001$, $\eta^2_G = .30$). Miscalibration was greater before training ($M = .33$, $SD = .12$) than after training ($M = .19$, $SD = .08$) and greater in the interval estimation task ($M = .37$, $SD = .04$) than in the binary choice task ($M = .14$, $SD = .01$). The main effects were qualified by a significant interaction ($F[1, 69] = 75.56$, $p < .001$, $\eta^2_G = .17$). Assessing the simple effect of training within task, we found that training significantly reduced miscalibration in the interval estimation task ($t[69] = 8.80$, $p < .001$, $d = 1.05$), but training did not significantly affect miscalibration in the binary choice task. In fact, in the latter task, miscalibration was worse after training, although the decrement was not statistically significant, ($t[69] = 1.80$, $p = .076$, $d = 0.22$). Figure 2 presents the effect of training on calibration as a function of task.

3.3 | Bias

We similarly examined the influence of training and task type on participants' bias using repeated measures ANOVA. There were main effects of both training ($F[1, 69] = 113.44$, $p < .001$, $\eta^2_G = .20$) and task ($F[1, 69] = 355.83$, $p < .001$, $\eta^2_G = .54$). Bias was greater before training ($M = .21$, $SD = .11$) than after training ($M = .01$, $SD = .10$) and greater in the interval estimation task ($M = .34$, $SD = .01$) than in the binary choice task ($M = -.12$, $SD = .01$). These

