Subjective Probability Interval Estimates:
A Simple and Effective Way to Reduce Overprecision in Judgment

Uriel Haran

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy to the Tepper School of Business at Carnegie Mellon University

Committee:
Don Moore, Haas School of Business, University of California-Berkeley
Ido Erev, Faculty of Industrial Engineering and Management, Technion
Carey Morewedge, Tepper School of Business, Carnegie Mellon University
Baruch Fischhoff, Department of Social and Decision Sciences, Carnegie Mellon University

Table of Contents

Acknowledgements....................3

Abstract....................4

Introduction....................7

Part I: SPIES Produce Estimates with Better Calibration 13

Part II: SPIES Reduce Overprecision in Other Elicitation Formats....................29

Conclusion 52

References....................58

Tables....................64

Figures....................70

Acknowledgements

I wish to thank the members of my committee Don Moore, Ido Erev, Carey Morewedge, and Baruch Fischhoff for their advice and suggestions. Additional thanks go out to my advisor and co-author Don Moore for his guidance and financial support for these projects as well as to Ido Erev for his generous support and thoughtful guidance in the latter parts of the work on this dissertation. I am grateful for the resources and facilities provided by the Center for Behavioral Decision Research. I also acknowledge the financial support of the William Larimer Fellowship. I would like to thank Nir Kerem, Dafna Shahaf and Lior Lipshitz for their help with the development of the algorithm for transforming SPIES into confidence intervals, Jeff Easter for programming one of the studies in this paper, Carolina Vélez, Aliza Haider, Maya Hasson and the CBDR summer interns for their help with data collection and Lawrence Rapp for his administrative aid. Thanks to seminar attendees at the Tepper School of Business, the Center for Behavioral Decision Research and the Faculty of Industrial Engineering and Management at Technion, as well as conference attendees at the Academy of Management Annual Meeting and the Society for Judgment and Decision Making Annual Conference for their thoughtful feedback. I also am grateful for the comments and suggestions of editors and anonymous reviewers at *Psychological Science* and *Judgment and Decision Making*. I appreciate the patience of my family and all the help they provided while I was half way around the world taking this long journey.

Abstract

Overprecision in judgment is the most robust type of overconfidence, and the one least susceptible to debiasing. It refers to people's excessive certainty in the accuracy of their estimates, predictions or beliefs. Research on overprecision finds that confidence intervals, estimated ranges that judges are confident will include the correct answer, tend to include the correct answer significantly less often than what their assigned confidence level would suggest. For example, 90% confidence intervals typically include the correct answer about 50% of the time (Klayman, Soll, González-Vallejo, & Barlas, 1999). By this standard, confidence intervals appear too narrow, or overprecise. This dissertation focuses on effectively reducing this bias.

In this dissertation, I present a novel elicitation method which can reduce overprecision, sometimes eliminating the bias. This method, called Subjective Probability Interval Estimates, or, in short, SPIES, presents the judge with the entire range of possible values, divided into intervals. The judge estimates, for each interval, the probability that it includes the correct answer. Since these intervals include the entire range of possible values, the sum of these subjective probabilities is constrained to equal exactly 100%.

This work presents six experiments, organized in two parts. Part I focuses on the use of SPIES for eliciting quantitative estimates, and tests it against other elicitation methods in three experiments. Experiment 1 included a within-subject comparison of SPIES and two other elicitation methods, namely 90% confidence intervals and 5th and 95th fractile estimates, and found that SPIES produce interval estimates with significantly higher hit-rates than the other two methods. Experiment 2 varied the range which the SPIES task spanned and the number of intervals included in it, and found that SPIES outperformed the confidence interval method across all configurations. Experiment 3 tested the robustness of this effect to different value

scales, and to variations in the extremity of true values on the range. SPIES again produced consistently more inclusive and better calibrated estimates than confidence intervals.

In Part II, I tested whether SPIES can improve estimates in other elicitation formats. Participants made multiple estimates, using SPIES for some and confidence intervals for others. Participants in Experiment 4 produced confidence intervals with better calibration with their assigned confidence level after having used SPIES in a prior estimate than before having practiced with SPIES. This effect held even when the two estimates had no shared content, suggesting that SPIES influence the estimation process, rather than merely increase the amount of relevant information already present in memory when making the second estimate.

Experiment 5 tested the effect of SPIES on subsequent confidence intervals in two types of estimates. When participants could retrieve a relatively homogeneous set of values but were asked to estimate likelihoods of values across a wide range of possible outcomes, they responded by improving the inclusiveness and calibration of their subsequent confidence intervals. However, when the value set in the first estimate was diverse, such that retrieving evidence for the entire range of the SPIES was easy, no effect was observed in subsequent estimates. This suggests that judges do not simply generalize the SPIES process to subsequent confidence intervals. Rather, they might react to the conflict between their knowledge and the estimates they had to make. This conflict may increase doubt, leading to an adjustment in subsequent estimates to account for this uncertainty.

Experiment 6 manipulated the existence of a conflict between participants’ knowledge of the distribution of possible values and the structure of the SPIES task, by varying the value set’s exposure time. When exposure time was very long, participants could assign each interval in the

SPIES task its according likelihood without conflict, and when it was very short, participants' knowledge of the value set came mainly from the SPIES task itself. SPIES did not improve subsequent confidence intervals in either of these two conditions. Rather, only when exposure time was moderate, as it was in Experiment 5, did SPIES result in improved calibration of subsequent confidence intervals.

Together, results of all experiments show that SPIES is an effective method for reducing overprecision in judgment. It allows for the elicitation of more inclusive and better calibrated estimates than those produced by the confidence interval method for a wide variety of estimate types. In addition, it can enact changes in judges' estimation process, such that their subsequent estimates, elicited by traditional methods, display better accuracy. These features make SPIES an effective tool to reduce one of Judgment and Decision Making's most robust biases.

One of the painful things about our time is that those who feel certainty are stupid, and those with any imagination and understanding are filled with doubt.
(Bertrand Russell)

Subjective Probability Interval Estimates

Introduction

The Federal Home Loan Mortgage Corporation, otherwise known as Freddie Mac, provides an online calculator on its website to help potential clients determine whether they should buy a home or rent one[1]. Among the factors included in this calculation is the estimated appreciation rate of the home in question, defined by the website as “the yearly percentage rate that an asset increases in value”. The user provides a percentage value by which, according to her best judgment, her potential home will increase or decrease. However, until recently, entering a negative value (i.e., a forecast that the house’s value will go down) produced an error message: “Please fix the following errors: Appreciation rate must be a number between 0.00 and 100.00.” The design of this online calculator conveyed Freddie Mac’s belief that housing prices can change only between 0% and +100%, with any rate outside this range being improbable. However, according to the Federal Housing Finance Agency (2011), the average yearly appreciation rate of houses in the United States was consistently outside this range from the second quarter of 2007 through the first quarter of 2010, falling as low as −12.03% (and even lower than −28% in some areas). This forecasting error, among others, resulted in Freddie Mac’s near failure. (In December 2010, Freddie Mac finally changed its online calculator to account for house value depreciation.)

[1] (http://www.freddiemac.com/homeownership/calculators)

The failure of Freddie Mac to anticipate a depreciation of U.S. house prices is but one of many examples of overprecision in judgment. Overprecision is a form of overconfidence. A widely studied topic, overconfidence had been considered for decades a unitary phenomenon. Recent research, however, has identified three distinct types of overconfidence (Moore & Healy, 2008), which, under certain conditions, lead to different outcomes. One type of overconfidence is overestimating one's actual ability, performance, level of control, or chance of success. For example, many people overestimate the amount of time they can hold their breath under water. This form of overconfidence is called *overestimation.* The second form, termed *overplacement,* is people's erroneous or exaggerated belief that they are better than others. The famous finding that over 90% of drivers believe they are better than the average driver is an example of this bias, which has also been termed the *better-than-average* effect (Alicke, Klotz, Breitenbecher, & Yurak, 1995; Svenson, 1981). *Overprecision* is the third type of overconfidence. Also referred to as overconfidence in estimates (e.g., Soll & Klayman, 2004), overprecision is the excessive certainty that one knows the truth. It is in this form that overconfidence tends to be the most consistent and extreme (Juslin, Wennerholm, & Olsson, 1999; Klayman et al., 1999). Also, unlike the two other forms of overconfidence, which are negatively correlated with each other and can be attenuated by varying task difficulty, overprecision is particularly impervious to debiasing (Moore & Healy, 2008).

The literature on overprecision has documented numerous consequences of the bias. Physicians and mental health professionals display excessive certainty in their diagnoses, leading to mistreatment of patients (Christensen-Szalanski & Bushyhead, 1981; Oskamp, 1965). Excessive faith in the accuracy of their forecasts and market predictions can lead market traders to engage in too many trades (Daniel, Hirshleifer, & Subrahmanyam, 1998; Odean, 1999).

Research has also documented overprecision among leading experts in the field of climate research, finding that their predictions about the rate of climate change vary greatly, yet are so precise that there is virtually no overlap between them (Morgan & Keith, 1995; Zickfeld, Morgan, Frame, & Keith, 2010). This leads to a lack of consensus regarding the proper response to climate change, and hinders progress toward action. As these findings indicate, overprecision occurs among experts and laypeople alike, a result that has been confirmed in controlled studies, which varied judges' levels of expertise but found overprecision at all levels (Clemen, 2001; Hilary & Hsu, 2011; Juslin, Winman, & Hansson, 2007; McKenzie, Liersch, & Yaniv, 2008).

Most studies of overprecision include laboratory experiments. These experiments typically measure participants' judgments by eliciting a confidence interval—a range of values that the judge is confident, to a certain degree, will include the true value (Alpert & Raiffa, 1982). A perfectly calibrated judge should produce confidence intervals which include the correct answer at a frequency equal to the level of confidence associated with them. For example, if the assigned confidence level is 90%, then nine out of ten confidence intervals should include the correct answer. However, people generally do not display perfect calibration; far from it. Research repeatedly finds that the percentage of confidence intervals which include the true values is lower than their assigned confidence level, suggesting that these confidence intervals are too narrow (i.e., overly precise, see Soll & Klayman, 2004).

Research has proposed several ways to decrease overprecision, most of which concentrate on two revisions in the elicitation of the estimates: a) unpacking the estimate into multiple judgments; b) separating the interval from its fixed assigned confidence level, and letting the judge determine both the interval and her confidence in its accuracy. The effects of unpacking have been achieved by eliciting the confidence interval's low and high boundary values separately (Soll and Klayman, 2004), by eliciting a point estimate for the exact value in question before the production of a confidence interval around it (Block and Harper, 1991), by asking for the assessment of more alternative outcomes to the target outcome (Fischhoff, Slovic, and Lichtenstein, 1978), and by unpacking an estimate of a value one year in the future into multiple estimates of shorter time frames (i.e., 3 months and 6 months, see Bearden, Gaba, Jain, and Mukherjee, 2011).

