Transferability of calibration training between knowledge domains

Christopher Babadimas, Christopher Boras, Nicholas Rendoulis, Matthew Welsh & Steve Begg
([Christopher.babadimas; Christopher.boras; Nicholas.Rendoulis] @student.adelaide.edu.au; [matthew.welsh; steve.begg] @adelaide.edu.au)
Australian School of Petroleum, University of Adelaide, Adelaide, SA 5005, Australia

Abstract

Many industry professionals are poorly calibrated, overestimating their ability to make accurate forecasts. Previous research has demonstrated that an individual's calibration in a specific domain can be improved through calibration training in that domain; however devising a training program for each specific domain within a field is laborious. A more efficient method would be if individuals from different disciplines could undertake the same general training and transfer the skills learnt to their respective, specific domains. This study investigated whether calibration training in a general domain was transferable to the specific domain of petroleum engineering. The results showed that, whilst the feedback training was effective within the general domain, there was only limited transfer to the specific domain. This is argued to be due to recognition failure, where the participants failed to recognise that the skill learnt through training in the general domain could be transferred to the specific domain.

Keywords: calibration; overconfidence; training; skill transfer.

Introduction

In technical disciplines and industries, individuals are required to provide range estimates, such as 80 percent confidence intervals, for uncertain parameters used in modeling and decision making (see, e.g., Capen, 1976). The accuracy of the individual's estimates can greatly influence decisions, with significant impacts on company bottom lines (see, e.g., Welsh, Begg & Bratvold, 2007). Calibration is the measure of how well individuals' estimates match real world outcomes (Lichtenstein, Fischhoff & Phillips, 1982). For example, if a weather forecaster makes multiple predictions of an 80% chance of rain, and on 80% of those occasions it does rain, they are well calibrated (for 80%), meaning they have a higher likelihood of providing more accurate estimates, which lead to more informed decisions.

Poor calibration in range estimation tasks can result from cognitive biases (e.g., biases from the anchoring and availability heuristics; Tversky & Kahneman, 1974) and is described as overprecision - one type of overconfidence bias (Moore & Healy, 2008). Overprecision describes the observation that individuals provide overly narrow ranges that do not represent their true degree of knowledge (Moore, 2008). The tendency for individuals to be over-precise in estimation has been demonstrated repeatedly (e.g., Soll & Klayman, 2004; Lichtenstein *et al.,* 1982) and seems to affect experts similarly to novices (McKenzie, Lierch & Yaniv, 2008). (NB, many studies use the term 'overconfidence' rather than overprecision and, in order to stay consistent with past literature, this will be done hereafter.)

Calibration Training

Past research has shown calibration can be improved through debiasing techniques, the most effective being domain-specific performance feedback training, wherein a subject receives timely feedback on the accuracy of their estimates within a particular area of knowledge (e.g., a field like petroleum engineering or meteorology; see, e.g.: Adams & Adams, 1958; Fischhoff, 1981) or learns this over an extended period in an amenable environment (see, e.g., Tetlock & Gardner, 2016). Whilst domain-specific training may be effective, devising training programs for numerous specific domains within a wider field or industry is laborious. For example, oil industry personnel include engineers and geoscientists across various specialties and a generalised training program, with calibration training learnt in a general domain and learnings transferred to specific domains, would be a more efficient method of improving calibration for a company employing these people.

Despite this previous research on domain-specific performance feedback training, it has seldom extended to the idea of creating generalised performance feedback training. Adams and Adams (1961) showed that training a subject's calibration in a series of tasks lead to an improvement in calibration in a separate task, an idea termed "generalisation"; although the degree of improvement in the untrained task was lower than in the trained task. Lichtenstein & Fischhoff (1980) showed that calibration training in a base task improved calibration in other, similar, tasks but not on dissimilar tasks, which was attributed to the subjects' inability to spontaneously relate the new task to the base task.

Similarly, Bornstein & Zickafoose (1999) demonstrated that individuals' confidence and accuracy were stable across domains of general knowledge and eyewitness memory, and that training using general knowledge questions reduced overconfidence in eyewitness memory. Improvements in calibration and resolution, however, were not observed, implying no improvement in accuracy. Thus, the above studies suggest that generalised training could be effective but, given inconsistent results and the fact that this was not their primary focus, the question of whether generalised training transfers to specific domains remains open.