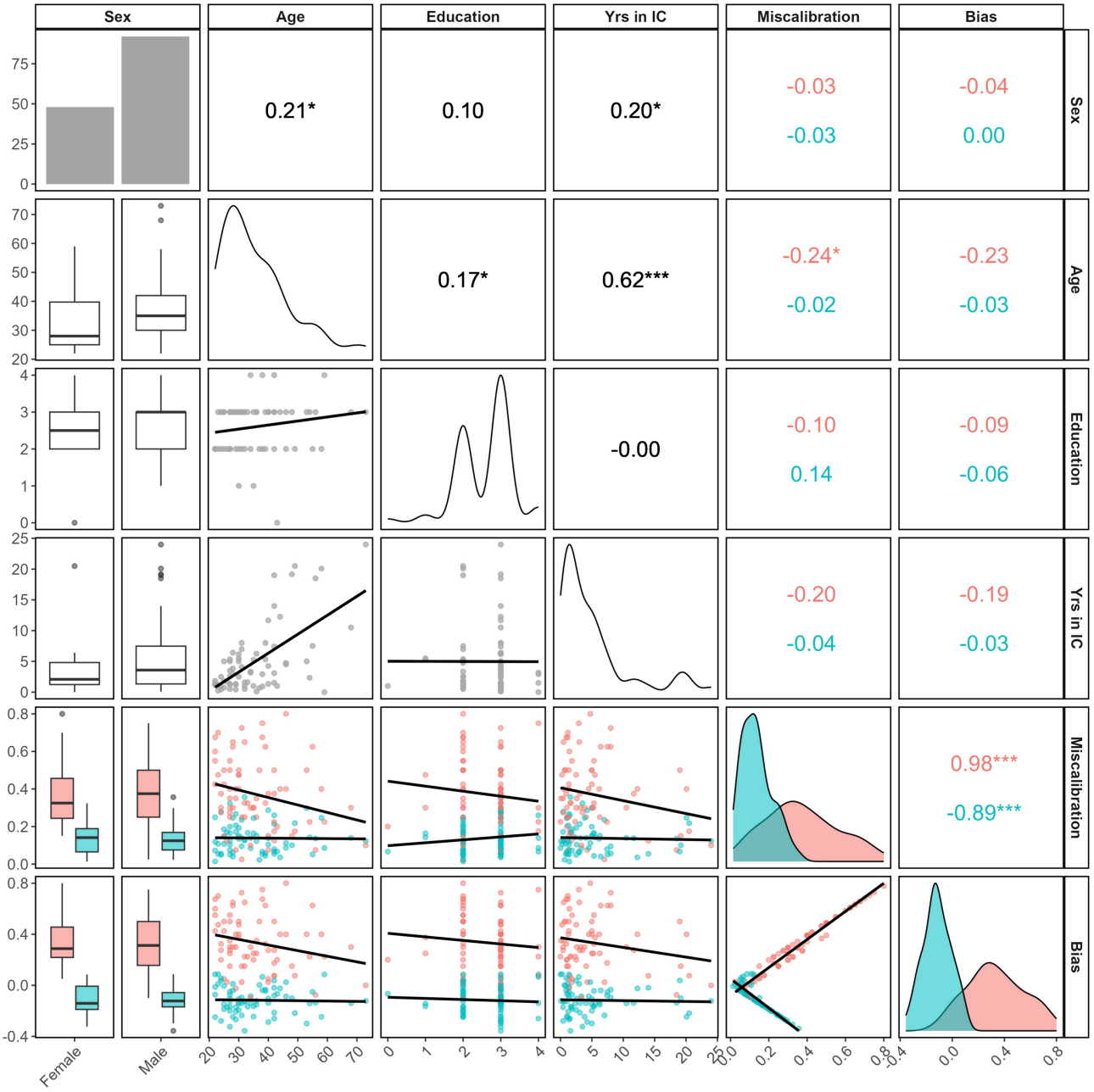

FIGURE 1 Bivariate analyses of key demographics and key dependent variables by task. Bivariate analyses of main variables by task (Pink = interval estimation task; Blue = binary choice task) where relevant. The upper-right panels present the Pearson correlations (Point biserial when involving the demographic variable of sex) and significance (*p < .05, **p < .01, ***p < .001). The diagonal displays (from top to bottom) the proportion of females and males (respectively), the distribution of age, education level, and years in the intelligence community (demographics not colored by task as all individuals completed both tasks), and the distribution of calibration and bias (colored by task). The left column presents box plots of the demographic variables by sex and box plots of the dependent measures as a function of sex and task. The lower-left panels (excluding the first column) present the bivariate distributions and slopes.

main effects were qualified by a significant interaction (F[1, 69] = 60.60, p < .001, η^2_G=.13). When assessing the simple effect of training within task, we found that training significantly increased bias in the binary choice task (t[69] = 3.11, p = .003, d = 0.37) and significantly reduced bias in the interval estimation task (t[69] = 9.83, p < .001, d = 1.17). Figure 2 presents the effect of training on bias as a function of task.

3.4 | Training generalizability

Finally, we examined the generalizability of calibration training across tasks. First, we computed the standardized difference between pre-training and post-training within each task separately for miscalibration and bias. Then, we tested whether these standardized differences correlated across task. These correlations were not statistically

