In What Sense is P(A|B) P(B) = P(A,B)? The Relationship between Distributional Format and Subjective Probability Estimates

Belinda Bruza, Matthew B. Welsh, Daniel J. Navarro & Stephen H. Begg
({belinda.bruza, matthew.welsh, daniel.navarro, steve.begg}@adelaide.edu.au)
The University of Adelaide, SA 5005, Australia

Abstract

The elicitation of uncertainty is a topic of interest in a range of disciplines. The conversion of expert beliefs into probability distributions can play a role in assisting key decisions in industry. However, elicitation methods can be prone to bias. In this paper we investigate the effect of changing the presentation of stimulus information and question format on elicited judgments of marginal, conditional and joint probabilities. Participants taught a probability distribution in one structure were expected to have difficulty assessing the distribution in another structure. While this pattern was not found, it turned out that training participants on the more difficult task (learning from a conditional structure) improved overall performance.

Keywords: decision making; cognitive biases; elicitation; probability learning

The "elicitation of uncertainty" is a general term that is often used to refer to methods for translating a set of implicit beliefs into an explicit probability distribution (Wolfson, 2001). The reason for using these methods is to allow researchers to incorporate subjective expert knowledge into a quantitative model that makes predictions about future events (Morgan & Keith, 1995). In view of this, good elicitation methods can play an important role in guiding decision making in a range of industries in which uncertain outcomes are central.

One of the main impediments to widespread use of elicitation techniques in applied settings is the inherent difficulty of the task. This difficulty is caused by the many well-known decision-making heuristics and biases, which can distort the estimates of the underlying beliefs. For instance, anchoring and adjustment, representativeness, availability, base rate neglect and overconfidence (see Tversky & Kahneman, 1974; Bar-Hillel, 1980; Lichtenstein, Fischoff, & Phillips, 1982) have all been found to influence the judgments people make in an elicitation context, in both lay and expert populations (see, e.g., Eddy, 1982; Welsh, Bratvold & Begg, 2005). Moreover, people often mistake conditional probabilities for joint probabilities (Pollasek et al., 1987) since these are easier to compute (Lewis & Keren, 1999), and often experience difficulties with characterizing the conditioning event (Bar-Hillel & Falk, 1982). People may confuse one conditional probability $P(A \mid B)$ with another $P(B \mid A)$, or have difficulties interpreting instructions related to probability (Bar-Hillel, 1980; Fiedler et al., 2000).

Problem Representation

A consistent finding in the decision-making literature is that people are sensitive to the surface representation of a problem. For instance: options described in terms of gains are evaluated differently to the same options when described in terms of losses (Kahneman & Tversky, 1979); changing the surface form of the Tower of Hanoi problem can alter the difficulty of the task (Gunzelmann & Blessing, 2000); and statistical problems expressed in terms of frequencies seem to be easier than the same problems described in terms of probabilities (Gigerenzer & Hoffrage, 1995).

One interesting variation on the question of problem representation arises when people need to learn about and report on the joint distribution of two variables, *A* and *B*. Mathematically, we can describe the distribution to be learned and subsequently elicited in three formally equivalent ways, by noting that:

$$P(A, B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A) \qquad (1)$$

For the current purposes we refer to each of these three variations as a "problem format", and note that while all three formats describe to the same distribution over *A* and *B*, there is no guarantee that people will treat them as such. Indeed, in view of the known differences in how people estimate marginal probabilities, conditional probabilities and joint probabilities, we would expect to observe fairly substantial differences between formats.

In this paper we describe an experiment that examines (1) whether one format for the problem leads to superior learning and subsequent probability estimation in general, and (2) whether learning in one format makes it easier to report on questions framed in the same format. Should either of these two effects be observed, a natural method for improving elicitation in an applied context would be to alter the presentation format to be more suited to the expectations of the expert whose beliefs are to be elicited.

Method

Participants

Participants were 60 students (18 male) studying at the University of Adelaide, aged 18 to 37 years, and were paid $15 for their time.