An argument supporting the plausibility of the generalizable calibration training is analogical transfer,

where studies have shown transfer of knowledge (although not calibration training) across domains. Analogical transfer involves the use of a familiar problem to solve a novel problem of the same structure (Reeves, 1994). By identifying similarities in the structure of base and target problems, a subject can transfer the principles of the base problem to solve the target problem (Glick & Holyoak, 1983). Analogical transfer is argued to be the main method used to solve novel problems in all domains (Rumelhart, 1989). Given this, if the process of improving calibration training can be stripped down to its base structure, analogical transfer may facilitate transfer of calibration training across domains.

Whilst the structural similarities of the base and analogue problems are essential to facilitate transfer, they do not guarantee *recognition* of the relationship, which would prevent spontaneous transfer from one problem to the next without instructions or help (Day & Goldstone, 2012). Recognition failure is argued to occur largely as a result of dissimilar surface elements in the respective problems (Day, 2012) – for example, questions drawn from different domains - but may be improved by providing multiple base problems, as this will allow the subject to derive a more general analogy (Glick & Holyoak, 1983). Recognition failure may provide an explanation for the limitations in generalisation seen in Lichtenstein & Fischoff (1980), and Bornstein & Zickafoose (1999). Conversely, Adams and Adams (1958) achieved moderate generalization - using training in multiple, different tasks.

Given the paucity of research into the generalization of calibration training and the apparent absence of research connecting transferability of calibration training to analogical transfer, this paper has the opportunity to fill a distinct research gap. The research is further warranted by the paper's focus on the practical issue of how best to provide training. That is, seeing whether analogical transfer facilitates calibration training transferring to a new domain is both practically and theoretically interesting.

Aims

Given the unclear evidence in the literature, this paper's primary aim is to see whether generalised training in calibration can be developed to enable transfer of improved calibration to problems in a different, specific knowledge domain – specifically, petroleum engineering. This leads to two main hypotheses, as shown below:

H1: Calibration training will improve calibration within the domain in which the training is given.
H2: Improvements in calibration training will transfer to a new, specific domain.

It is important to highlight that the term "general domain" is used to describe a domain, unrelated to the specific domain, in which training will be given. The term generalized training thus refers to training applied in the general domain. In the context of a real world application it makes sense for the general domain selected to be general knowledge, as this domain is accessible to all, and is clearly separate from a subject's specific domain of expertise. A general domain in this context could, however, be any domain other than the participant's specific domain.

Methodology

Participants

Participants were 54 (15F and 39M) recent (n=7) and current (n=47) students of the Australian School of Petroleum, University of Adelaide, ranging in age from 18 to 35 (M=22, SD=3.0). Previous experience with calibration varied amongst the participants, with 31 participants having previously undertaken a course that taught calibration, and 15 who had not undertaken the course but who indicated (prior to the study) that they understood what calibration was and how it affects decision making. Participants entered a draw (1 in 6 chance) to win one of several $200 gift cards.

Materials

Testing materials consisted of three questionnaires - two general knowledge, and one in the domain of petroleum engineering - and a feedback/training package (described below). In this scenario, petroleum engineering is the specific domain, and general knowledge the general domain. Petroleum engineering was chosen to be specific domain due to the higher level of expertise in this field (compared to the general populace) shared by all participants. This higher level of experience is expected to elevate their knowledge of this domain above the participant's understanding of more general knowledge; separating it from the general domain. In terms of knowledge transfer, the assumption is that participants may think differently about their area of specialty than general knowledge questions and, thus, that recognition failure across the two domains may be more likely.

The first general knowledge test – designated "Pre-Training" - contained 30 questions; however, the number of questions in the remaining tests (designated "Post-Training" and "Domain Specific") were reduced to 20 each following participant feedback. The tests consisted of questions that had definite numerical answers and were sufficiently difficulty for participants not to simply know the true answer. An example of a question used in the general knowledge domain (i.e., the Pre-Training and Post-Training questionnaires) was "How many countries does the Nile River cross over?" For comparison, an example question used in the Domain Specific questionnaire was "How many times greater is the Young's Modulus of a stiff sandstone compared to the Young's Modulus of coal?"

In all cases, participants were asked to provide a low and a high value such that they were 80% confident their range would contain the true value. (The initial page of each test provided information about how to answer the questions, including an example question.)

While the second test is designated "Post-Training",

feedback materials were provided only to the Experimental Group. This was a pdf document, consisting of: information about calibration and overconfidence, a calibration curve illustrating the subject's under/overconfidence, a histogram showing the subject's calibration score relative to other participants, and a graphical depiction of the subject's confidence intervals, plotted against the corresponding true answers. These figures were intended to help participants understand the degree of overconfidence they had shown in the Pre-Training test. Each figure was accompanied by a short explanation, and information on methods for improving calibration on the remaining tests – including recognition of their current calibration in order to prompt them to give wider ranges.