Studies separating confidence intervals from fixed confidence levels have done so in several ways. Some elicited probability judgments for predetermined intervals (Abbas, Budescu, Yu, & Haggerty, 2008; Glaser & Weber, 2007; Seaver, von Winterfeldt, & Edwards, 1978). Others had one group of judges estimate an interval for a fixed confidence level and then asking another group to assess their confidence in the accuracy of these intervals (Teigen & Jørgensen, 2005). Other studies asked judges to adjust their confidence intervals until they fit the requested level of confidence (Cesarini, Sandewall, & Johannesson, 2006; Winman, Hansson, & Juslin, 2004), or simply asked them to estimate “reasonable” intervals and then ask for their confidence in these intervals (Speirs-Bridge et al., 2010). In all of these studies, there was a modest reduction of overprecision, but never to the point of eliminating the bias. Moreover, none of these methods had an influence on participants’ judgments outside of the specific format offered. In other words, when participants switched back to the traditional confidence interval production method, overprecision returned to its prior high levels.

THE SPIES METHOD

This work introduces a novel method for reducing overprecision in estimates. This method, called *Subjective Probability Interval Estimates* (SPIES), does not directly elicit a confidence interval. Rather, it presents the judge with the entire range of values, divided into intervals of equal width. These intervals can span the entire range, or, in case the value scale does not have pre-determined high and low bounds, a range that includes all plausible values, with an additional interval at each end representing all extreme values which lie outside this plausible range. The judge then estimates a probability for each interval. In case of estimating one true value, this probability is the likelihood that the interval includes the correct answer. For estimates of a population's properties, this probability is the proportion of the population that is included in the interval. Since the SPIES range includes all possible values, the sum of these probabilities is constrained to equal exactly 100% (see Figure 1 for an illustrative example).

The judge's output, then, is a series of subjective probabilities that total 100%. These subjective probabilities can be computed for numerous types of estimates. In addition to the estimated distribution, it is possible to calculate confidence intervals of virtually any width and confidence level, by combining the SPIES' intervals within the SPIES task. For example, from the estimate of a future temperature in Pittsburgh, as in Figure 1, it is possible to calculate the most likely 10-degree and 20-degree intervals, as well as the judge's 70% and 90% confidence intervals, all without having to elicit the judgment from the judge multiple times. The SPIES method, then, offers great versatility and flexibility to the recipient of the estimate.

In this paper I make the argument, and present data to support it, that SPIES can significantly reduce overprecision in two ways. As an elicitation method, SPIES forces the judge to consider all possible values, including ones that often go ignored in the estimation process of other, more instantiated methods. This enables judges to produce confidence intervals of greater width and better calibration. As an intervention for reducing bias, SPIES influences subsequent estimates in other elicitation formats, by inducing judges to revise their estimation process. The remainder of this dissertation will be organized in two parts, each presenting data from three laboratory experiments. Part I will focus on SPIES as an elicitation method. It will present a comparison between estimates made using this method and estimates made using other methods and tests of the robustness of this difference. In Part II, I will explore how making an estimate with the SPIES method can improve the calibration of subsequent confidence interval estimates. Finally, I will discuss implications, theory extensions and possible applications of the SPIES method, within and outside the realm of cognitive research.

Part I

SPIES produces estimates with better calibration

This part of the dissertation focuses on the question of whether SPIES can produce estimates with lower levels of overprecision than traditional confidence interval estimates. The SPIES method includes a number of features which can potentially lead to a reduction in overprecision. First, in the SPIES task, the judge is forced to consider all possible values, including those which she may overlook when estimating confidence intervals. In confidence interval production, the judge thinks of the most likely outcomes, and builds her confidence interval around these outcomes, without an explicit attempt to consider all but the least likely ones. However, in a reverse process, by which the judge considers all possible values and dismisses only the least likely ones, overconfidence can significantly decrease (Yaniv & Schul, 1997). In SPIES, not only does the judge consider all values, she must also assign each of them a probability of being correct. This probability can, of course, be zero, but this would not be due to unintended ignorance of this value by the judge, but rather after some consideration of its likelihood.

The SPIES method also includes other features that are instrumental in reducing overprecision, such as the use of multiple judgments for producing one estimate. Soll and Klayman (2004), unpacked estimates into two or three judgments, whereas Speirs-Bridge et al. (2010) used four judgments to elicit estimates. These methods resulted in lower levels of overprecision. In SPIES, the estimate is unpacked into as many judgments as the number of intervals that make up the task. Also, building on findings of research on format dependence (Juslin et al., 2007; Teigen & Jørgensen, 2005), SPIES elicit probability judgments of fixed intervals, which appear to display less overprecision than do interval estimates for assigned confidence levels. This reduction may be further enhanced by constraining the total probability assigned to outcomes to equal 100%, limiting the tendency to overstate subjective probabilities (Fox & Rottenstreich, 2003; Tversky & Koehler, 1994).

Part I includes three experiments. Experiment 1 compared estimates produced by SPIES to those produced by other, more instantiated methods. Experiments 2 and 3 tested the robustness of this difference to different configurations of SPIES, to forecasts and general knowledge estimates and to bounded, as well as unbounded value ranges. These three studies were published in Haran, Moore, & Morewedge (2010).

EXPERIMENT 1

Experiment 1 compared SPIES to two common methods of eliciting interval estimates. One is the traditional 90% confidence interval, the most widely used method of interval production. The other is the fractile method, which elicits the judge's estimated 5th and 95th fractiles of the distribution and infers the judge's 90% confidence interval from the distance between these high and low bounds. This method was found to produce confidence intervals that include the correct answer more frequently than confidence intervals produced in one estimate (Soll & Klayman, 2004).

Method

Participants. One-hundred three Pittsburgh residents responded to an email solicitation, sent to past participants in studies of the Center for Behavioral Decision Research, inviting them to participate in an online study. One of the participants was randomly selected to receive a $100 prize.

Procedure. In a within-subjects design, participants estimated the high temperature in Pittsburgh one month from the day on which they completed the survey, in three different formats. In a *90% confidence interval* format, participants were instructed to "please enter two numbers (separated by a dash), one low and one high, so that you are 90% sure the actual high temperature in one month will lie inside the range." In a *fractile* format, participants were asked to "please specify a number sufficiently high that you are 95% sure that the high temperature will be BELOW this value one month from today" and to "please specify a number sufficiently low that you are 95% sure that the high temperature will be ABOVE this value one month from today." In addition, participants made *Subjective Probability Interval Estimates* (SPIES)—they were presented with the following temperature ranges: below 40°F, 40-49, 40-59, 60-69, 70-79, 80-89, 90-99, 100-109, and 110°F or above, and responded to the following question: "What is the likelihood, based on what you know, that the high temperature a month from today will fall in each of the following ranges?" The sum of these nine probabilities was constrained to equal 100%. Presentation order of the three formats was randomly determined, and not recorded.

Results

Because the assigned confidence level for intervals produced by the first two methods was 90%, this was also the target confidence level for intervals produced by SPIES. My collaborators and I used an algorithm to calculate these confidence intervals, which identifies the temperature interval with the highest subjective probability and adds its neighboring intervals until the sum of probabilities reaches closest to, but not more than 90%. The algorithm then adds the proportion of the adjacent interval with the next highest probability (or the two intervals on both sides of the aggregated interval, when they are assigned equal probabilities) needed to reach 90%. The resulting confidence interval is referred to henceforth as 90% SPIES.[2] This is a conservative calculation of 90% SPIES, designed to produce a confidence interval out of the

[2] The full algorithm used to calculate 90% SPIES can be viewed online at http://www.sjdm.org/~baron/journal/10/101027/jdm101027.html (see appendix).

fewest possible subjective probability intervals. In cases where an extreme interval (i.e., below 40°F, 110°F or above) was included in a participant's 90% SPIES, that interval's width was calculated as 10°F.

The true temperatures on the days for which participants made their estimates ranged between 67°F and 73°F. A repeated-measures ANOVA comparing the accuracy of participants' estimates across the three methods revealed a significant difference, $F(2, 101) = 4.98$, $p = .009$. 90% confidence intervals and intervals produced by the 5th and 95th fractiles did not differ in their accuracy, both including the correct answer 73.79% of the time ($SD = 44.19$).[3] 90% SPIES, however, included the correct answer in 88.35% of the estimates ($SD = 32.24$), a significantly higher hit-rate than both 90% confidence intervals, $t(102) = 2.88$, $p = .005$, and fractiles, $t(102) = 2.69$, $p = .008$. While 90% confidence intervals and fractiles displayed significant overprecision of 16.21%, $ts(102) = 3.72$, $p < .0005$, the accuracy level produced by SPIES was not significantly different from the 90% confidence level assigned to them, $t < 1$, meaning that these estimates did not exhibit overprecision (see Figure 2).

The SPIES method does not seem to have improved participants' intuition regarding the precise temperature, as measured by the distance between an interval's midpoint and the true answer. A repeated-measures ANOVA revealed a significant method effect, $F(2, 101) = 3.49$, $p = .03$, but the midpoints of 90% SPIES were not significantly closer to the true answer than either those of 90% confidence intervals, $t(102) = 1.39$, $p = .17$, or those between the 5th and 95th fractiles, $t < 1$.

[3] The identical result for 90% confidence intervals and the fractile method appears to be coincidental, as 63 participants were accurate in both their 90% confidence intervals and fractile estimates, whereas 26 were accurate in only one of the two formats.

I also compared the widths of the intervals generated by the three methods. A repeated-measures ANOVA revealed a significant effect of method on interval width, *F*(2,101) = 21.71, *p* < .0005. Within-subject contrasts show that 90% SPIES were significantly wider (*M* = 31.81, *SD* = 11.96) than 90% confidence intervals (*M* = 23.58, *SD* = 14.42), *t*(102) = 5.73, *p* < .0005, but slightly, and non-significantly, narrower than fractiles (*M* = 33.15, *SD* = 22.48), *t* < 1. The fractile estimates' relatively large mean width, as well as their high variability, can be accounted for by the fact that eight of these estimates reached either below 30°F or above 119°F (the boundary values we set for calculating 90% SPIES), and resulted in relatively wide intervals.[4]

Discussion

Of the three methods tested in this experiment, the SPIES method was the only one in which intervals' hit-rates matched their assigned confidence levels. Although 90% confidence intervals and fractile estimates produced higher hit-rates than those which research typically finds (see Klayman et al., 1999), the hit-rate of SPIES was significantly higher than both of these methods. Moreover, the SPIES method not only produced better accuracy, it eliminated overprecision.

Another noteworthy finding is that SPIES produced a significantly higher hit-rate than did fractile estimates. This result suggests that although in both methods judges unpack their estimates into multiple judgments, this feature might not be the primary driver of the superior calibration found in SPIES.