Procedure

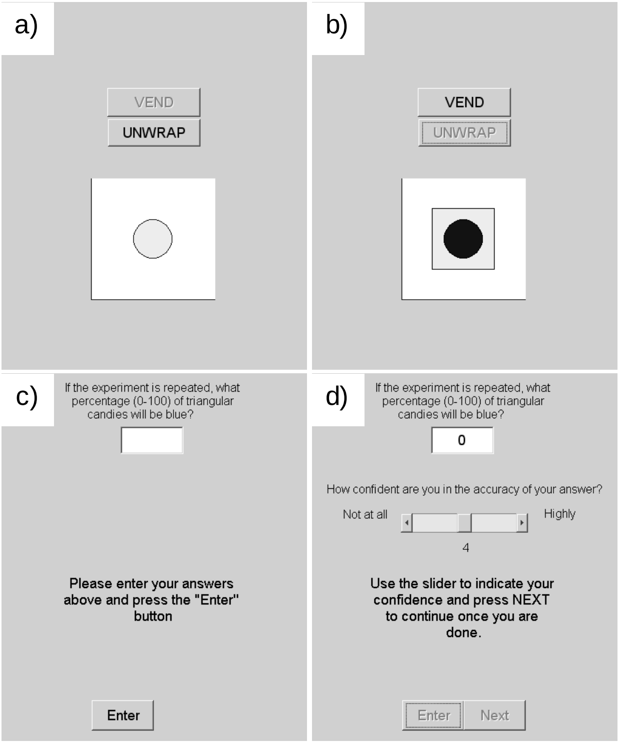
The experiment involved three learning tasks, and two testing conditions, and the measurement of several key covariates. All participants completed all three learning tasks, but were tested in only one of the two testing conditions (based on a random assignment to one of two groups). The basic procedure was as follows. Participants were individually tested in a quiet, well-lit room in front of a computer. Firstly, basic demographic data were collected. Participants then did a simple practice task to demonstrate how the interface works and to illustrate what they would be tested on. Participants then undertook all three learning-plus-elicitation tasks in a random order, with the covariate measurement tasks (APM & MHV; see later) used as filler tasks to help prevent order effects and learned probabilities from previous urn distributions affecting recall of later distributions. Participants were not allowed to use external resources (e.g., pen and paper, calculator) to aid calculations.

The learning tasks

The experiment involved showing participants 20 "candies" which could vary in color (red or blue) and shape (circle or triangle). The participants' task was to learn the distribution over colors and shapes. The experiment was conducted on computer, and the interface was designed so that the stimuli could be presented to participants in all three formats (i.e., *P(A, B)*, *P(A | B) P(B)* and *P(B | A) P(A)*). The cover story told participants that they had encountered a "vending machine" (which we refer to as the urn) filled with candies, which was varied slightly between conditions. Participants were shown the 20 candies one at a time: each candy appeared after the participant clicked on a "vend" button (see Figure 1). After viewing all candies, they were asked various elicitation questions (described later).

In the *wrapped* candy condition, participants were told that the candy was covered in a yellow wrapper. As a result, when they clicked on the "vend" button (see Figure 1) they would be able to see the shape of the candy but not its color. If they then clicked the "unwrap" button, the color would be revealed. Because of the sequential way in which the stimulus characteristics were revealed, the format in which "the world" presents the items is naturally described in terms *P(color | shape) P(shape)*.

In the *masked* candy condition, the distribution was also shown to people in a sequential fashion. However, the color of the candy was shown before the shape, so that participants would see items in a *P(shape | color) P(color)* format. The cover story in this case implied that the participants were initially viewing the candies through a small window, so they could see the color but not the shape. In this condition, the "unwrap" button was replaced by a "retrieve" button, which then revealed the shape.



If the experiment is repeated, what percentage (0-100) of triangular candies will be blue?

Please enter your answers above and press the "Enter" button

If the experiment is repeated, what percentage (0-100) of triangular candies will be blue?

How confident are you in the accuracy of your answer?

Use the slider to indicate your confidence and press NEXT to continue once you are done.

Figure 1: GUI of wrapped candy condition. Vended circular candy (a) unwrapped to reveal blue color (b). Percentage estimate requested (c) before confidence rating (d). All GUIs presented the same basic layout.

The *unveiled* candy condition was the simplest of the three, and presented the two features together as soon as the participants clicked on the "vend" button. As a consequence, participants observed the joint distribution *P(color, shape)* in a more direct fashion.

To allow for between-participant comparisons, the base rate for each type of candy was preset in all three conditions (see Table 1). The shape and color of each candy was randomly determined at each trial. After completing 20 trials, the elicitation questions were asked.