Procedure

After registering their interest, participants were provided with links to access the Pre-Training questionnaire online (on SurveyMonkey) with instructions to complete each question by providing 80% confidence interval estimates. Based on the results of the Pre-Training questionnaire, participants were divided into two groups with similar levels of calibration. Feedback training was then distributed to the Experimental Group via email, at most two weeks after completing the test, with instructions to read and understand the material completely before continuing to the general knowledge Post-Training questionnaire). To test whether participants understood the feedback, a four-question quiz was given on the material covered in the training package. Participants who scored less than 3 out of 4 (2 participants) were moved from the Experimental Group to the control group, as it was adjudged they had not read the material and hence not received the feedback (NB – while recognizing that removing the participants may have been a more appropriate, this choice was made in light of the already small sample). Links to the Post-Training and the Domain Specific questionnaires were then provided to participants straight after the feedback training was distributed.

Improvements in calibration due to the feedback training were measured by comparing the Experimental Group's Post-Training and Domain Specific questionnaires to baselines of the Experimental Group's Pre-Training questionnaire and the Control Group's Post-Training and Domain Specific questionnaires. Comparisons were made under the assumption that the tests were of equal hardness and both groups were equally well calibrated. This yields measures of both the effectiveness of the feedback training and the transferability of the training across domains.

Results

Descriptive Statistics

Table 1 shows the demographics for the experimental and control groups while Figure 1 shows the mean calibration achieved by each group under each condition (with 95% CIs - recalling that questions asked for 80% ranges meaning numbers under 80% reflect overconfidence). Prior experience refers to the knowledge the participants had acquired regarding calibration and overconfidence prior to this experiment's start. Participants who answered 'Yes' indicated they had received prior training or learning regarding calibration and overconfidence. 'Partial' referred to participants who believed they understood the concepts at least vaguely. 'No' referred to the participants believing they had no understanding of calibration or overconfidence.

Table 1. Descriptive statistics for demographic variables including prior knowledge of calibration by group

	Overall	Control	Experimental
N	54	29	25
Gender (%)	M: 72 F:28	M: 72 F: 28	M: 72 F: 28
Age (SD)	22.0 (3.0)	22.7 (1.3)	22.3(4.2)
Prior Experience (%)	Yes: 57 Partial:28 No:15	Yes: 72 Partial:24 No: 4	Yes: 40 Partial:28 No:32

Looking at the figure, ones sees clear evidence of overconfidence across both groups and tests with none of the 95% CIs containing the 'expected' 0.8 proportion correct. The two groups seem to show similar levels of calibration on the Pre-Training questionnaire and Domain Specific Test but differ on the Post-Training questionnaire.

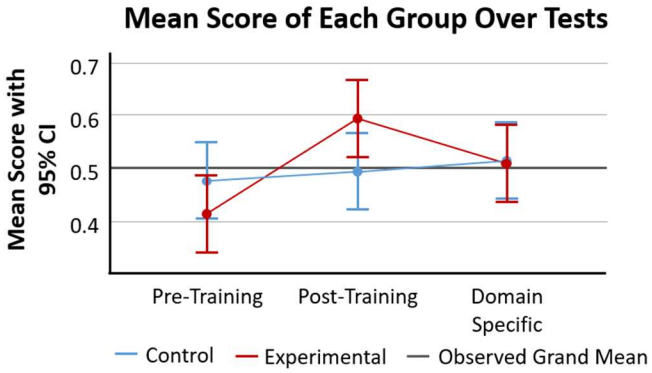

Figure 1. Calibration by group and condition.

Repeated Measures ANOVA

A Repeated Measures ANOVA was used in SPSS to test the two hypotheses simultaneously. Table 2 summarises the significant results from this.

As shown in Table 2, participant's calibration scores differ across the three tests and there is also an interaction between test and group – supporting the observations made above. Independent samples t-tests were used, post-hoc, to compare the mean calibration scores of the Control Group and Experimental Group for each of the three tests as shown in Table 3. The tests indicated that, for both the Pre-Training questionnaire and Domain Specific questionnaire, the difference observed in mean calibration score between the Experimental and Control Group was not significant. However, there was a significant difference in the means of the Experimental and Control Groups on the Post-Training

questionnaire with the mean calibration score of the Experimental Group noticeably higher. This supports Hypothesis 1 – that the feedback training improved the experimental group's Post-Training questionnaire results.