The results of this experiment are not conclusive about why SPIES were more accurate. On the one hand, interval midpoints did not differ between the three estimation formats in their distance from the true value, suggesting the better hit-rate is due to 90% SPIES being more

[4] Only one 90% confidence interval exceeded these boundary values.

inclusive than intervals in the other formats. On the other hand, 90% SPIES achieved a higher hit-rate than fractile estimates without being significantly wider. As noted, this may be due to the constraint put on including extreme values in the SPIES, but not in the other estimates. This issue was addressed in Experiment 3.

One noteworthy difference between the structure of SPIES and the two other methods tested in this experiment is that the confidence interval and fractile tasks do not change when estimating different values. The questions to which the judge responds in these tasks are identical for any kind of estimate. In SPIES, however, this is not the case, and the way in which the SPIES task is structured is up to (and the responsibility of) the person eliciting the estimate. The range on which the true value of an estimate lies is different for every estimate, as is the variability of estimated values. Therefore, the SPIES task should be configured specifically for each estimate. There is reason to suspect that different configurations of the SPIES task may lead to differences in estimate performance. While in Experiment 1 the SPIES task was structured rather arbitrarily (i.e., ranges of 10 degrees in width between two round values, one very low and the other very high), in Experiment 2 my collaborators and I systematically varied the configuration of the task.

EXPERIMENT 2

One rather cumbersome feature of SPIES is that the task needs to be structured specifically for every estimate. For example, the SPIES task employed in Experiment 1 ranged from 40 to 110, a reasonable range for estimating the weather in Pittsburgh in degrees Fahrenheit, but not, for example, the size of the United States' national debt. Variations in the configuration of the SPIES task might then affect the quality of the resulting estimate. Experiment 2 examined this hypothesis more closely, and compared confidence interval estimates to estimates elicited by SPIES in various configurations.

Configuring the SPIES task includes making two choices: how big to make the range of possible responses and into how many intervals to divide this range. These variations may influence the amount of attention given by the judge to the values she considers, and, subsequently, affect the quality of the estimates she produces. In order to test the robustness of the results obtained in Experiment 1 to these variations, Experiment 2 varied the configuration of the SPIES tasks in two ways: one was the width of the range for which estimates were made; the other was the number of intervals into which this range was divided. While these variations may lead to differences in the resulting 90% SPIES, the basic estimation process remains the same for all of them. Therefore, we hypothesized that 90% SPIES will outperform 90% confidence intervals, regardless of the width of their range, or of how many intervals are included in each SPIES task.

Method

Participants. One-hundred twenty-five U.S. participants were recruited through Amazon.com Mechanical Turk to participate in a "Weather Forecasting Survey". They completed the experiment online in exchange for $0.05 each. There were nine instances of multiple responses from the same IP address. The second record from each of these duplicates was stricken from the data file. The final sample consisted of 116 participants.

Procedure. Participants estimated the day's high temperature in Washington, DC exactly one month after the day on which they took the survey. In a 2 x 2 between-subjects design, participants made SPIES with a narrow range (-15°F to 84°F)[5] or with a wide range (-65°F to 134°F), which were divided into either ten or twenty intervals. These divisions resulted in three interval grain-sizes: fine (5°F), medium (10°F) and coarse (20°F). Two intervals of extreme

[5] The highest and lowest temperatures, respectively, ever recorded in Washington, DC in February, the target month for participants' forecasts.

values were added at both ends of these ranges: “-16°F or lower” or “-166°F or lower” at one end, and “85°F or higher” or “135°F or higher” at another end (see Table 1). To compare SPIES with conventional interval estimates, an additional group of participants produced a 90% confidence interval.

Results

Actual temperatures on the days for which participants provided their estimates fell between 31°F and 40°F. First, we compared the hit-rate of 90% confidence intervals to that of estimates made using SPIES. Similar to Experiment 1, 90% SPIES achieved a significantly higher hit-rate (M = 73.91%, SD = 44.15) than 90% confidence intervals (M = 29.17%, SD = 46.43), $t(114) = 4.38$, $p < .0005$. As expected, 90% SPIES of all four configurations produced accurate estimates at a significantly higher rate than 90% confidence intervals, ts ≥ 2.28, ps ≤ .03 (see Figure 3).

Second, we tested whether the different configurations of the SPIES task affected participants’ estimates. A 2 (range width: 100°F / 200°F) x 2 (number of intervals: 10 / 20) between-subjects ANOVA on the hit-rates of 90% SPIES revealed no significant effects of either range width, $F < 1$, or number of intervals, $F(1,88) = 3.23$, $p = .08$, nor was there a significant interaction, $F < 1$ (see Table 2). In order to perform a more conservative test of the effect of range width on participants’ estimates, we compared the two conditions in which participants made SPIES with a medium, 10ºF grain size (see Table 1). These two conditions differed only in range width: one group was presented with a 100ºF range, whereas for the other group, the SPIES task spanned 200ºF. The comparison between these two groups revealed no significant effect of range width on hit-rates (100ºF range: M = 80.95%, SD = 40.24%; 200ºF range: M = 69.23%, SD =47.07%), $t < 1$.

We did, however, find that the width of 90% SPIES was affected by the configuration of the task. We conducted a similar ANOVA on *estimate width*, which revealed significant main effects of the overall SPIES' range width and the number of intervals it included, *F*(1,88) = 12.52, *p* = .001 and *F*(1,88) = 12.25, *p* = .001, respectively, with no interaction, *F* < 1 (see Table 3). However, a comparison of the two 10ºF grain size groups found no effect of range width on estimate width (100ºF range: *M* = 33.40, *SD* = 16.58; 200ºF range: *M* = 33.50, *SD* = 12.93), *t* < 1.

As in Experiment 1, the manipulations did not affect the estimated intervals' midpoints. The distances of 90% SPIES' midpoints from their respective true values did not vary with range width, *F*(1, 88) = 1.47, *p* = .23, or with grain size, *F* < 1, nor was there an interaction, *F* < 1. No significant difference in midpoint accuracy was found between 90% SPIES and 90% confidence intervals, either, *t* < 1.

In light of the significant effects on estimate width and the large, though only marginally-significant, effect of number of intervals on hit-rates, we sought to examine the extent to which participants were sensitive to the different SPIES configurations. We tested this by measuring the number of intervals to which participants assigned some probability higher than zero in their estimates. A 2 (range width) x 2 (number of intervals) ANOVA found a significant effect of interval number, wherein participants for whom the SPIES task consisted of twenty intervals gave significantly more intervals (*M* = 6.36, *SD* = 3.81) non-zero probabilities than those who were presented with only ten intervals (*M* = 4.24, *SD* = 2.30), *F*(1, 88) = 14.94, *p* < .0005. The ANOVA also revealed a significant range width effect, *F*(1, 88) = 22.69, *p* < .0005, but the direct comparison of the two 10ºF grain size groups found no effect of range width on the number of intervals with non-zero probabilities (100ºF range: *M* = 5.14, *SD* = 2.83; 200ºF range: *M* = 4.62, *SD* = 1.79), *t* < 1. Together, these results suggest that participants who made estimates with the finer-grained SPIES were aware of the need to use a larger number of intervals and adjusted their estimates, but not sufficiently to fully equate their estimates' width to those who made estimates with coarser-grained SPIES.

Discussion

As in Experiment 1, SPIES had a significantly higher hit-rate than standard 90% confidence interval estimates. More important, this difference was consistent across various configurations of SPIES. Participants made their estimates using SPIES that spanned a very wide range, including unreasonable values, or a moderate range, including only values previously observed. The SPIES ranges were divided either coarsely divided (up to 20 degrees per interval) or more finely (with intervals as thin as 5 degrees). Inevitably, the inclusiveness of the resulting 90% SPIES varied from condition to condition, but in all conditions, their inclusiveness and calibration was better than what was observed in 90% confidence intervals.

One common feature of the first two experiments is that both included estimates of values on an unbounded scale (i.e., temperatures), for which we did not specify a minimum or a maximum value. In the absence of such explicit bounds, the highest and lowest intervals in the SPIES task may be perceived by the judge as cues, or reasonable bounds, between which the experimenters expect the true answer to lie. Previous studies have shown that providing judges with the relevant range of possible answers improves calibration (Rakow, Harvey, & Finer, 2003). Because such information could have been inferred from the SPIES tasks, but not from the confidence interval estimates, this may account for some of the difference in performance between the two methods.

Also, in both experiments, the true values eventually fell closer to the middle of the range than to any of its ends. One of the proposed advantages of SPIES over other methods is that it prevents the judge from overlooking extreme values when these values are plausible (that is, can potentially be the true answer), but also not highly likely (and thus do not receive much attention in confidence interval production). Therefore, a comparison between SPIES and confidence intervals should also include estimates which true answers are closer to one of the ends of the possible value range than to its center. These issues were addressed in Experiment 3, in which the high and low bounds of the range were specified in all conditions, and the estimates' true values were pre-determined, and thus could be manipulated to be across the entire value range.

Another open question is whether the difference between SPIES and confidence interval production is solely due to the different elicitation format, or whether SPIES enact a change in the underlying process by which estimates are generated. I argue that the ways in which SPIES reduce overprecision are not external to the estimation process, as in, for instance, Winman et al.'s (2004) Adaptive Interval Assessment method, which repeatedly elicits judges' assessments of intervals of varying widths until their confidence in a certain interval's accuracy matches the desired level. Rather, by making judges consider, and assess the likelihoods of values across the entire range, plausible and implausible, SPIES activate different processes of sampling relevant evidence and inferring the most likely values from these pieces of evidence. Therefore, we hypothesized that using SPIES for estimating uncertain quantities may have effects beyond the specific elicitation method, and will affect subsequent estimates made in different formats. This hypothesis was also tested in Experiment 3.

EXPERIMENT 3

Experiment 3 differed from the previous two experiments in several aspects: First, rather than estimating a single value, participants made a series of estimates, each of a different value. Second, participants made estimates rather than forecasts, that is, they estimated items of general knowledge rather than make predictions about future values. They estimated the years in which all 20th Century U.S. presidents were first elected to office. These estimates constitute a third difference between this and previous studies, which is that the values being estimated were on a bounded range, ranging from 1900 to 1999. This limit was made known to all participants in all types of estimates, eliminating a potential alternative explanation for the results of Experiments 1 and 2. In addition, since election years for all presidents of the 20th Century were estimated, the true values fell at various points on the range, both near the ends and closer to the middle.

Last, the elicitation methods used for estimates in this experiment were varied systematically between estimates. Participants used confidence intervals for half of their estimates and SPIES for the others. This design enabled us to test for the influence of SPIES on subsequent confidence interval estimates, by measuring differences in 90% interval widths between confidence intervals produced before SPIES and those produced after. If format dependence is solely responsible for the reduction in overprecision exhibited in SPIES, then, similar to the findings of Winman et al. (2004), confidence intervals will not be affected after switching from SPIES. If, as we suggest, SPIES change the process by which judges make confidence estimates, then 90% confidence intervals should include a wider range of values if made after SPIES than when made beforehand.