The elicitation questions

Participants answered 10 possible questions about the percentage of particular candies in a *future* urn distribution (two regarding marginal probabilities, four conditional probabilities, and four joint probabilities). The questions were asked in a random order. Participants in group 1 were asked to give estimates in terms of a "shape preceding color" structure. These estimates were therefore elicited in the same format in which the distribution of candies was learnt in the wrapped candy condition (e.g., *P*(*circle*), *P*(*red* | *circle*), *P*(*red*, *circle*) etc). Participants in group 2 were requested to give estimates in terms of a "color preceding shape" structure, hence estimates were elicited in the same format in which the distribution of candies was learnt in the masked candy condition (e.g., *P*(*red*), *P*(*circle* | *red*), *P*(*circle*, *red*) etc).

Thus, in order to produce estimates, participants in group 1, for example, needed to "flip" the probability distribution (using Bayes' theorem) that they learnt for candies in the masked candy condition (see Table 1). As shown in Figure 1, the elicited percentage was typed in an editable text box. Additionally, for every probability judgment that participants were asked to make, they were subsequently asked rate their confidence in their accuracy, using a horizontal scroll bar to enter a value that ranged from 1 (*not at all*) to 7 (*highly*). This process was repeated for each elicitation question. All GUI controls were sequentially locked and unlocked to prevent backtracking and to ensure that the participant answered questions in the prescribed order.

Covariate controls

Given that participants with higher cognitive functioning have been found to perform better on tasks involving conditional reasoning (Stanovich & West 1998) and to be less susceptible to overconfidence (Pallier et al., 2002), intelligence measures were included as controls. Bors and Stokes' (1998) short form of Raven, Court and Raven's (1988a) Advanced Progressive Matrices (APM) was used to measure fluid intelligence. Crystallized intelligence was measured using Senior Form 1 of the Mill Hill Vocabulary Scale (MHV) (Raven, Court, & Raven, 1988b). Finally, information regarding participants' TER (percentile Tertiary Entrance Rank derived from students' performance in the final year of secondary education in several Australian states) was collected.

Results

The accuracy of any given judgment was assessed in terms of the absolute error – the magnitude of the difference between the empirical probability experienced by the participant, and the participant's subjective estimate of that probability. Since the distribution of absolute errors was skewed to the right, a log transformation was performed on absolute error data points prior to model fitting (with the addition of 1 to each data point to prevent negative values).

Order, format and question type effects

It was hypothesized that participants taught a probability distribution in one conditional structure would have difficulty estimating probabilities in another conditional structure. Since group 1 participants were asked to answer questions consistent with the format learnt in the wrapped candy condition (i.e., a shape preceding color structure), they were expected to give estimates closer to the empirical rate than would group 2 participants. The same was expected for group 2 participants in the masked candy condition (i.e., a color preceding shape structure). Because a joint distribution was presented in the unveiled candy condition, question format was expected to have no effect on performance in either group.

Table 1: Base rates of candy color (red or blue) and shape (circle or triangle) and consistency of question format with presentation of candy features in each of the three conditions for group 1 and group 2. Since the unveiled candy condition contained a joint distribution, question format was neither consistent nor inconsistent.

	Condition		
	Wrapped	Masked	Unveiled
	Average base rate (%)		
Color			
Red	10	30	30
Blue	90	70	70
Shape			
Circle	30	90	30
Triangle	70	10	70
	Format consistent		
Group 1			
Shape, color	Yes	No	–
Group 2			
Color, shape	No	Yes	–

Examination of the relationship between questions of conditional probability and log absolute error in Figure 2a) showed what may be weak evidence for the predicted effect. That is, group 1 produced better conditional probability estimates in the wrapped candy condition, and group 2 produced better conditional probability estimates in the masked candy condition. There was also an effect of question type with the log absolute error score highest on questions of conditional probability (see Figure 2b).

Note that in the experimental phase there were four sets of questions that should sum to 100%: questions 1 and 2, which concerned marginal probabilities; 3 and 4; 5 and 6, which concerned conditional probabilities; and 7 to 10, which asked for joint probabilities. Errors within each set should therefore be positively correlated (e.g., if a participant estimated 50% of candies would be circular when the true value was 25%, the absolute error would be 25% and a similar absolute error score would thus be expected in their estimate of triangular candies). Moreover, there were participant-level correlations – some participants consistently had poorer or better performance than others. Linear mixed effects models were therefore fitted to further investigate the effect of condition (wrapped candy, masked candy and unveiled candy), group (1 or 2) and question type (marginal, conditional or joint) on absolute error while adjusting for interdependence of the data.