Table 2: Significant results of RM ANOVA

Comparison	F(df)	F-value	P-value
Questionnaire	F(2,104)	10.4	<0.001
Questionnaire*Group	F(2,104)	6.1	0.003

Table 3: Independent t-tests between Experimental and Control Group for each questionnaire.

Questionnaire	t(52)	p
Pre-Training	0.925	.359
Post-Training	-2.189	.033
Domain Specific	0.164	.871

Paired samples t-tests were used, post-hoc, to compare the relative difficulty of the tests, and to verify improvements in calibration observed in the independent samples t-tests. The tests, shown in Table 4, indicated that differences in the means between all the Control Group's tests were non-significant. That is, the tests were equally difficult for the Control group. Conversely, the tests indicated that differences in the means between all Experimental Group tests were significant – as shown in Table 5.

Table 4: Paired t-tests between each questionnaire for the Control Group.

Comparison	t(28)	P
Pre-Training– Post-Training	0.852	.402
Post-Training– Domain Specific	0.627	.536
Domain Specific– Pre-Training	1.315	.199

Table 5: Paired t-tests between each questionnaire for the Experimental Group.

Comparison	t(24)	p
Pre-Training– Post-Training	-5.368	0.000
Post-Training– Domain Specific	3.633	0.001
Domain Specific– Pre-Training	-2.439	0.023

The results of the ANOVA and t-tests, along with observation of Figure 1, suggest no significant difference in calibration scores in the Control Group – as would be expected. However, the figure and analyses show that calibration score for the Post-Training questionnaire of the Experimental Group is significantly higher than both the Experimental Group's Pre-Training questionnaire, and the Control Group's Post-Training questionnaire, which is taken as evidence that feedback training improved calibration.

The near-identical scores of the Control and Experimental groups on the Domain Specific questionnaire, however, suggests this benefit did not transfer to the new domain. That is, despite all tests using the same question format (80% confidence intervals), the change in domain was seemingly sufficient to prevent the training transferring, meaning Hypothesis 2 was not supported. A caution to this interpretation, however, is the observation that the Experimental Group's calibration in the Domain-Specific questionnaire was significantly higher than in the Pre-Training questionnaire, but statistically no different to Control Group's calibration in the Domain-Specific questionnaire. This discrepancy is explored further, below.

Discussion

Experimental Findings

Baseline Measure

The performance on the Pre-Training questionnaire between the Experimental Group and the Control Group suggested both groups were similarly calibrated, indicating that the method for dividing participants into two groups was successful and that the control group can, justifiably, be compared to the experimental group as a baseline.

The consistent results of the Control Group across all tests similarly showed that each test was of similar difficulty, justifying comparisons between tests within a group.

Feedback Effectiveness

The comparison between the mean calibration scores of the Pre-Training questionnaire and Post-Training questionnaire of the Experimental Group shows that the feedback was effective - to a degree. This was reinforced through the comparison of the Experimental Group and the Control Group for the Post-Training questionnaire, which also found a significant result. Between the Pre-Training questionnaire and the Post-Training questionnaire for the Experimental Group, calibration scores improved by 17% (from 42% to 59%). This improvement in calibration was expected, as a wealth of previous research has shown that performance feedback training improves a subject's calibration (Adams & Adams, 1961; Lichtenstein & Fischhoff, 1980; Moore et al., 2017; Stone & Opel, 2000).

Transfer

As noted above, the comparison of the Experimental and the Control Groups on the Domain-Specific Test showed no significant difference, suggesting the Experimental Group was *not* able to transfer their knowledge of calibration to a different domain and thus arguing for recognition failure.

Comparing the Experimental Group's Pre-Training and Domain Specific results, however, showed a significant result driven by an approximately 8% increase (from 42% to 50%) in the Experimental group's mean calibration score. Considering the two tests were of similar difficulty – according to the baseline measure from the Control group - the significant increase in calibration suggests participants *were* able to, at least partially, transfer their skills from the general knowledge domain to the specific domain of petroleum engineering and that this is being obscured in the analyses above by the Experimental Group's slightly lower scores on the Pre-Training questionnaire.

To quantify the extent of the transfer, the Post-Training

questionnaire was compared to the Domain-Specific for the Experimental Group. This showed an ~10% decrease in mean calibration score from the Post-Training questionnaire to the Domain-Specific Test (60% compared to 50%, respectively). This significant difference suggests participants were not able to transfer all of what they had learnt about improving calibration to the new domain. Looking solely at the Experimental group's results in Figure 1 suggests that about half of the improvement seen following training transferred to the Domain Specific Test.