Method

Participants. Three-hundred thirty-four Pittsburghers (169 women, *M* age = 22.6, *SD* = 6.79) completed a survey in the lab in exchange for $3 or course credit.

Procedure. Participants answered a 16-item quiz, estimating the years in which all 20th Century U.S. presidents were first elected to office[6]. For each president, participants estimated either a 90% confidence interval or SPIES. The SPIES task included all years from 1900 to 1999, divided into ten intervals, each representing a decade, with no end intervals for more extreme values. Similarly, in the confidence interval production condition, any estimate that included years outside the 20th century could not be submitted, and the participant was instructed to revise it. We used a mixed design, in which half of the participants provided 90% confidence intervals for the first eight estimates and SPIES for the last eight, whereas for the other half this order was reversed. Items appeared in a different random order for each participant.

Results

We calculated 90% SPIES the same way as in Experiments 1 and 2. Next, we conducted a 2 (elicitation method: SPIES / confidence intervals) x 2 (elicitation order: first eight estimates / last eight estimates) mixed ANOVA[7] on *hit-rates*, which showed that 90% SPIES had a significantly higher hit-rate than 90% confidence intervals. SPIES included the correct answer 76.91% of the time (*SD* = 20.17), compared with 54.34% (*SD* = 26.26%) in 90% confidence intervals, $F(1,332) = 192.34$, $p < .001$. This result supported our prediction that SPIES would provide greater accuracy for estimated values in bounded ranges, regardless of where on the

[6] Not including William McKinley, who was first elected in 1896, and Gerald Ford, who was never elected president.

[7] Since we counterbalanced elicitation order between the two groups (i.e., one group estimated SPIES for the first eight estimates and confidence intervals for the last eight, whereas the other group made estimates in the reverse order), the group means are equal to the method x order interaction.
In formal terms, the group main effect is: H_0: $(\text{SPIES}_1 + \text{Conf. Int}_2) – (\text{Conf.Int}_1 + \text{SPIES}_2) =$
$\text{SPIES}_1 + \text{Conf. Int}_2 – \text{Conf. Int}_1 – \text{SPIES}_2 = 0$.
Method x order interaction is: H_0: $(\text{SPIES}_1 – \text{SPIES}_2) – (\text{Conf. Int}_1 – \text{Conf. Int}_2) =$
$\text{SPIES}_1 – \text{SPIES}_2 – \text{Conf. Int}_1 + \text{Conf. Int}_2 = 0$. As you can see, these two equations are the same. Therefore, a difference in the estimates of the two groups would imply a significant interaction between the elicitation method and order (i.e., first eight estimates vs. last eight).

range the true value eventually falls. As in Experiments 1 and 2, we found no significant effect of elicitation method on interval midpoint accuracy, *F*(1, 332) = 1.11, *p* = .29.

A similar ANOVA on *estimate width* yielded a significant effect of SPIES on subsequent confidence interval width. SPIES produced significantly wider estimates (*M* = 36.27, SD = 20.09) than 90% confidence intervals (*M* = 18.17, *SD* = 14.84), but there was also a significant Elicitation method x Elicitation order interaction, *F*(1,332) = 3.97, *p* = .04. Simple effects tests revealed that 90% confidence intervals produced after having taken the SPIES task were significantly wider (*M* = 20.77 years, *SD* = 16.13) than those produced in the first set of estimates (*M* = 15.57, *SD* = 12.95), *t*(332) = 3.25, *p* = .001, whereas 90% SPIES did not differ between the two groups, *t* < 1. This result suggests that SPIES had a carryover effect on subsequent confidence interval estimates, leading judges to consider a wider range of values in their estimates. To rule out learning and time effects, we conducted a repeated-measures ANOVA on confidence interval widths for each item participants estimated. The last confidence interval estimate in each set was not, on average, wider than the first estimate in the set, *F* < 1, suggesting the greater width of confidence intervals made after SPIES than of those made before SPIES was not due to a simple improvement with experience or time within the same elicitation method (see Figure 4).

Discussion

Results of Experiment 3 lead to two main conclusions. First, this experiment extended the findings of Experiments 1 and 2, that SPIES produce interval estimates of reduced overprecision, to different conditions. Specifically, SPIES proved to be a superior elicitation method for estimates of general knowledge as well as forecasts of future values, and for values on all parts of the range. Also, this advantage proved not to be due to a signal provided by the extreme intervals in the SPIES method, as the minimum and maximum possible true values were made explicit to judges during estimates in both formats. These results suggest that judges can benefit from using SPIES for estimates of uncertain quantities in a wide variety of contexts, with robust and reliable results.

Furthermore, this experiment found that SPIES have a carryover effect on subsequent confidence interval estimates. Participants who first made a series of SPIES before switching formats produced 90% confidence intervals that were significantly more inclusive than participants who used 90% confidence intervals in the first set of estimates. This suggests that the reduced bias in SPIES is not only due to format dependence, but also demonstrates a change in the process by which judgments are made. The more extensive consideration of values in SPIES prompted judges to generate wider confidence intervals in later estimates.

SUMMARY

Part I of this dissertation aimed to present the SPIES method as an alternative to the standard elicitation methods of quantitative estimates. In three experiments, the SPIES method consistently produced interval estimates that were more inclusive, and better calibrated, than the standard confidence interval production method. This difference in inclusiveness and calibration was robust to different configurations of SPIES, i.e., different widths of ranges included in the task and different numbers of intervals which the judges had to estimate. This difference was observed in estimates of values on unbounded ranges (temperatures), as well as values on bounded ranges, with specified minimum and maximum values (the years of the 20th Century). Also, whether the estimate was a forecast of a future value or an estimate of general knowledge items did not affect the results, as SPIES provided better estimates for all types of questions.

In addition to demonstrating the power of SPIES as an elicitation method, Experiment 3 provided an interesting result regarding the lasting effect of SPIES on subsequent estimates in other formats. While in prior studies (e.g., Winman et al., 2004) the effect of the bias-reducing intervention was limited to the specific elicitation format, SPIES appear to have an effect on judges' estimation process, such that subsequent confidence interval estimates become more inclusive after having practiced with SPIES.

Part II will focus on examining this carryover effect more closely. In order to fully understand this effect, we must first identify the stage of the estimation process at which the effect takes place, and the nature of the change it induces in the process of making subsequent estimates.

Part II

SPIES reduce overprecision in other elicitation formats

INTRODUCTION

Part I reveals an interesting finding about the SPIES method. Not only can this method produce better estimates than other methods, it also has a lasting effect on participants' judgments, such that their subsequent estimates, even when elicited in a different format, are more inclusive, relative to those who did not use SPIES in prior estimates. This suggests that, unlike previously developed methods of reducing overprecision, the SPIES method influences the cognitive process by which judges produce their estimates, and its effect is not format dependent (Winman et al., 2004)—an artifact of the elicitation format.

There are a number of potential reasons for the lasting reduction in overprecision achieved by SPIES. In the following section, I will discuss the processes by which judges succumb to overprecision and then speculate on the ways by which SPIES can help judges attenuate the influence of these processes.

CAUSES OF OVERPRECISION

When a value is estimated, assuming it is not known and cannot be retrieved from memory, the judge must make an uncertain estimate by inferring the value inductively from knowledge of similar observations. This is essentially a two-step process. The first step includes the retrieval from memory of some known fact about the value in question and a sample of relevant observations. For example, to estimate the population of London, a person may retrieve the fact that London is a European capital, along with a sample of other European capitals and their respective populations (Hansson, Juslin, & Winman, 2008). The second step includes transforming the sampled knowledge into an actual estimate, be it the most probable value (i.e., best guess), a statement about how the value compares against a specific reference point (e.g., whether the population of London is higher or lower than the population of Berlin), a probability judgment (e.g., the probability that the population of London is between 4 million and 6 million) or a confidence interval. This step includes adjusting the amount of knowledge retrieved for the level of uncertainty the judge feels about his or her estimate (Teigen & Jørgensen, 2005), and for how informative the judge wants the estimate to be (Ackerman & Goldsmith, 2008; Yaniv & Foster, 1995, 1997).

Overprecision can result from flaws in either of these two steps (Pleskac, Dougherty, Rivadeneira, & Wallsten, 2009). Incomplete or biased sampling of information from memory can place confidence intervals inaccurately on the range of possible values. Tversky and Kahneman (1973, 1974) demonstrated how values that are accessible in memory at the time of estimating affect the estimates. For example, evidence that is more salient (e.g., evidence that was encountered more recently, or remembered more vividly) tends to be perceived as more frequent or more representative of the population, and thus bias the estimates. Some have proposed that point estimates, "best guesses" made by the judge or another person, produce overprecision by serving as an anchor. This anchor may either bias the judge's search for information in favor of information consistent with the her initial judgment (Ditto & Lopez, 1992), or make it harder to adjust the confidence interval's boundaries far enough from it (Clemen, 2001; Epley & Gilovich, 2004; Seaver et al., 1978). However, empirical tests of this proposition have not found consistent support for it (Block & Harper, 1991).

Overprecision can also arise during the inference process. Soll and Klayman (2004) argue that random variability in the width of confidence intervals will result in more overprecision than underprecision, even when judgment errors themselves are not, on average, biased. Let us assume a single-peak, symmetric distribution of likelihoods for all possible values. The most likely value is inside the confidence interval, while the likelihood of any other value decreases the farther away the value is from the distribution's peak. The variability in setting the boundaries of a confidence interval should then result in asymmetric effects, such that an inward error should result in a loss of higher likelihoods than an outward error of the same size. See, for example, Figure 5 (Soll & Klayman, 2004). Interval I is the hypothetical, perfectly calibrated confidence interval for a certain time estimate, ranging from 1800 to 1840. Intervals J and K were produced by two individual judges. Interval J is 10 years narrower than the perfectly-calibrated interval I while interval K is 10 years wider than this interval. While the average of these two intervals is equal to the perfectly calibrated interval I, the number of observations erroneously included in interval K is less than the number of observations erroneously excluded from interval J. Therefore, interval J will display a greater bias than interval K, resulting in overall overconfidence.

Fiedler and Juslin (2006) also proposed that judges employ generally unbiased retrieval processes. However, they argue, judges fail to account for the differences between their retrieved samples and the populations which these samples attempt to represent, such as the effects sampling strategies, biased estimators, the origins of the samples and the level of variation. According to the Naïve Sampling Model, or NSM (Juslin, Winman, & Hansson, 2007), judges make two errors in transforming their retrieved sample into an estimate of the properties of the population. One is that they perceive the sample to be an exact, unbiased representation of the population, despite the fact that distributions of small samples are, on average, dislocated relative to the population distribution. The second error is the failure to acknowledge that sample variances are smaller than population variances. In order for a sample distribution to be an unbiased estimator of a population's distribution, its variance needs to be corrected by n/n-1 (Kareev, Arnon, & Horwitz-Zeliger, 2002).