To adjust for the dependence in estimates within the same question set and within estimates from the same participant for a condition, random effects for participant and question set × condition × participant were added to the linear mixed effects models. Condition, group and question type were treated as fixed effects (predictor variables) in the model. The three-way interaction

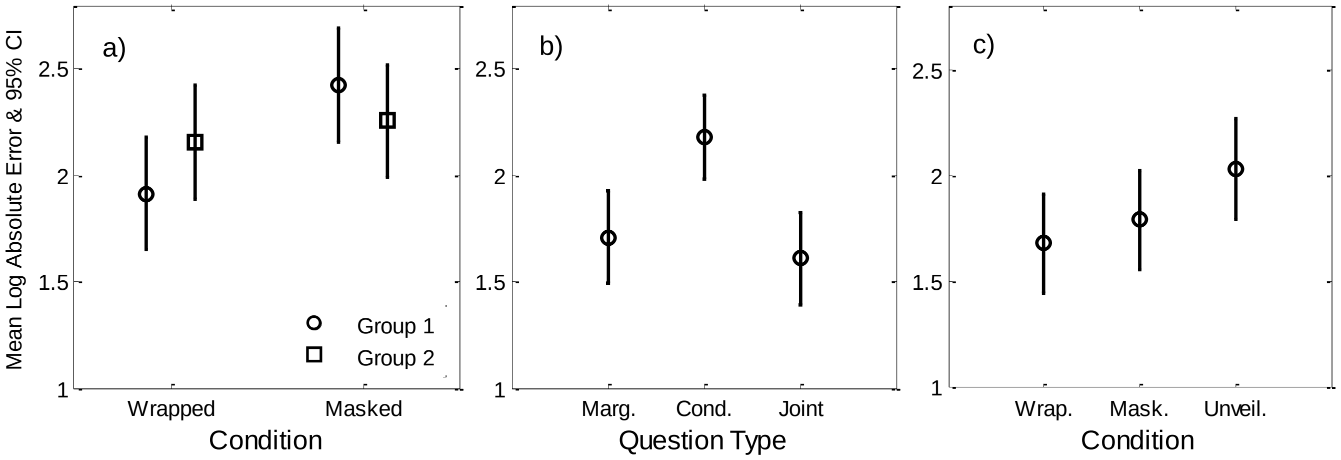


Figure 2: Mean log absolute error scores, with 95% confidence intervals, for (a) group 1 and 2 estimates of conditional probability in the wrapped and masked candy conditions; (b) combined estimates of marginal, conditional and joint probability; and (c) combined estimates in wrapped, masked and unveiled candy conditions. Group 1 *N* = 30, Group 2 *N* = 30. Sample size of estimates is *N* = 240 in each candy condition in (a); *N* = 120 for marginals, *N* = 240 for conditionals, *N* = 240 for joint in (b); and *N* = 600 for each candy condition in (c).

between these variables and all two-way interactions were examined. APM, MHV and TER scores were also included as fixed effects in the models to assess their influence on absolute error. Degrees of freedom were calculated using the containment method (see Littell et al., 1996).

There were no significant interactions so interaction effects were removed from the model. Significant main effects were found for question type *F*(2, 1061) = 31.38, *p* <.001; and condition (i.e., urn type), *F*(2, 1061) = 4.75, *p* < .01, as can be seen in Figures 2b) and 2c).

Bonferroni post-hoc tests indicated questions of conditional probability (adjusted *M* = 2.18, *SE* = .10) were associated with higher log absolute error relative to questions of marginal probability, (adjusted *M* = 1.71, *SE* = .11), *F*(1, 1061) = 25.40, *p* <.001; and questions of joint probability, (adjusted *M* = 1.61, *SE* = .11), *F*(1, 1061) = 55.20, *p* <.001.

The unveiled candy condition (adjusted *M* = 2.03, *SE* = .12) had a significantly higher log absolute error than the wrapped candy condition, (adjusted *M* = 1.68, *SE* = .12), *F*(1, 1061) = 9.12, *p* < .01 and masked candy condition, (adjusted *M* = 1.79, *SE* = .12), *F*(1, 1061) = 4.12, *p* = .04.