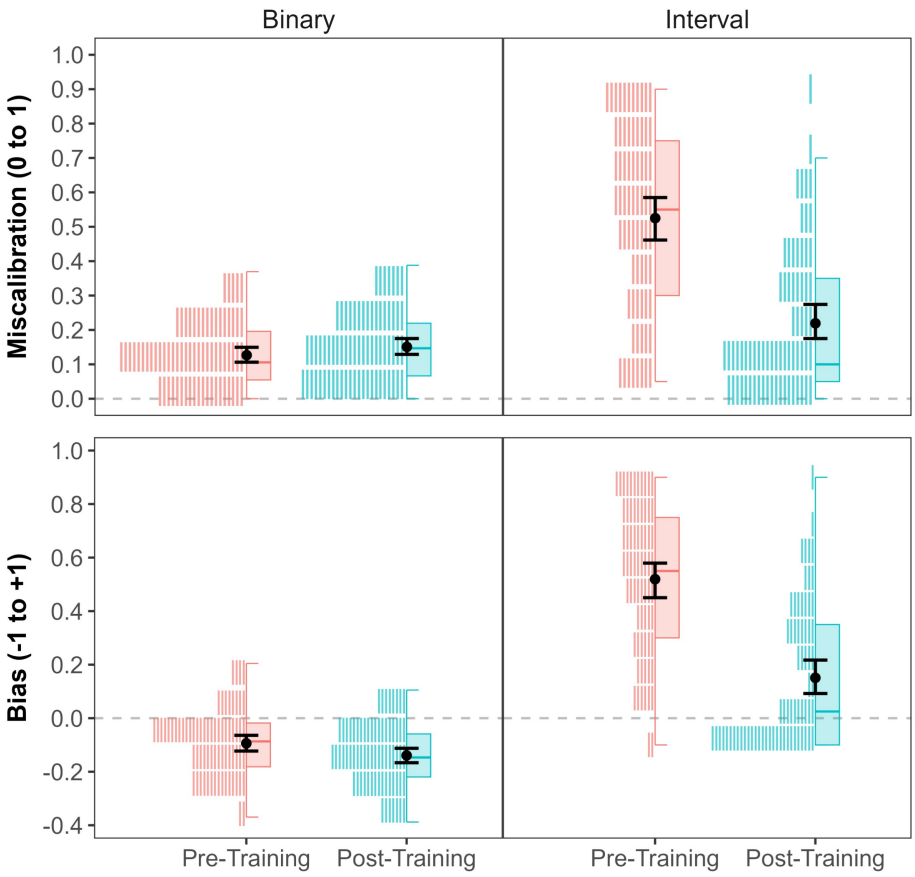

FIGURE 2 Effect of training (pre-training vs. post-training) on each task (binary vs. interval) on both calibration and bias. Histograms are sample data with each stroke representing one participant and where bin width is .10; Box plots are sample data; Group-level means are depicted by the black points and error bars are bootstrap bias-corrected accelerated 95% confidence intervals using 10,000 samples. Perfect calibration is represented by the horizontal dashed gray line.

significant: for miscalibration, $r(68) = -.06$, $p = .638$; for bias, $r(68) = -.10$, $p = .411$.

4 | GENERAL DISCUSSION

Calibration is a vital skill component of judgment accuracy (Lichtenstein & Fischhoff, 1980; Mandel & Barnes, 2014). Commercial calibration training offers one possible route for organizations that seek to develop the calibration skill of their employees and tamp down on judgment biases (e.g., overconfidence and underconfidence) that take the form of miscalibration. Although the prospect of commercially available solutions to improve judgment quality may be enticing, such training seldom receives rigorous intervention testing by independent researchers.

Presently, we examined the influence of commercial calibration training offered by HDR in the form of its CPA course on the degree of calibration and bias of intelligence analysts working across Canadian government departments. Generally, the CPA course significantly reduced miscalibration and bias. However, the effect of training on metacognitive performance depended on the task. Calibration performance increased after training in the case of the interval estimation task, but it decreased after training in the binary choice task such that analysts became more biased (i.e., underconfident). Training improved calibration and reduced bias only on the task for which analysts were initially overconfident (the interval estimation task) while training exacerbated the extant bias on the task for which analysts were initially underconfident (the binary choice task). We did not expect one task to yield overconfidence and the other to yield underconfidence. Instead, the tasks were selected because they reflected the types of judgments that were featured in the CPA course. It was fortuitous that the two tasks yielded opposing biases, however, for the effect would not otherwise have been detected. Had we only administered the interval estimation task, training would have appeared to be highly effective at improving calibration. Conversely, had we only administered the binary choice task, training would appear detrimental to good judgment.

4.1 | Confidence reduction

With both tasks yielding opposing biases, the findings indicate that the principal effect of training is to shift bias toward becoming less confident. After the present training, individuals initially overconfident become better calibrated after becoming less confident while individuals who are initially underconfident are at risk of becoming more underconfident. One might infer from these results that an initially perfectly calibrated judge will become miscalibrated following training provided an expected shift towards underconfidence. These findings bear not only on the efficacy of the CPA course, but raise a fundamental question about all calibration training: does calibration training reduce miscalibration by improving metacognitive monitoring of

confidence in one's judgments, or does it *appear* to reduce miscalibration because of a simple shift in bias during a context prone to initial overconfidence? This question is of both practical and theoretical importance.

If the effect of calibration training is driven by bias shift, then it is vital to know what, if any, bias is inherent in the judge or what bias might be encouraged by a given task. A bias-shifting intervention may still be beneficial in terms of accuracy. However, such benefit will be contingent on conditions remaining the same. Should the judge or task (or the judge-task interaction) prompt an opposing bias or even an attenuation of the same directional bias, then the debiasing intervention might harm rather than improve judgment quality. As Chang et al. (2018) and Mandel and Irwin (2023) have noted, intelligence community methods aimed at debiasing analysts' reasoning processes and judgments ought to consider the fact that most cognitive biases are bipolar (e.g., overshooting as in overconfidence or undershooting as in underconfidence). Yet, the intelligence community seldom takes bias direction into account and, quite often, the direction of bias is presumed (e.g., "overconfidence needs to be reduced"; Chang et al., 2018; Mandel & Irwin, 2023).