This, of course, contradicts the previous results and the discrepancy between these means that no strong conclusion can be drawn regarding whether the transfer of knowledge between domains did or did not occur. However, one conclusion that can be drawn is that, if the transfer occurred, it is well under 100%, in agreement with Adams & Adams (1961). This is also reminiscent of Glick & Holyoak's (1983) work on analogical transfer, where they argue that incomplete transfer may be due to recognition failure; that is, a failure to recognise the similarities in the problem structure and, hence, to recognise that the skills used successfully in one problem are applicable to the other.

While the question formats used in the three tests herein were identical – asking for 80% confidence intervals - it is possible that having experience in a domain evokes a different thought process to that which may be used to solve general knowledge type questions - reminiscent of knowledge partitioning (Lewandowsky & Kirsner, 2000) - and suggesting that individuals' domain specific knowledge could be separated from their general knowledge and thus processes used to access one may not work for another.

This may have caused the participants to not recognise the similar structure between the general knowledge and specific domain type questions. That is, participants may have simply not recognised that their calibration training should also be applied to the specific-domain questions.

External Factors

Initial Calibration and prior knowledge

Simple comparisons showed participants, regardless of their stated prior experience with calibration (trained, aware or unaware) had similar calibration, and similar improvement after feedback. This is likely due to participants with prior knowledge not being able to apply the knowledge they learnt previously when setting confidence intervals. These results suggest that participants with previous experience with calibration were unsuccessful in reducing their overconfidence long-term, likely due to the fact they did not receive frequent calibration training or regularly practice calibration – as has been observed in previous research (see, e.g., Welsh, Bratvold & Begg, 2005).

Caveats

Sample Size

The sample size was smaller than hoped, as a result of strict time constraints for the project, meaning that statistical power is low. A larger sample might, for example, have helped determine to what extent transfer was actually occurring or whether the effect is an artefact of differences between groups and tests aligning coincidentally. As noted above, the low sample size also resulted in the decision to move participants from the experimental group and control group.

Expertise

The type of questions asked throughout the Domain-Specific Test were designed to relate to the expertise of the participants. As petroleum engineering students, participants have increased knowledge about the petroleum engineering field, but would not be classified as 'experts'. This is doubly true, as the sample includes student participants from different year levels and thus with differing amounts of learning within the field. This concern is somewhat alleviated by the fact that the majority of participants were final year students or recent graduates, who could be expected to have similar levels of understanding of the field (which might, in fact, be less true of professionals further into their careers who tend to specialise into a sub-field). The selection of students of all year levels as the sample, however, meant that, despite all of the questions being related to the oil and gas industry, they had to be kept general enough that all participants could reasonably understand what they referred to – rather than being specific, technical questions that only a fully trained petroleum engineer could understand. That is, while the questions were *about* petroleum engineering, they did not truly test fundamental skills learnt by the participants. Questions more central to the petroleum engineering domain would provide a more accurate measure of knowledge transfer across domains but would require an expert sample.

Testing Conditions

As noted, all tests were online, meaning participants were unable to ask clarifying questions if they did not understand the point of the test - or may have approached the test in unanticipated ways. Although instructions indicated that questions should take no longer than 30 seconds, many participants spent much longer than that on some questions, which may have repercussions on the consistency of the answers. Future work could, therefore, be conducted face to face, or more time be spent explaining the purpose of the test, possibly with the aid of a video. This would make it easier to see if participants are engaged in the test and answering the questions as expected. Conducting training feedback sessions in person could also be beneficial in ensuring that the main points of the training session are highlighted to the participant, so that they can better learn how to improve their calibration for future tests.

Questions

Another concern related to the amount of time available to pilot the general knowledge questions with people similar to the expected participants. As noted elsewhere, the study was conducted as part of the student authors' coursework and, as

such, had to be completed within semester, meaning that time for piloting was limited. While efforts were made to include a wide variety of questions participant feedback indicated there were several questions where some participants did not understand what the question was asking, and hence had no point of reference.

Future Research

As noted in the acknowledgements, this experiment was conducted as part of the first three authors' final-year, undergraduate, research project. As such, there were strict time and budgetary constraints which dictated the approach taken and resulted in unavoidable limitations. Given this, and the equivocal evidence observed herein, larger, more rigorous follow-ups are warranted.