Intelligence and accuracy

Participants with higher APM, MHV and TER scores were expected to provide more accurate probability estimates and a significant main effect was found for APM, (*F*(1, 1061) = 3.20, *p* = .04). Looking at Table 2, it seems that MHV scores were also weakly related to accuracy on the estimation task, with 8 of 9 correlations in the predicted direction (*p* = .002 by a sign test), four of which were significant in their own right. TER scores, however, had no predictive power. Independent samples t-tests confirmed that there was no significant difference between groups on the covariates, specifically: the APM (group 1 *M* = 10.47, *SD* = 2.16; group 2 *M* = 10.77, *SD* = 2.93; *t*(58) = –.45, *p* = .65); and MHV (group 1 *M* = 58.37, *SD* = 10.48; group 2 *M* = 56.40, *SD* = 10.53; *t*(58) = .73, *p* = .47).

Table 2: Spearman correlations between MHV score, APM score, TER score and log absolute error broken down by question type.

		Condition		
Question type	Score	Wrapped	Masked	Unveiled
Marginal	MHV	–.08	–.11	.01
	APM	–.29**	–.23**	–.20*
	TER[a]	–.13	.02	.20
Conditional	MHV	–.09	–.11*	–.08
	APM	–.06	–.11	–.26**
	TER[a]	.12	.05	.02
Joint	MHV	–.16**	–.15**	–.12*
	APM	–.10	–.12*	–.19**
	TER[a]	–.06	.03	–.02

*Note.***p* < .05, ***p* < .01, one–tailed. *N* = 60, unless otherwise indicated. [a]*n* = 41. Sample size of estimates is *N* = 120 for marginals, *N* = 240 for conditionals, *N* = 240 for joint.

Confidence and accuracy

It was predicted that confidence ratings would decrease as absolute error scores increase. All correlations were significant and in the expected, negative direction (see Table 3).

Linear mixed effects models were also fitted to assess the relationship between confidence rating and absolute error. The relationship between confidence rating and log absolute error was highly significant – with every one unit increase in confidence rating, log absolute error was expected to decrease by –.19 units. That is, an approximately 20% reduction in absolute error, *t*(1559) = –7.31, *p* <.001. No significant interaction effects were found but a significant main effect was found for condition, *F*(2, 1061) = 14.69, *p* <.001.

Table 3: Spearman correlations between confidence rating and log absolute error broken down by question type and condition.

	Condition		
Question type	Wrapped	Masked	Unveiled
Marginal	$–.32^{**}$	$–.22^{**}$	$–.31^{**}$
Conditional	$–.25^{**}$	$–.20^{**}$	$–.13^{*}$
Joint	$–.27^{**}$	$–.16^{**}$	$–.19^{**}$

*Note.***p* < .05, ***p* < .01, one–tailed. *N* = 60. Sample size of estimates is *N* = 120 for marginals, *N* = 240 for conditionals, *N* = 240 for joint.

Bonferroni post-hoc tests indicated confidence ratings were significantly lower for the unveiled candy condition (*M* = 3.92, *SE* = .20) compared to the wrapped candy condition (*M* = 4.57, *SE* = .20), *F*(1, 1061) = 28.41, *p* < .001; and the masked candy condition (*M* = 4.35, *SE* = .20), *F*(1, 1061) = 12.53, *p* = <.001.

Discussion

In this study we found no significant evidence to suggest that performance on probability estimation tasks changes as a function of the order in which information is acquired. When items were presented in the *P(A | B) P(B)* format, there was no advantage to eliciting participants' knowledge in this same format, as compared to eliciting the knowledge in the *P(B | A) P(A)* format. However, we did find that participants who were shown the stimuli in the *P(A, B)* format actually had significantly higher error than participants taught in either of the other two formats, regardless of what type of question was asked. Given that joint probabilities are presumably easier to process than conditionals, one possibility is that this is a depth of processing effect (Craik & Lockhart, 1972). Recall that, when studying urns with a conditional structure, participants were presented with one characteristic of the candy at a time. This two stage learning process presumably required more attention, involvement and time spent to process each stimulus than the one stage learning process of the joint distribution. This may have contributed to the improvement in overall performance, precisely because the task is harder.