The present findings raise important questions about the causal bases of calibration training effects. On tasks that yield overconfidence, a reduction in overconfidence can signal either a genuine improvement in calibration skill (i.e., a better ability to metacognitively monitor the appropriate level of confidence to assign to a judgment or choice) or a simple shift in bias. As noted, a simple shift in bias is of unreliable benefit because any changes that prompt a change in the direction of bias or that eliminate the bias would lead to impairment of judgment quality. A genuine improvement in the metacognitive ability to monitor appropriate confidence levels in judgment should result in confidence reduction where overconfidence is manifested, and confidence increases where underconfidence is manifested.

Moreover, if metacognitive monitoring were strengthened from training, one might also expect that changes in performance on the two tasks would be correlated. However, changes in performance across tasks were uncorrelated in this study, consistent with previous work showing that calibration skill does not easily generalize across judgment tasks (Keren, 1985 as cited in Keren, 1987; Solomon et al., 1985). That the effect of CPA training did not appear to generalize between the tasks we used, even though both tasks drew on similar knowledge (i.e., UA populations), supports a bias-shifting mechanism rather than genuinely improved metacognitive monitoring of confidence in judgment. Future research could profitably examine this issue, for instance, by systematically varying task difficulty in the training and test phases of an interventions-testing study. If one were trained on difficult items and learned mainly that they should be less confident, then when presented with easy items, they may be expected to exhibit underconfidence. Another avenue for future work would be to assess the effect of calibration training when participants are allowed to select their confidence level as some research indicates that individuals may be better calibrated when they can choose the confidence levels for their judgments (Soll & Klayman, 2004; Teigen & Jørgensen, 2005).

4.2 | Training content

It is also important to consider the potential partiality within the training materials themselves. The content of the current CPA course, in particular, may have caused a shift in bias towards confidence reduction considering much more attention is paid to reducing overconfidence than to reducing underconfidence. More direct references to overconfidence rather than underconfidence were made and training examples focused on overconfidence. Moreover, a few times throughout training, trainees are told that "if they are like most individuals, they will exhibit overconfidence, especially at first" and, the tables used throughout the course to explain degrees of calibration show two levels of overconfidence—slightly and extremely, while only showing one level of under confidence (slightly). Finally, the strategies encouraged throughout the training arguably promote underconfidence. For example, Klein's premortem encourages trainees to think of why their initial estimate is wrong which might encourage confidence reduction even if individuals are already calibrated or underconfident.

We have not reviewed alternatives such as Good Judgment Incorporated's (2022) Superforecasting Fundamentals course to assess whether there is a similar bias toward reducing confidence. However, we recommend that organizations seeking to improve calibration skills review products with this question in mind and also consider whether they have reason to care more about minimizing one type of miscalibration over another. If, for instance, they know that their experts are overconfident, then training that encourages less confidence may still have the desired effect.

It is possible that training developed to emphasize both the perils of over- and underconfidence could improve the efficacy of courses designed to improve metacognitive calibration. The current study also begs the question of whether calibration training is possible or if it is merely lowering confidence. Future research is needed to determine what cues intelligence analysts are using to arrive at their judgments of accuracy (their confidence) given that it is unclear how the extant literature might translate to real-world judgments made by such analysts.

4.3 | Task-expert interactions

Note that the present sample of intelligence analysts may be more underconfident than samples of non-analysts or, for that matter, from other groups of experts. For instance, Mandel and Barnes (2014, 2018) found that strategic intelligence forecasts made by Canadian intelligence analysts were underconfident. Forecasts made by intelligence analysts often bear much similarity to the binary choice task in the present study, which also demonstrated underconfidence. In contrast, the interval estimation task has less of a deep-structure resemblance to the typical judgments by analysts which might explain why analysts were overconfident. The binary choice task might not prompt underconfidence in nonexpert samples, suggesting that careful attention to potential task-expertise interactions is needed. Future work could examine task-expert interactions by comparing the effects of calibration training in expert and nonexpert samples across different tasks.

Finally, the tasks implemented in the current work are likely more straightforward than those faced by professional analysts "in the wild." While this may cause some concerns about external validity, featuring generic and simple tasks seemed necessary in the current work provided that the commercial calibration training tasks are generic and simple, and we aimed to test the claims of improved calibration made by commercial training courses. Further, the generic topics covered in the training seemed reasonable provided wide-ranging expertise of professional analysts across domains. While it is important to be aware of the potential risks in overextending the current results to the "real world," we think that our tasks being more straightforward than those tasks faced by professional analysts would likely only serve as stronger evidence that the commercial training courses are unlikely to improve calibration, given that we did not see consistently increased calibration in simple tasks based on those featured in the training materials.