HOW SPIES MIGHT REDUCE OVERPRECISION IN INTERVAL PRODUCTION

The SPIES method can potentially reduce overprecision in either of the two steps described above. The retrieval step might be improved by inducing the judge to scan the entire range of possible answers more systematically, which may make subsequent estimates more evidence-rich and thereby more inclusive (Soll & Klayman, 2004). This effect was observed in the studies reported in Part I: when using SPIES, participants' judgments included reference to more possible values than when they produced confidence intervals directly.

SPIES might also influence the inference step of the estimation process. This can happen in one of two ways. One is by training the judge to generalize the wide scan and multiple assessment processes of SPIES to other elicitation formats. In the SPIES task, the judge scans the entire range of possible values and encounters more values deemed likely enough for inclusion in the estimate. The judge may then realize that these search tactics identify plausible values that are not discovered in prior methods and make a more systematic, and perhaps more effortful search for evidence (see Galinsky, Moskowitz, & Skurnik, 2000). By eliciting more information-rich estimates, the SPIES method makes subsequent confidence interval estimates more information-rich as well (Hirt & Markman, 1995; McKenzie, 1997, 1998; Morewedge & Kahneman, 2010).

Another way by which SPIES might affect the inference process is by highlighting a discrepancy between the judge's knowledge or retrieved sample and the multiple assessments requested by the SPIES task. This discrepancy may cause increase the judge's doubt, or prime exertion of more cognitive control in subsequent estimates. Prior research has shown that when people are in a state of low certainty, they tend to engage in greater information processing than in high certainty states (Tiedens & Linton, 2001; Weary & Jacobson, 1997). Botvinick, Braver, Barch, Carter, & Cohen (2001) demonstrated that a response conflict in one trial of a task leads participants to exert more cognitive control, and make less errors in a subsequent trial. Other studies (Simmons and Nelson 2006) suggested that intuitive biases arise when people feel very confident in their intuitions and are unmotivated or unable to doubt them; when this confidence in intuition is undermined, these biases can be reduced. (see also Chaiken, Liberman, & Eagly, 1989; Petty & Cacioppo, 1986). While the spontaneous search process might be the same for SPIES and interval production, in SPIES the judge is asked about all possible values, including those that did not come up in the search. This may lead the judge to be either more thorough or more conservative in her estimation process thereafter.

The remainder of this part of the dissertation includes three experiments, which attempt shed light on how SPIES improve the process of confidence interval production. Experiment 4 tested the robustness of the carryover effect of SPIES and attempted to identify the stage of the estimation process in which this effect takes place. Experiment 5 compared two potential accounts for the change which SPIES enact in judges' estimation processes, and Experiment 6 directly tested the hypothesis that SPIES when there is a conflict between the judge's knowledge or retrieved sample and the response format.

EXPERIMENT 4

Results of Experiment 3 showed that the reduction of overprecision in SPIES was not limited to the specific elicitation format, but also carried over to confidence interval estimates produced by judges who used SPIES in earlier estimates. In Experiment 4, I sought to examine the mechanism by which this carryover effect happens. Specifically, this experiment was designed to test whether the effect of SPIES influences only the amount of evidence retrieved before producing the estimated interval, or whether it also affects the inference stage of the estimation process. In order to test this, I used a novel experimental procedure of eliciting estimates and measuring calibration, called the value population paradigm.

The value population paradigm

For this experiment, as well as the other experiments reported in this part, my collaborators and I developed a novel paradigm for eliciting estimates and measuring their calibration. This paradigm includes presenting participants with a population of 100 numerical values, each corresponding to a percentile point on a certain distribution of numbers on a specified, 100-number range. The numbers are presented in a random order (see Figure 6 for an example).

After the presentation of the value population, participants make an estimate in one of the following two formats: in a 90% confidence interval format, participants estimate a range that includes exactly 90% of the values they saw; in a SPIES format, participants are presented with ten intervals, each spanning a ten-number range, and estimate the percentage of the values they saw that is included in each interval (see figure 7).

This paradigm has a number of distinct advantages. First, it provides a normative standard for precision at the individual estimate level. In interval estimates of general knowledge, the only standard for calibration is whether or not the proportion of estimated intervals which include the true value matches the target confidence level of these intervals. This measure can be misleading. For example, a judge who produces ten 90% confidence intervals, nine of which spanning the entire range of possible values and one interval including only one value (not the correct value) has the same calibration score as a judge who truthfully conveys her subjective 90% confidence in all ten of her 90% confidence intervals, and hits the mark in nine of them. The value population paradigm avoids this problem by setting the goal of each estimate to represent as best as the participant can her knowledge about the population that was presented to her, rather than to include one true value.

Another advantage of this paradigm is that it minimizes the effect of participants' prior knowledge. In general knowledge questions, participants who know the exact correct answers to most of the questions should estimate intervals of minimal width. Their data might not be very informative with regard to their estimation process. In this paradigm, all participants start with the same minimal level of knowledge.

The third and final important feature of the value population paradigm is that it controls the amount of information that can be transferred from one estimate to the next. When each estimate is made on different value populations, placed on non-overlapping number ranges, no specific evidence from one estimate can be relevant for subsequent estimates. Therefore, this paradigm was especially useful in Experiment 4, which focused on cases when there is no overlapping information between estimates.

Experiment 4 was designed to examine whether the effect of SPIES on subsequent confidence interval estimates is limited to the information retrieval stage of the estimation process, or whether it also affects the inference stage, and thus can be applied even in cases where the test estimate for which knowledge from the practice estimate is irrelevant. One argument is that the SPIES method only eases information retrieval, by making estimate-relevant evidence, already retrieved in the initial SPIES estimate, more accessible during information retrieval for the next estimate. However, by doing so, SPIES does not influence the inference stage of subsequent estimates. If this is true, then confidence intervals will be affected by SPIES only when at least some evidence retrieved during the SPIES task can be useful for the next estimate. This argument could explain the results of Experiment 3 by positing that participants who made SPIES in the first set of estimates had already scanned the entire 20th century range a few times, and had accessed evidence that was useful for estimating election years in the second set of estimates. However, had the two types of estimates not shared relevant evidence (e.g., had one set of estimates included election years of early 19th century Presidents), then confidence intervals would not have been any different after SPIES than before SPIES.

Alternatively, SPIES may influence the inference stage of the estimation process, and induce a change in how judges use their sampled knowledge to produce estimates. The process of the SPIES task—the systematic scan of the entire range of possible values and the production of multiple, mutually-exclusive probability judgments—may have a lasting effect on how judges estimate other, unrelated values. Previous research (Hirt, Kardes, & Markman, 2004) has shown that alternative-generation-based exercises can debias judgments across different knowledge domains. Therefore, the SPIES effect might not be dependent on the content retrieved during the initial estimate, and exist also in settings in which both SPIES and subsequent estimates share relevant information.

To test these two competing hypotheses, Experiment 4 varied whether the SPIES and confidence interval estimates share relevant content. All participants made two estimates, one using SPIES and the other using 90% confidence intervals. Half of the participants made both estimates on the same value population, whereas the other half made each estimate on a different population. If SPIES influence the inference stage of the estimation process, then we should expect 90% confidence intervals produced after SPIES to be better calibrated than those produced before SPIES in both content conditions. If SPIES work by increasing the amount of accessible information during the retrieval process and do not affect the inference process, then we should observe an effect of SPIES on subsequent confidence intervals only when both estimates are made on the same population.

Method

Participants. One-hundred sixty participants were invited via Amazon.com Mechanical Turk to complete a "Cognitive Survey" online. They were paid $0.10 each for their participation. There were five instances of multiple responses from the same IP address. The second record from each of these duplicates was stricken from the data file. The final sample consisted of 155 participants.

Procedure. Participants in the *same population* group received the following instructions:

> "In this experiment we will present you with a string of numbers. The string includes 100 numbers, which were sampled from a certain distribution. This means that some numbers can appear more than once and others may not appear at all.
>
> The string will appear on the screen for two seconds.
>
> After the presentation of the string of numbers, we will ask you two different questions about it."

Participants in the *different populations* group were given these instructions:

> "In this experiment we will present you with two strings of numbers. Each string includes 100 numbers, which were sampled from a different distribution each time. This means that some numbers can appear more than once and others may not appear at all.

> Each string will appear on the screen for two seconds.
>
> After the presentation of each string of numbers, we will ask you a question about it.”

Next, participants in the *SPIES-first* group then received the following instructions:

> “In the first question, we will present you with all number ranges that make up the string of numbers. Your job will be to estimate which percentage of the numbers in the string you saw is included in each range. Since these ranges include all numbers in the string, the sum of your estimated percentages must equal 100%.
>
> In the other question, we will ask you to estimate a number range that will include a certain percentage of the numbers you saw in the string. For example, if we ask you for a 50% range, then you will need to estimate a range that will include exactly 50 of the 100 numbers that were presented in the previous screen.”

Participants in the *SPIES-second* group received the instructions about each method in reverse order.

Before the presentation of each population, participants were told the minimum and maximum values of the value population’s range and that the population will be presented for 2 seconds. Participants in the *same population* group were presented with one value population, which included all 100 percentiles of a distribution (alpha = 2, beta = 2), ranging from 341 to 440 and made both estimates in sequence. Participants in the *different populations* group were first presented with a value population with the same distribution properties, but on a different range (731 – 830), and made their first estimate on this population. Then, they were presented with the same distribution as their *same population* counterparts, on which they made their second estimate (see Table 4).

Results

To measure the effect of SPIES on subsequent 90% confidence intervals, I conducted a 2 (method order: before SPIES / after SPIES) x 2 (populations: same population / different populations) between-subjects ANOVA on the inclusiveness of 90% confidence intervals, or the percentage of values from the population included in these intervals. The ANOVA revealed a significant effect of method order, $F(1,151) = 10.06$, $p = .002$, but no effect of population or an interaction, Fs < 1. Simple effects tests show that 90% confidence intervals produced after SPIES were significantly more inclusive, and were better calibrated, meaning the proportion of the population they included was significantly closer to 90%, than confidence intervals produced before SPIES, both in the *same population* condition, $t(79) = 2.19$, $p = .03$, and in the *different populations* condition, $t(72) = 2.28$, $p = .03$ (see Figure 8). This effect was not a simple order effect, whereby second estimates were more inclusive than first estimates, as 90% SPIES produced second did not include more values than those produced first. In fact, the opposite trend was observed, as 90% SPIES seemed to be more inclusive (albeit not significantly so) when produced before 90% confidence intervals than when produced after, both in the *same population* condition ($M_{First} = 88.23$, $SD = 18.56$; $M_{\text{Second}} = 83.29$, $SD = 23.48$; $t(79) = 1.06$, $p = .29$) and in the *different populations* condition ($M_{First} = 95.50$, $SD = 3.21$; $M_{\text{Second}} = 88.85$, $SD = 22.05$; $t(79) = 1.89$, $p = .06$). 90% SPIES that followed 90% confidence intervals were not affected by whether they were made on the same population or a different one, $t(69) = 1.02$, $p = .31$.