Analogical Transfer

The results from this paper could be extended to determine the true extent at which analogical transfer of calibration training can occur. As shown by Glick & Holyoak(1983), one method to overcome recognition failure and improve transfer is to provide hints about applying the solution of the base problem to solve the analogue problem. In terms of this study, providing hints could simply entail telling participants to apply the training to the specific domain. The purpose of these hints is to remind the participants to use the knowledge and skills learnt from the training on the Post-Training questionnaires, in order to improve calibration. Directly reminding them to incorporate these skills when providing their ranges, would show how much of the training could be transferred in optimal conditions.

Individual Differences

An interesting approach would be to examine responses to a larger study of this type at an individual level – in order to determine whether the group-level improvements are driven by the majority of people improving a small amount or a smaller number of people showing a large improvement in calibration. Which of these better represents the true state of nature has implications for how to improve training processes. If the first, one might consider that better, or more intensive training is required to get participants closer to optimal calibration. If some participants are reaching optimal calibration with the current training, by comparison, the characteristics of or explanations provided by those participants might help improve current training to assist others in achieving similar benefits.

Initial Calibration

The results from this experiment suggested that having prior knowledge of calibration did not influence the participants calibration estimates at any point during the test (in line with previous research from Welsh et al, 2005). An extension to the research could thus conduct a second, Post-Training questionnaire at a later date to determine if the effectiveness of the feedback training remained over time for participants who either had or had not been provided continuing feedback aimed at maintaining better calibration. This could assist in determining how durable any benefits of training are and, thus, how often they need to be reinforced.

Conclusion

Participants in this experiment showed levels of miscalibration in the form of overconfidence (overprecision) consistent with previous literature. The Control group, who received no feedback on their performance, showed very similar levels of overconfidence across the three tests with around half of their (theoretically) 80% interval estimates containing the true value on each test – suggesting that they were appropriately matched for difficulty and that the participants degree of expertise within a specific domain did not alter their degree of calibration relative to the general knowledge domain. Additionally, no benefit was seen for participants who reported having prior experience or knowledge of calibration and overconfidence.

The feedback training provided to participants in the Experimental group proved effective, increasing the number of their ranges containing the true value from 42% to 60%. Whether this benefit transferred to the Domain-Specific Test, however, was less clear, with different analyses pointing in different directions. The Experimental group did not significantly outperform the control group on the Domain Specific Test (in fact, they performed very slightly but not significantly worse). This may, however, reflect their having started from a somewhat lower base – as their Domain Specific Test results *were* significantly better than their own Pre-Training questionnaire results.

Given this conflict, the strongest conclusion that can be drawn is that, while it seems that transfer may have occurred, it was less than complete and that future research is needed to more accurately determine the bounds on the efficiency of transfer of expertise in calibration across domains.

Acknowledgements

This study was conducted as part of the first three named authors' final year Petroleum Engineering Project at the Australian School of Petroleum. These authors contributed equally to the project and this paper and are named alphabetically. MBW is supported by ARC Linkage Grant LP160101460, which was awarded to SHB and includes support from Santos and Woodside. Please address correspondence regarding this article to MBW.

References

Adams, P. A., & Adams, J. K. (1958). Training in confidence-judgments. *The American Journal of Psychology*, *71*(4), 747-751.

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological review*, *68*(1), 33.

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

domains. *Journal of experimental psychology: Applied*, *5*(1), 76.

Capen, E. C. (1976). The Difficulty of Assessing Uncertainty (includes associated papers 6422 and 6423 and 6424 and 6425). *Journal of Petroleum Technology*, *28*(08), 843-850.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, *47*(3), 153-176.

Fischhoff, B. (1981). *Debiasing* (No. PTR-1092-81-3). DECISION RESEARCH EUGENE OR.

Glick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology*, *15*(1), 1-38.

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory & cognition*, *28*(2), 295-305.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational behavior and human performance*, *26*(2), 149-171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In *Decision making and change in human affairs* (pp. 275-324). Springer, Dordrecht.

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.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., ... & Tenney, E. R. (2016). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552-3565.

Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological bulletin*, *115*(3), 381.

Rumelhart, D. E. (1989). of human reasoning. *Similarity and analogical reasoning*, 298.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Welsh, M. B., Begg, S. H., & Bratvold, R. B. (2007, January). Modelling the economic impact of common biases on oil and gas decisions. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Welsh, M. B., Bratvold, R. B., & Begg, S. H. (2005, January). Cognitive biases in the petroleum industry: Impact and remediation. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.