5 | CONCLUSION

Our examination of the effect of commercial calibration training on the calibration of professional intelligence analysts across two judgment tasks found that there was an overall reduction in miscalibration and bias after training. However, in a task wherein experts tended to already exhibit underconfidence before training, underconfidence increased rather than decreased after training. Taken together, the findings indicate that the training led to a bias shift toward lower confidence rather than an improved ability to metacognitively monitor confidence in judgment.

AUTHOR CONTRIBUTIONS

Megan O. Kelly: Conceptualization; formal analysis; data curation; visualization; writing – original draft; writing – review and editing. **David R. Mandel:** Conceptualization; supervision; writing – review and editing; funding acquisition; data curation; formal analysis.

ACKNOWLEDGMENTS

We wish to thank Nicole Herz and Daniel Irwin for their assistance with this research.

FUNDING INFORMATION

This research was supported by the Accelerate Command, Control, and Intelligence Project Activity #AC2I-019 and Canadian Safety and Security Program project #2018-TI-2394 under the direction of the second author.

CONFLICT OF INTEREST STATEMENT

Neither researcher has professional or financial connections to HDR; the course licenses used in this study were procured at regular cost by the Government of Canada. HDR's CPA course was chosen for this study because it has previously been used by one of the intelligence organizations from which we recruited participants and that organization had requested an investigation of the CPA course's efficacy. HDR did not provide any support for this study aside from reviewing our description of the CPA course for accuracy and no member of the team was asked to review a draft of this article prior to journal submission. None of the raw data from this study was shared with HDR (nor was it requested).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/yqxpw/.

ORCID

Megan O. Kelly https://orcid.org/0000-0001-6497-4400

David R. Mandel https://orcid.org/0000-0003-1036-2286

ENDNOTES

[1] Note that Western intelligence organizations typically communicate probabilistic judgments verbally (e.g., "*x* is *likely* to occur"; "We assess with *low* confidence…"), despite the strong empirical case for using numeric quantifiers (see Dhami & Mandel, 2021; Friedman, 2019; Irwin & Mandel, 2023; Mandel & Irwin, 2021).

[2] One participant did not answer this question, but the results of analyses with or without this participant are qualitatively the same and the participant was included in the analyses.

[3] Results are qualitatively the same with and without the 13% of participants who did not report completing training within 5 days of their self-reported CPA course completion (see Data S1 at https://osf.io/yqxpw/?view_only=634724b75ae642c0b189406d50e3e79d for details).

[4] The United Nations Department of Economic and Social Affairs (2018b) defines an urban agglomeration as "the population contained within the contours of a contiguous territory inhabited at urban density levels without regard to administrative boundaries." Participants were shown this definition at the start of both question sets.

[5] The default position of each confidence slider was 50. Participants who wanted to indicate the default position as their response still had to click the slider. Whereas HDR's CPA course elicited confidence in 10-point increments (except for 95% confidence), our sliders increased in one-point increments.

[6] See examples in Exhibit 5.1 in Hubbard, 2014, p. 96.

REFERENCES

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review*, *68*(1), 33–45. https://doi.org/10.1037/h0040274

Adams, P. A., & Adams, J. K. (1958). Training in confidence-judgments. *The American Journal of Psychology*, *71*, 747–751. https://doi.org/10.2307/1420334

Alpert, M., & Raiffa, H.. (1969/1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.022

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 495–515). MA7 Kluwer Academic Publishers. https://doi.org/10.1007/978-0-306-47630-3_22

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379–384. https://doi.org/10.3758/BF03192707

Ben-David, I., Graham, J. R., & Harvey, C. R. (2013). Managerial miscalibration. *The Quarterly Journal of Economics*, *128*(4), 1547–1584. https://doi.org/10.1093/qje/qjt023

Benjamin, D., Hey, S., MacPherson, A., Hachem, Y., Smith, K., Zhang, S., Wong, S., Dolter, S., Mandel, D. R., & Kimmelman, J. (2022). Principal investigators over-optimistically forecast scientific and operational clinical trial outcomes. *PLoS One*, *17*(2), e0262862. https://doi.org/10.1371/journal.pone.0262862

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. https://doi.org/10.1016/0169-2070(92)90066-I