Discussion

Results of experiment 4 lead to two conclusions. First, they replicate the findings of Experiment 3 that 90% confidence intervals estimated after SPIES are significantly more inclusive than intervals produced without having previously completed a SPIES task. In this experiment, where the measurement of calibration at the individual estimate level was possible, individual 90% confidence intervals proved to be better calibrated with the 90% confidence standard when produced after SPIES than before SPIES.

In addition, these results demonstrate that the carryover effect of SPIES occurs even in cases where no relevant knowledge can be transferred between estimates. Confidence intervals produced after SPIES were improved, relative to those produced before, even when the two estimates were made on different value populations with no shared content. This suggests that the effect of SPIES is not limited to increased information retrieval, or on the amount of evidence sampled to form the interval estimate. Participants in the *different populations* could not use any content retrieved in the first estimate for making the second estimate, but they, too, achieved higher calibration of their 90% confidence intervals after SPIES than before SPIES. Thus, SPIES influenced the inference process of their confidence interval estimates. The next two experiments were designed to identify the nature of this influence.

EXPERIMENT 5

Experiment 4 demonstrated that the carryover effect of SPIES can occur even without the transfer of relevant content between estimates, suggesting that this effect goes beyond the evidence retrieval stage of the estimate and also affects the inference stage. As previously mentioned, there are two ways by which SPIES may affect the inference the judge makes from her sampled evidence. One is by generalizing the SPIES estimation process into the process of confidence interval production. This may include the presentation of more value categories than the judge would spontaneously generate, and the decomposition of the estimate into multiple judgments, both of which have been found to reduce biases in estimates (Erev, Shimonowitch, Schurr, & Hertwig, 2008; Soll & Klayman, 2004). By carrying out the process of systematically scanning the full range of possible values, the judge assesses various value categories and perceives more values as plausible. This realization then carries over to the confidence interval production process in subsequent estimates.

However, the process of considering all possible values may have a different effect on the judge: by scanning the entire range of values, the judge may encounter many possibilities about which she is unsure. Forcing the consideration of these values in the SPIES task may create a conflict between the judge's knowledge (or the stimuli presented to her) and the estimate elicitation format. This discrepancy may create doubt in the judge, who might revise her estimation process to account for this added uncertainty, or exert more cognitive control in later estimates. A number of studies report having reduced overestimation by creating similar discrepancies. Some studies (Hoch, 1985; Koriat, Lichtenstein, & Fischhoff, 1980; Sieck & Arkes, 2005) have asked participants to think of reasons for revising their estimates. Arkes, Christensen, Lai, and Blumer (1987) manipulated participants' expectations about the difficulty of a task, and found that when participants completed a task that was more difficult than they had expected, their confidence coming into a subsequent task decreased. Bloomfield, Libby, and Nelson (1999) simply made uninformed judges aware that their available information suffers from low statistical reliability, which helped in debiasing their choices. Similar effects may occur on overprecision as well.

Experiment 5 sought to distinguish between the two potential effects of SPIES on the estimation process of subsequent confidence intervals. Specifically, I wished to examine whether the effect occurs in settings where it is easier to generalize the systematic scan and multiple assessment procedures from the SPIES task to subsequent confidence intervals, or in settings that create a discrepancy between the product of evidence retrieval and the estimate production process. To that end, I varied the properties of the population on which participants made their initial estimates. All participants made initial estimates using SPIES that included the same range, but for half of the participants the value population was distributed narrowly, including mostly values placed around the middle of the range, whereas the other participants were presented with a very diverse population of values, more evenly distributed across all parts of the range. Then, all participants made a second estimate on a new value population (the same population for all groups).

A flat, even distribution of values in the population facilitates the retrieval of values from all parts of the range and incorporating them into the estimate. This makes the generalization of the SPIES process fairly easy, and may lead to an enhanced evidence retrieval process in later estimates as well. Conversely, when the distribution of the retrieved sample is very different from that of the intervals being assessed, such as when the sample is rather homogeneous and narrowly distributed but the SPIES task includes assessments of intervals across the entire range, a conflict may arise in the judge, which may prompt her to revise her estimation process, but will not encourage generalization of the same process to subsequent estimates.

Method

Participants. One-hundred sixty participants were invited through Amazon.com Mechanical Turk to participate in a “Cognitive Survey”. They completed the experiment online in exchange for $0.10 each. There were seven instances of multiple responses from the same IP address. The response from each of these duplicates was taken out of the data file. Results of eleven participants were not recorded on account of a technical error in the program.[8] The final sample, then, consisted of 142 participants.

Procedure

Similar to Experiment 4, all participants in this experiment made two estimates—an initial, practice estimate and a second, test estimate. Each estimate was made on a different value population, presented for 2 seconds. Instructions for both estimates were the same as in Experiment 4.

In a 2 x 2 orthogonal design, half of the participants used SPIES for their practice estimate whereas the other half used the 90% confidence interval method. For their test estimates, all participants produced 90% confidence intervals. Practice estimates for all participants were made on a value population ranging from 731 to 830. For half of the participants (the *narrow distribution* condition) the value population was narrowly distributed (alpha = 10, beta = 10), and included mostly similar values with many repetitions; for the other half (the *flat distribution* condition) the value population was distributed almost uniformly (alpha = 1.3, beta = 1.3), including values from across the entire range. Test estimates for all participants were made on the same value population used in Experiment 4 (ranging from 341 to 440, distribution properties: alpha = 2, beta = 2). The sequence of estimates for each group is illustrated in Table 5.

Results

Manipulation check. Participants were sensitive to the distributions of the value populations they saw in their practice estimates, and adjusted their estimates accordingly. The

[8] The error was in the naming of variables in the Flash program we used to run the experiment. It affected eleven of the first thirty participants in the experiment, and was corrected when we updated variable names in the program during data collection for ease of use. This change did not affect the experimental procedure, and the error was discovered only after data collection was complete.

distribution width manipulation did not significantly affect the number of values included in these estimates, either in the form of 90% confidence intervals (M_{Narrow} = 59.05, *SD* = 35.55; M_{Flat} = 67.09, *SD* = 29.06, *t*(70) = 1.046, *p* = .30) or 90% SPIES (M_{Narrow} = 88.84, *SD* = 26.07; M_{Flat} = 83.64, *SD* = 17.50, *t* < 1). Consistent with findings reported in Part I of this paper, 90% SPIES were significantly more inclusive than 90% confidence intervals in both the narrow distribution and the flat distribution conditions, *t*s ≥ 3.00, *p*s ≤ .004, displaying significant overprecision only in the flat distribution condition, *t*(39) = 2.27, p = .03, and not in the narrow distribution condition, *t* < 1.

Effect of SPIES on subsequent confidence intervals. A 2 (method used in practice estimate) x 2 (distribution width in practice estimate: narrow/wide) between-subjects ANOVA on the inclusiveness of test confidence interval estimates reveal a significant interaction, *F*(1,138) = 5.72, *p* = .02. Contrary to the generalization hypothesis, making SPIES on an evenly distributed value population did not result in improved confidence intervals in subsequent estimates; in fact, 90% confidence intervals after SPIES included a slightly lower proportion of the value population than those estimated after a confidence interval, but this result did not significantly deviate from chance, *t* < 1. Alternatively, when first presented with a narrow distribution, participants who had used SPIES for their practice estimates produced 90% confidence intervals that were significantly more inclusive and better calibrated than those who had used confidence intervals, *t*(66) = 2.84, *p* = .006, suggesting that the discrepancy between the distribution of the population and the structure of the task in the first estimate contributed to the improvement of confidence intervals in the second estimate (see Figure 9).

Discussion

The results presented here refute the hypothesis that judges generalize the SPIES process to estimates made in other formats. When the values presented were fairly evenly distributed across all parts of the range, participants were able to sample values from many parts of the range and use them in their SPIES, but this did not have any positive effect on their subsequent estimates. Conversely, in the *narrow distribution* condition, participants' retrieved samples included, primarily, values from the middle of the range, but they were asked in the SPIES task to estimate frequencies of values from across the entire range. This resulted in improvement of the calibration of subsequent estimates, which raises the possibility that confidence intervals were enhanced by the conflict between the stimuli and the SPIES task. This conflict may have instilled doubt in participants, leading them to produce wider confidence intervals to accommodate for this uncertainty, or caused them to exert more cognitive control in producing their subsequent estimate, in response to this conflict.

It is interesting to note that participants who produced confidence intervals in their practice estimates displayed the opposite pattern: practice estimates on a flat distribution of values led to wider confidence intervals in the test estimate than practice estimates on the narrow distribution, although this difference was non-significant, *t*(70) = 1.22, *p* = .23. It appears that in the confidence interval condition generalization could occur: a widespread population of values generates a wider confidence interval than a narrowly distributed one, leading the judge to produce wider confidence intervals in subsequent tasks as well. On the other hand, the confidence interval production task does not produce the doubt present in the SPIES task, because it does not force the judge to consider any values she did not voluntarily take into account, and this is no different for either distribution condition.

EXPERIMENT 6

Experiment 5 showed that the effect of SPIES on subsequent estimates does not occur when the practice estimate is made on a heterogeneous population that is equally represented by all intervals of the SPIES task. This suggests that the SPIES effect is not caused by the mere presentation of extreme possible values, even when these values are as plausible as ones which are closer to the mean of the distribution. Also, this experiment showed that successful sampling of multiple values across the entire range, which leads to the retrieval of a wide range of values, does not result in an improved estimation process in subsequent estimates. Rather, in order for SPIES to influence subsequent estimates, there needs to be some discrepancy between the retrieved sample and the assessments which the task elicits, which may lead the judge to revise her estimation process.

In Experiment 6, I sought to test whether the mere discrepancy between the value population and the structure of the SPIES task can improve subsequent estimates. If doubt is involved in creating the effect, such discrepancy may not always be enough. For example, if a current reader of this paper were to use SPIES to estimate his or her own age, this is unlikely to undermine the strength of the judge's belief in her knowledge of the true answer. Similarly, using SPIES to estimate the number of veal plates that were ordered at restaurants in Sligo, Ireland on January 25th, 1994, should not create much doubt or conflict, assuming the judge's knowledge in this domain is already approximating zero. In both of these cases, a discrepancy exists between the properties of the distribution of likelihoods and the estimation task, but differences in knowledge prevent a real conflict from taking place. Thus, when the judge's knowledge sufficiently high not to be fazed by the SPIES task, or sufficiently low that the SPIES task itself provides most of the information regarding the estimate, a change in the estimation process might not occur.