The expected effect of question type was also observed. Absolute error was smallest on questions related to marginal probabilities, and largest on questions related to conditional probability. This was observed regardless of question format or distribution format. These findings are consistent with previous research (see, e.g., Lewis & Keren, 1999), as is the relationship between accuracy and intelligence (see Stanovich & West, 1998). Finally, participants did seem to be aware of how accurate their performance was, since confidence and accuracy were related in a sensible fashion.

Future directions

Our finding that training on the more difficult task improves elicitation warrants further investigation. Future research could determine whether performance is improved by only the two stage learning process used here or by any training format that fosters increased depth of processing.

Limitations

Before concluding, it is worthwhile considering the limitations of this study. It should be noted, for example, that participants provided estimates for each urn distribution based on only 20 trials, which may not have been sufficient for them to form strong beliefs about the distribution. Increasing the number of trials to 100 might allow participants to get a better sense of the underlying distributions, while a larger sample size would enable a clearer understanding of the results; for example, clarifying whether the suggestive results seen in Figure 2a) actually reflect the hypothesized interaction between learnt distributional formats and probability estimates.

A secondary concern is the level of control over the empirically observed rates; although the "true" base rate for each urn was the same, random draws from the true distribution contain sampling error that results in participants observing slightly different empirical rates from each other, diluting control over the experiment. One solution to this would be to use a pseudo-random distribution with a fixed empirical rate, rather than the truly probabilistic approach taken here.

A third possibility is that the sequential presentation method did not have a strong effect because only one stimulus (the candy) was perceived. That is, the nature of the task may have undermined the experimental manipulation to some extent. A task in which *A* and *B* refer to distinct but causally related stimuli (instead of two features of a single object) might provide a better test of the hypothesis.

Conclusions

Although one of the main predicted effects did not appear, the overall results paint an intriguing picture of the potential impacts that training format can have on elicited probability estimates. For example, the fact that training people on the harder task improves estimates is interesting, and of potential applied value. The longer-term goal is thus to see how well these findings can be adapted to improve the elicitation of uncertainty in real world contexts.

Acknowledgments

MBW and SHB were supported by ExxonMobil and Santos' contributions to the CIBP at the Australian School of Petroleum. DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). We thank Thomas Sullivan for his assistance with the statistical models.

References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211-233.

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition, 11*, 109-22.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*, 382-398.

Bratvold, R. B., Bickel, J. E., & Lohne, H. P. (2007) Value of information in the oil and gas industry: past, present and future. *Paper presented at the 2007 SPE Annual Technical Conference and Exhibition, 11-14, November, Anaheim, California.*

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11, 671-684.*

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General, 129,* 399-418.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704.

Gunzelmann, G., & Blessing, S. B. (2000). Why are some problems easy? New insights into the Tower of Hanoi. In L. R. Gleitman and A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (p. 1029). Mahwah: NJ: Lawrence Erlbaum Associates.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263-292.

Lewis, C., & Keren, G. (1999). On the difficulties underlying Bayesian reasoning: a comment on Gigerenzer and Hoffrage. *Psychological Review, 106,* 411–416.

Lichtenstein, S., Fischoff, 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*. Cambridge: Cambridge University Press.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfsinger, R. D. (1996). *SAS system for mixed models*. SAS Institute.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press.

Morgan, M. G., & Keith, D. W. (1995). Subjective judgments by climate experts. *Environmental Science and Technology, 29*(10), 468A-476A.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology, 129*, 257–299.

Pollatsek, A., Well, A. D., Konold, C., Hardiman, P., & Cobb, G. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40,* 255–269.

Raven, J. C., Court, J. H., & Raven, J. (1988a). Manual for Raven's Progressive Matrices and Vocabulary Scales. London: H. K. Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1988b). *The Mill Hill Vocabulary Scale Form 1 Senior*. Oxford: Oxford Psychologists Press Limited.

Stanovich, K. E., & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking and Reasoning, 4*, 289-317.

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

Welsh, M. B., Bratvold, R. B., & Begg, S. H. (2005). Cognitive biases in the petroleum industry: Impact and remediation. *Paper presented at the Proceedings of the Society of Petroleum Engineers 81st Annual Technical Conference and Exhibition, Houston, Texas.*

Wolfson, L. J. (2001). Elicitation of probabilities and probability distributions. In E. Science (Ed.), *International encyclopedia of the social and behavioral sciences*. Elsevier Science.