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5), S2–S23. https://doi.org/10.1016/j.amjmed.2008.01.001

Biais, B., Hilton, D., Mazurier, K., & Pouget, S. (2005). Judgmental overconfidence, self-monitoring and trading performance in an experimental financial market. *Review of Economic Studies*, *72*, 287–312. https://doi.org/10.1111/j.1467-937X.2005.00333.x

Bier, V. M. (2004). Implications of the research on expert overconfidence and dependence. *Reliability Engineering & System Safety*, *85*(1–3), 321–329. https://doi.org/10.1016/j.ress.2004.03.020

Bolger, F., & Onkal-Atay, D. (2004). The effects of feedback on judgmental interval predictions. *International Journal of Forecasting*, *20*(1), 29–39. https://doi.org/10.1016/S0169-2070(03)00009-8

Brinkman, D. J., Tichelaar, J., van Agtmael, M. A., de Vries, T. P., & Richir, M. C. (2015). Self-reported confidence in prescribing skills correlates poorly with assessed competence in fourth-year medical students. *The Journal of Clinical Pharmacology*, *55*(7), 825–830. https://doi.org/10.1002/jcph.474

Bum, C. H., Choi, C., & Lee, K. (2018). Irrational beliefs and social adaptation of online sports gamblers according to addiction level: A comparative study. *Sustainability*, *10*(11), 4314. https://doi.org/10.3390/su10114314

Busby, L. P., Courtier, J. L., & Glastonbury, C. M. (2018). Bias in radiology: The how and why of misses and misinterpretations. *Radiographics*, *38*(1), 236–247. https://doi.org/10.1148/rg.2018170107

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, *11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Chang, W., Berdini, E., Mandel, D. R., & Tetlock, P. E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security*, *33*(3), 337–356. https://doi.org/10.1080/02684527.2017.1400230

Chang, W., & Tetlock, P. (2016). Rethinking the training of intelligence analysts. *Intelligence and National Security*, *31*(6), 903–920. https://doi.org/10.1080/02684527.2016.1147164

Charba, J. P., & Klein, W. H. (1980). Skill in precipitation forecasting in the National Weather Service. *Bulletin of the American Meteorological Society*, *61*(12), 1546–1555. https://doi.org/10.1175/1520-0477(1980)061<1546:SIPFIT>2.0.CO;2

Dhami, M. K., & Mandel, D. R. (2021). Words or numbers? Communicating probability in intelligence analysis. *American Psychologist*, *76*(3), 549–560. https://doi.org/10.1037/amp0000637

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science*, *10*(6), 753–757. https://doi.org/10.1177/1745691615598511

Doussau, A., Kane, P., Peppercorn, J., Feustel, A. C., Ganeshamoorthy, S., Kekre, N., Benjamin, D., & Kimmelman, J. (2023). The impact of feedback training on prediction of cancer clinical trial results. In *The impact of feedback training on prediction of cancer clinical trial results*. Advanced Online Publication. https://doi.org/10.1177/17407745231203375

Du, N., & Budescu, D. V. (2007). Does past volatility affect investors' price forecasts and confidence judgements? *International Journal of Forecasting*, *23*(3), 497–511. https://doi.org/10.1016/j.ijforecast.2007.03.003

Du, N., Shelton, S., & Whittington, R. (2012). Does supplementing outcome feedback with performance feedback improve probability judgments? *International Journal of Financial Research*, *3*(4), 19–32. https://doi.org/10.5430/ijfr.v3n4p19

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Emory, B., & Luo, T. (2022). Metacognitive training and online community college students' learning calibration and performance. *Community College Journal of Research and Practice*, *46*(4), 240–256. https://doi.org/10.1080/10668926.2020.1841042

Erceg, N., & Galić, Z. (2014). Overconfidence bias and conjunction fallacy in predicting outcomes of football matches. *Journal of Economic Psychology*, *42*, 52–62. https://doi.org/10.1016/j.joep.2013.12.003

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning*, *12*, 1–19. https://doi.org/10.1007/s11409-016-9158-6

Friedman, J. A. (2019). *War and chance: Assessing uncertainty in international politics*. Oxford University Press. https://doi.org/10.1093/oso/9780190938024.001.0001

Gentry, J. A. (2017). The intelligence of fear. *Intelligence and National Security*, *32*(1), 9–25. https://doi.org/10.1080/02684527.2016.1199348

Good Judgment Incorporated. (2022). Superforecasting fundamentals course. https://good-judgment.thinkific.com/courses/Superforecasting-Fundamentals

Gruetzemacher, R., Lee, K. B., & Paradice, D. (2023). Calibration training for improving probabilistic judgments using an interactive app. *Futures & Foresight Science*, *e177*.