Experiment 6 was designed to test this proposition. In this experiment, participants made the same estimates that displayed the SPIES carryover effect in Experiment 5. However, their knowledge of the initial estimate's content matter was manipulated by varying the time for which participants saw the value populations before making their estimates. When participants could observe the population for a long time, their knowledge about it was high enough that the structure of the SPIES task should not cause much of a conflict. When the time allotted to observe the population was very short, participants' information about the population was low, such that the intervals making up the SPIES may have provided a source of information, rather than created a conflict with, or doubt in existing knowledge. However, when the time participants had to observe the population is neither too long nor too short, as may have been the case in Experiment 5, the SPIES task should conflict with participants' knowledge of the population's distribution, and may cause them to revise their estimation process.

Method

Participants. One-hundred forty participants were recruited through Amazon.com Mechanical Turk to participate in a "Cognitive Survey". They completed the experiment online in exchange for $0.10 each. There were four instances of multiple responses from the same IP address. The latter response of each of these duplicates was stricken from the data file. The final sample consisted of 136 participants.

Procedure. In a three-group design, all participants made a practice estimate on one value population using SPIES and then a test estimate, in a 90% confidence interval format, on a different value population. Participants were given the same instructions as in Experiments 4 and 5. The values for both the first and second estimates were sampled from the same ranges and distributions as in the narrow-range condition in Experiment 5: the practice estimate included numbers sampled from the 731 – 830 range, on a narrow beta distribution (alpha = 10, beta = 10), whereas the test estimate was made on numbers sampled from the 341 – 430 range, on a moderate (alpha = 2, beta = 2) distribution. The groups varied in their exposure time to the value population before making the practice estimate. One group (the *medium exposure* group) saw the numbers for 2 seconds, same as in Experiment 5, where SPIES led to a revision in subsequent confidence interval estimates. This group served as a control group, providing a test for the replication of the findings of Experiment 5. A second group (the *short exposure* group) was presented with the value population for a quarter of the time allotted to the control group—500 milliseconds—such that they could identify a very low number of values in the population, if any. For the third group (the *long exposure* group) the value population was presented for a time four times as long as the control group—8 seconds—such that participants could observe most, if not all, the values in the population. Therefore, their knowledge of the value population's properties should be high enough that completing a SPIES task should not cause them to doubt it.

After the practice estimate, all participants made a 90% confidence interval estimate on the same new population of values. Presentation times of the new estimate were also equal between the groups (2 seconds).

Results

Manipulation check. Table 6 shows that none of the groups displayed significant overprecision in their 90% SPIES, as might be expected of estimates of a population so narrowly distributed. In fact, the medium exposure group was even underconfident. However, one expected difference between these groups was that the more time one is given to sample the population, the better sense one should have about the distribution of the population, resulting in smaller errors in interval assessments. A one-way ANOVA comparing the three groups on average error, or mean difference between the subjective probability participants gave each interval in the SPIES task and the interval's true likelihood, revealed a significant difference: the longer the participant could observe the population, the smaller the participant's average error was, *F*(2,133) = 6.71, *p* = .002 (see Table 6). In addition, a longer exposure time should improve the judge's ability to distinguish between frequent and infrequent values, and produce a confidence interval that includes the most likely values and excludes very unlikely ones. This results in confidence intervals that are denser, meaning they include the same proportion of values from the population, but are narrower, in absolute terms, than the one produced by an less-informed judge. This density measure can be calculated by measuring the ratio between the inclusiveness of the confidence interval (the proportion of values from the population included in the estimate) and the width of the confidence interval, in absolute terms. A one-way ANOVA comparing the density of participants' 90% SPIES shows a significant between-group difference: the longer the participant could observe the population, the higher the density of their 90% SPIES, *F*(2,133) = 3.78, *p* = .02 (see Table 6).

Effect of SPIES on subsequent 90% confidence intervals. To measure the effect of stimulus exposure time in practice estimates on participants' test estimates, I conducted a one-way ANOVA to compare the inclusiveness of participants' confidence intervals between the three groups. Results reveal a significant difference, *F* (2,133) = 3.32, *p* = .04. Consistent with my prediction, while the calibration of confidence intervals produced by the medium exposure group replicated the results of Experiment 5, in the two other groups, where SPIES were assumed to not create doubt for participants, calibration of confidence intervals in the test estimates was significantly poorer (Short exposure: $t(93) = 2.39$, $p = .02$; Long exposure: $t(86) = 2.18$, $p = .03$, see Figure 10).

Discussion

Results of this experiment provide further support for the prediction that SPIES influence subsequent confidence interval estimates only when there is a discrepancy between the judge's knowledge and the structure of the task and that knowledge is neither too high nor too low to create a conflict with the task, or increase the judge's doubt. In the two conditions of this experiment where participants' knowledge of the population in the practice phase was very high or very low (as seen by the different levels of error and density of their estimates), their estimates in the test phase demonstrated lower calibration, at a degree similar to that of confidence intervals produced without a preceding practice phase (see results of Experiment 4).

The SPIES task in the short and long exposure conditions in this experiment included features which, in theory, could have contributed to the improvement of calibration of subsequent estimates. A very short exposure time should increase participants' uncertainty and reduce their confidence in their ability to make an accurate prediction (Erev, Wallsten, & Budescu, 1994; Peterson & Pitz, 1986, 1988). Conversely, a long exposure time in the first estimate may make subsequent estimates, for which exposure times are shorter, seem more difficult. This may create a contrast effect and reduce participants' confidence in their estimates (Arkes et al., 1987; Mussweiler, 2003). However, if this were true, then a difference in the opposite direction should have been observed in the *short exposure* condition, in which the test estimate was given a longer exposure time than the practice estimate. The fact that the two groups did not differ in their performance on the test estimate, and that both were outperformed by participants in the *medium exposure* condition, provides further support for the proposition that conflict and doubt are required for subsequent estimate improvement.

Despite a significant improvement in calibration, 90% confidence intervals produced by the medium exposure group were still overprecise: they included, on average, of 77.57% (*SD* = 21.84) of the population, which was significantly lower than the standard 90%, *t*(46) = 3.901, *p* < .0005. This incomplete improvement in calibration can be explained by prior findings of research on revision of opinion, that judges tend to be too conservative in the extent to which they update their beliefs in light of new information (Erev et al., 1994; Fischhoff & Beyth-Marom, 1983). According to these studies, judges take into account new information that is inconsistent with prior estimates, but underweight this information vis-à-vis their prior beliefs, relative to predictions of Bayes's theorem. A similar process may have taken place in this experiment, where participants were influenced by the SPIES task and adjusted their subsequent confidence intervals, but not enough, reflecting an overreliance on their prior estimation process.

SUMMARY

Data presented in Part II of this dissertation demonstrates that SPIES enact a change in the cognitive process judges use to produce interval estimates. In three experiments, as well as in Experiment 3 of Part I, participants who had initially used SPIES and then switched to confidence intervals produced intervals that were more inclusive than those produced by participants who had not practiced with SPIES.

The value population paradigm used in the last three experiments enabled the comparison of not only the width of confidence intervals, but also their calibration with the 90% confidence standard. This was done by having participants make estimates on a finite population of values, and eliciting estimates of intervals that include a proportion of the population (i.e., 90% of the values in the population). Using this method, we learned that the SPIES method not only increases the width of subsequent confidence intervals, it also improves their calibration.

Although the exact mechanism by which the confidence interval production process changes as a result of SPIES is not fully clear yet, this work provides some interesting insight about this mechanism. First, this change occurs, at least in part, in the inference stage of the process, and does not depend on the accessibility of specific content. As demonstrated in Experiment 4, confidence interval estimates improved after SPIES, even when the two estimates were made on two different populations with no shared values. Second, for this effect to take place, there must exist a conflict between the judge's knowledge of the distribution of possible values and the structure of the SPIES task. Judges did not display improvement in confidence intervals after SPIES when their sampled values corresponded equally with each of the SPIES task's intervals, but rather only when the SPIES task included estimates of value categories participants did not think they saw. Experiment 6 showed that the discrepancy between the distribution's properties and the SPIES task will create this conflict only when judges' knowledge about the distribution is low enough that the task provides some information about the distribution, and high enough that the SPIES task is not the only source of information.

Conclusion

Overconfidence is a widespread and widely studied phenomenon. While it has a number of distinct forms, overprecision seems to be the most consistent, potentially the most harmful, and possibly the least understood (Moore & Healy, 2008). Despite decades of research on overprecision, few studies have found ways to decrease it, and these attempts mostly resulted in modest improvements, limited to specific content and elicitation format.

One common point stressed by most of these studies is that overprecision stems from judges' failure to consider all alternative outcomes and relevant information, and that reducing the bias can be made possible by making judges pay attention to a wider range of evidence and possible answers. This dissertation presents a novel method which achieves just that. The SPIES method forces judges to consider all possible outcome categories and estimate the relative likelihood of occurrence of each of these categories. It presents the judge with the entire range of possible outcomes, and elicits probability judgments of occurrence of all of them. By transforming these probability judgments into an interval estimate, the SPIES method produces confidence intervals with consistently better calibration than intervals produced by other methods. This improved calibration is achieved by increasing the inclusiveness of these intervals, that is, making them span a wider range of values, rather than by improving the accuracy of their midpoints, as these were not, on average, closer to the true answer than the midpoints of traditional confidence intervals. This suggests that the SPIES method does not teach the judge anything new about the content matter of the estimate, but rather causes her to realize that more outcomes are possible, and prevents her from overlooking them.

Further, SPIES seem to enact a change in judges' estimation process, as demonstrated by the improved calibration of confidence intervals produced by judges who used SPIES in prior estimates. To our knowledge, SPIES is the first method to have achieved a lasting effect on estimates produced in different elicitation formats. Similar to interval estimates produced by the SPIES method, the improved calibration appears to have been achieved by increasing the inclusiveness of the confidence intervals. Further support for this argument is provided by the fact that confidence interval calibration is improved after SPIES even when the two estimates are made on different knowledge domains. Data reported in Part II suggest that after practicing with SPIES, judges do not simply generalize this estimation process to subsequent judgments, but rather revise their processes after some conflict between the evidence they sampled and the structure of the SPIES task, which may cause them to doubt their current knowledge and adjust their estimation process to account for this uncertainty. Accordingly, when conditions are such that the knowledge-task conflict is low, subsequent confidence intervals are not affected. These conflict-curbing conditions occur when evidence across the entire range is easily accessible, or when the judge’s knowledge of the estimate is too high for SPIES to cause doubt or too low that uncertainty is already at a maximum.

In addition to its ability to reduce bias, I believe SPIES can be an effective method thanks to its applicability. Relative to other elicitation methods, it is simple to use, from the judge’s point of view. Probability judgments are more intuitively comprehensible than value estimates (Teigen & Jørgensen, 2005) and while judges do provide multiple judgments in the SPIES task, these judgments are elicited as one general estimate. This method provides improved calibration without making the judge engage in meta-cognitive exercises or make multiple estimates of the same value. Although the SPIES method requires the designer of the task to specify the range on which the tasks’ intervals are placed, determining this range is not critical for achieving improved calibration, relative to the confidence interval production method.