Guiterrez, A. P., & Schraw, G. (2015). Effects of strategy training and incentives on students' performance, confidence, and calibration. *The Journal of Experimental Education*, *83*(3), 386–404. https://doi.org/10.1080/00220973.2014.907230

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of experimental psychology. Learning, Memory, and Cognition*, *11*, 719–731. https://doi.org/10.1037/0278-7393.11.1-4.719

Hoch, S. J., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 605–619. https://doi.org/10.1037/0278-7393.15.4.605

Hubbard, D. W. (2014). *How to measure anything: Finding the value of intangibles in business*. John Wiley & Sons. https://doi.org/10.1002/9781118983836

Hubbard, D. W., & Seiersen, R. (2016). *How to measure anything in cybersecurity risk*. John Wiley & Sons. https://doi.org/10.1002/9781119162315

Hubbard Decision Research. (2019). Calibrated Probability Assessments. https://hubbardresearch.com/shop/calibrated-probability-assessments/

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning*, *4*(2), 161–176. https://doi.org/10.1007/s11409-009-9042-8

Irwin, D., & Mandel, D. R. (2023). Communicating uncertainty in national security intelligence: Expert and nonexpert interpretations of and preferences for verbal and numeric formats. *Risk Analysis*, *43*(5), 943–957. https://doi.org/10.1111/risa.14009

Jaspan, O., Wysocka, A., Sanchez, C., & Schweitzer, A. D. (2022). Improving the relationship between confidence and competence: Implications for diagnostic radiology training from the psychology and medical literature. *Academic Radiology*, *29*(3), 428–438. https://doi.org/10.1016/j.acra.2020.12.006

Johnson, J. E., & Bruce, A. C. (2001). Calibration of subjective probability judgments in a naturalistic setting. *Organizational Behavior and Human*

Decision Processes, *85*(2), 265–290. https://doi.org/10.1006/obhd.2000.2949

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1038–1052. https://doi.org/10.1037/0278-7393.25.4.1038

Kelly, C. W., Peterson, C. R., Brown, R. V., & Barclay, S. (1975). *Decision theory research (technical program report 4)*. Decisions and Designs, Inc.

Keren, G. (1985). On the calibration of experts and lay people. Unpublished manuscript.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, *39*(1), 98–114. https://doi.org/10.1016/0749-5978(87)90047-1

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

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247. https://doi.org/10.1006/obhd.1999.2847

Klein, G. (2008). Performing a project premortem. *IEEE Engineering Management Review*, *36*(2), 103–104. https://doi.org/10.1109/EMR.2008.4534313

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–14. https://doi.org/10.1177/2515245920951503

Lawrence, M., Goodwin, P., O'Connor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518. https://doi.org/10.1016/j.ijforecast.2006.03.007

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*(2), 149–171. https://doi.org/10.1016/0030-5073(80)90052-5

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. 306–354). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.023

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology*, *6*(387), 1–12. https://doi.org/10.3389/fpsyg.2015.00387

Mandel, D. R. (2019). Can decision science improve intelligence analysis? In S. Coulthart, M. Landon-Murray, & D. van Puyvelde (Eds.), *Researching National Security Intelligence: Multidisciplinary approaches* (pp. 117–140). Georgetown University Press.

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(30), 10984–10989. https://doi.org/10.1073/pnas.1406138111

Mandel, D. R., & Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making*, *31*(1), 127–137. https://doi.org/10.1002/bdm.2055

Mandel, D. R., Barnes, A., & Richards, K. (2014). *A quantitative assessment of the quality of strategic intelligence forecasts* (Technical Report No. 2013-036). Defence Research and Development Canada.

Mandel, D. R., Collins, R. N., Risko, E. F., & Fugelsang, J. A. (2020). Effect of confidence interval construction on judgment accuracy. *Judgment and Decision making*, *15*(5), 783–797. https://doi.org/10.1017/S1930297500007920

Mandel, D. R., & Irwin, D. (2021). Facilitating sender-receiver agreement in communicated probabilities: Is it best to use words, numbers or both? *Judgment and Decision making*, *16*(2), 363–393. https://doi.org/10.1017/S1930297500008603

Mandel, D. R., & Irwin, D. (2023). /Beyond bias minimization: Improving intelligence with optimization and human augmentation. *International Journal of Intelligence and CounterIntelligence*, *1-17*, 649–665. https://doi.org/10.1080/08850607.2023.2253120

Martin, M., & Mandel, D. R. (2021). Effect of outcome and calibration feedback on the calibration of judgment. Unpublished data.

McKenzie, C. R., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, *107*(2), 179–191. https://doi.org/10.1016/j.obhdp.2008.02.007

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S., Moore, D., Atanasov, P., Swift, S., Murray, T., Stone, E., & Tetlock, P. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Meyer, A. N., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine*, *173*(21), 1952–1958. https://doi.org/10.1001/jamainternmed.2013.10081

Meyer, A. N., & Singh, H. (2017). Calibrating how doctors think and seek information to minimise errors in diagnosis. *BMJ Quality and Safety*, *26*(6), 436–438. https://doi.org/10.1136/bmjqs-2016-006071

Moore, D. A. (2011). *Managerial decision making*. Edward Elgar Publishing Limited. https://doi.org/10.4337/9781784713751

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552–3565. https://doi.org/10.1287/mnsc.2016.2525

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*, 489–500. https://doi.org/10.1080/01621459.1984.10478075

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, *1*(159), 159–179. https://doi.org/10.1007/s10409-006-9595-6

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research*, *95*(3), 131–142. https://doi.org/10.1080/00220670209596583

Niu, X., & Harvey, N. (2022). Outcome feedback reduces over-forecasting of inflation and overconfidence in forecasts. *Judgment and Decision making*, *17*(1), 124–163. https://doi.org/10.1017/S1930297500009050

O'Connor, M., & Lawrence, M. (1989). An examination of the accuracy of judgmental confidence intervals in time series forecasting. *Journal of Forecasting*, *8*, 141–155. https://doi.org/10.1002/for.3980080207

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Onkal, D., & Muradoglu, G. (1995). Effects of feedback on probabilistic forecasts of stock prices. *International Journal of Forecasting*, *11*, 307–319. https://doi.org/10.1016/0169-2070(94)00572-T

Phelps, R. H., Halpin, S. M., Johnson, E. M., & Moses, F. L. (1980). *Implementation of subjective probability estimates in army intelligence procedures: A critical review of research findings (research report 1242)*. US Army Research Institute for Behavioral and Social Sciences.

Rieber, S. (2004). Intelligence analysis and judgmental calibration. *International Journal of Intelligence and CounterIntelligence*, *17*(1), 97–112. https://doi.org/10.1080/08850600490273431

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, *33*, 7–17.

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology*, *33*(5), 918–929. https://doi.org/10.1002/acp.3556

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*(3), 271–283. https://doi.org/10.1016/0749-5978(88)90001-5

Soll, J. B., & Klayman, J. (2004). Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299–314. https://doi.org/10.1037/0278-7393.30.2.299

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

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes*, *83*(2), 282–309. https://doi.org/10.1006/obhd.2000.2910

Subbotin, V. (1996). Outcome feedback effects on under-and overconfident judgments (general knowledge tasks). *Organizational Behavior and Human Decision Processes*, *66*(3), 268–276. https://doi.org/10.1006/obhd.1996.0055

Teigen, K. H., & Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology*, *19*(4), 455–475. https://doi.org/10.1002/acp.1085

United Nations. (2018a). WUP2018-F12-Cities_Over_300K.xls. https://population.un.org/wup/Download/

United Nations. (2018b). How do we define "urban agglomeration". https://population.un.org/wup/General/FAQs.aspx#:~:text

Veinott, B., Klein, G. A., & Wiggins, S. (2010). Evaluating the effectiveness of the PreMortem technique on plan confidence. *Proceedings of the 7th International ISCRAM Conference*.

Zacharakis, A. L., & Shepherd, D. A. (2001). The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, *16*(4), 311–332. https://doi.org/10.1016/S0883-9026(99)00052-X

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kelly, M. O., & Mandel, D. R. (2024). The effect of calibration training on the calibration of intelligence analysts' judgments. *Applied Cognitive Psychology*, *38*(5), e4236. https://doi.org/10.1002/acp.4236