The SPIES elicitation method also provides flexibility in calculating confidence intervals. The amount of information elicited from the judge enables the calculation of confidence intervals of different confidence levels and different widths, all from the same estimate, without making the judge estimate the same value again. Take, for example, an estimate of the future price of a house. A decision maker who receives this estimate can determine, for example, the likelihood that the price will be above or below a certain value, the most likely $20,000 range for the house, the judge's 90% confidence interval, and the 70% confidence interval, all from the same estimate.

LIMITATIONS

The research on SPIES is hardly complete and several open questions remain. One is the applicability of the method to estimates that are not interval-based. Many estimates and forecasts attempt to predict a true outcome from a series of outcomes that are not placed on a scale. Can SPIES improve the calibration of categorical estimates as well? Prior research has shown that presenting participants with more possible outcomes can reduce base-rate neglect (Erev et al., 2008) and improve calibration of estimates of categorical outcomes (Fischhoff et al., 1978; Tversky & Koehler, 1994), giving reason to suspect that SPIES might have a similar effect. However, the studies presented here tested SPIES only on quantitative, interval-based estimates, and therefore cannot confirm this argument.

The studies in Part II on the carryover effect of SPIES have, admittedly, not entirely clarified the nature of the effect of SPIES on subsequent estimates. Experiments 5 and 6 showed that when conditions are such that when doubt in the judge's knowledge of the estimate is unlikely to increase, subsequent confidence intervals will not be affected. However, these two studies did not manipulate or measure doubt directly. Also, if doubt is driving the improvement in subsequent estimates, then this improvement could be achieved by other doubt-increasing methods, some simpler and more robust in generating doubt than SPIES. These questions should be investigated further.

The SPIES method should be also tested for carryover effects on different types of quantitative estimates, such as probability judgments and contingency estimates. These estimates have been shown to suffer from overprecision, and there is reason to believe that practicing with SPIES might lead to their improvement as well. One caveat is that calibration in these estimates is measured differently than the way it is observed in confidence intervals. However, the processes by which SPIES may improve confidence interval production should have a similar influence on the process of producing probability and contingency estimates as well.

CAN SPIES MAKE A DIFFERENCE FOR PRACTITIONERS?

The introduction of this dissertation tells the story of the demise of one of the largest financial institution in the United States, in part due to overprecision in judgment. At a conference at which this project was presented, one of the attendants suggested that a more common use of methods such as SPIES could have prevented the latest financial crisis from happening. While this last statement is probably too extreme, there is reason to believe using SPIES can reap benefits for professionals. In many fields, estimates are an integral part of the job for executives and policy makers, and accurate forecasts are a key determinant of their success, just as surprises and failure to prepare for certain outcomes lead to very harmful consequences. The SPIES method minimizes the failure to consider possible outcomes, and therefore could be helpful for practitioners who rely on accurate forecasting in their jobs.

There is, however, one major obstacle to the implementation of SPIES in the field. Most quantitative forecasts and estimates today use the point prediction format. The output of these estimates is a precise value. For example, future price forecasts provide a precise price, or a price change value; product managers who predict the demand for their product in the next quarter estimate a precise value for this demand. The use of this format has a number of inherent problems. One is that the criterion for accuracy is ambiguous. Exact accuracy of these predictions is, obviously, virtually impossible. Therefore, recipients of these estimates expect them to be approximately accurate, but the cutoff point beyond which an estimate is regarded as inaccurate is rarely specified. This, in turn, prevents forecasters from receiving meaningful and reliable feedback on their performance and may lead to communication problems between forecasters and recipients with regard to their performance expectations. Another problem is the inability to prepare for various scenarios. Take, for example, a forecast of the quarterly revenue a certain division is expected to generate. In order for the division to remain profitable for its company, it must generate at least $1 million in revenues in the next quarter. The quarterly revenue forecast, in a point prediction format, is $1.1 million. However, the level of uncertainty surrounding this estimate may make revenues between $800,000 and $1 million as likely, if not more than revenues between $1 million and $1.2 million. Knowing this fact may change the perceived risk associated with keeping the division, and thus also the company's strategic decisions about it, but the point prediction format does not provide this information. Similarly, medical patients rehabilitating from surgery are given prognoses for the time it should take them to fully recover. However, for a patient for whom the most likely recovery time is six months, a six week upward error might have very different implications than a six week downward error. Only an estimate in a range format can provide this important information.

The first step, then, on the way to improving forecasting in the field, is to move from a point prediction format to a range format. Once range estimates are accepted as the default estimate format, then methods for improving these estimates could be implemented. In this dissertation, I propose that the SPIES method, which my advisors and I have developed, can provide a simple, intuitive way to significantly improve range estimates.

References

Abbas, A. E., Budescu, D. V., Yu, H. T., & Haggerty, R. (2008). A Comparison of Two Probability Encoding Methods: Fixed Probability vs. Fixed Variable Values. *Decision Analysis*, *5*(4), 190-202. doi:10.1287/deca.1080.0126

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting--with and without satisfying knowledge. *Journal of experimental psychology. Learning, memory, and cognition*, *34*(5), 1224-1245. doi:10.1037/a0012938

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., & Yurak, T. J. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, *68*(5), 804-825. doi:10.1037//0022-3514.68.5.804

Alpert, M., & Raiffa, H. (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: Cambridge University Press.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, *39*(1), 133–144.

Bearden, J., Gaba, A., Jain, K., & Mukherjee, K. (2011). Unpacking the Future: A Nudge Toward Wider Subjective Confidence Intervals. *INSEAD Working Paper No. 2011/61/DS.*

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, *49*(2), 188-207. doi:10.1016/0749-5978(91)90048-X

Bloomfield, R., Libby, R., & Nelson, M. W. (1999). Confidence and the welfare of less-informed investors. *Accounting, Organizations and Society*, *24*(8), 623-647. doi:10.1016/S0361-3682(99)00025-2

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624-52.

Cesarini, D., Sandewall, O., & Johannesson, M. (2006). Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior & Organization*, *61*(3), 453–470.

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended Thought* (pp. 212-252). New York: Guilford Press.

Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians‘ use of probabilistic information in real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 928-935. doi:10.1037/0096-1523.7.4.928

Clemen, R. T. (2001). Assessing 10-50-90s: a surprise. *Decision Analysis*, *20*(1).

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. *The Journal of Finance*, *53*(6), 1839-1885. doi:10.1111/0022-1082.00077

Ditto, P. H., & Lopez, D. F. (1992). Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568-584.

Epley, N., & Gilovich, T. (2004). Are Adjustments Insufficient ? *Personality and Social Psychology Bulletin*, *30*(4), 447-460. doi:10.1177/0146167203261889

Erev, I., Shimonowitch, D., Schurr, A., & Hertwig, R. (2008). Base rates: how to make the intuitive mind appreciate or neglect them. In H. Plessner, C. Betsch, & H. Betsch (Eds.), *Intuition in Judgment and Decision Making* (Vol. 26, pp. 135-148). New York: Taylor & Francis.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527.

Federal Housing Finance Agency. (2011). Quarterly average and median prices for states and U.S.: 2000Q1 - Present. Retrieved September 19, 2011, from http://www.fhfa.gov/Default.aspx?Page=87

Fiedler, K., & Juslin, P. (2006). Taking the interface between mind and evironment seriously. In K. Fiedler & P. Juslin (Eds.), *Information Sampling and Adaptive Cognition2* (pp. 3-32). New York: Cambridge University Press.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*(3), 239.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 330–344.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*(3), 195–200.

Galinsky, A. D., Moskowitz, G. B., & Skurnik, I. (2000). Counterfactuals as self-generated primes: The effect of prior counterfactual activation on person perception judgments. *Social Cognition*, *18*(3), 252–280.

Glaser, M., & Weber, M. (2007). Overconfidence and trading volume. *The Geneva Risk and Insurance Review*, *32*(1), 1-36. doi:10.1007/s10713-007-0003-3

Hansson, P., Juslin, P., & Winman, A. (2008). The role of short-term memory capacity and task experience for overconfidence in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1027-1042. doi:10.1037/a0012638

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making, 5*(7), 467-476.

Hilary, G., & Hsu, C. (2011). Endogenous Overconfidence in Managerial Forecasts. *Journal of Accounting and Economics*. doi:10.1016/j.jacceco.2011.01.002

Hirt, E. R., & Markman, K. D. (1995). Multiple Explanation: A Consider-an-Alternative Strategy for Debiasing Judgments. *Journal of Personality and Social Psychology*, *69*(6), 1069-1086. doi:doi:10.1037/0022-3514.69.6.1069

Hirt, E. R., Kardes, F. R., & Markman, K. D. (2004). Activating a mental simulation mind-set through generation of alternatives: Implications for debiasing in related and unrelated domains. *Journal of Experimental Social Psychology, 40*(3), 374-383. doi:10.1016/j.jesp.2003.07.009

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 719. American Psychological Association. doi:10.1037//0278-7393.11.1-4.719

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. doi:10.1037/0278-7393.25.4.1038

Juslin, P., Winman, A., & Hansson, P. (2007). The Naıve Intuitive Statistician: A Naıve Sampling Model of Intuitive Confidence Intervals. *Psychological Review*, *114*(3), 678-703. doi:10.1037/0033-295X.114.3.678

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General*, *131*(2), 287-297.

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. doi:10.1006/obhd.1999.2847

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

McKenzie, C. R. M. (1997). Underweighting Alternatives and Overconfidence. *Organizational Behavior and Human Decision Processes*, *71*(2), 141-160. doi:10.1006/obhd.1997.2716

McKenzie, C. R. M. (1998). Taking Into Account the Strength of an Alternative Hypothesis. *Cognition*, *24*(3), 771-792. doi:10.1037/0278-7393.24.3.771

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

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502-17. doi:10.1037/0033-295X.115.2.502

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in cognitive sciences*, *14*(10), 435-440. doi:10.1016/j.tics.2010.07.004

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

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*(3), 472-489. doi:10.1037/0033-295X.110.3.472

Odean, T. (1999). Do investors trade too much? *The American Economic Review*, *89*(5), 1279-1298. doi:10.2139/ssrn.94143

Oskamp, S. (1965). Overconfidence in Case-Study Judgments. *Journal of consulting psychology*, *29*(3), 261-5. doi:10.1037/h0022125

Peterson, D. K., & Pitz, G. F. (1986). Effects of amount of information on predictions of uncertain quantities. *Acta psychologica*, *61*(3), 229–231.

Peterson, D. K., & Pitz, G. F. (1988). Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 85.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Psychology* (19th ed., pp. 123-205). New York: Academic Press.

Pleskac, T. J., Dougherty, M. R., Rivadeneira, A. W., & Wallsten, T. S. (2009). Random error in judgment: The contribution of encoding and retrieval processes. *Journal of Memory and Language*, *60*(1), 165-179. doi:10.1016/j.jml.2008.08.003

Rakow, T., Harvey, N., & Finer, S. (2003). Improving calibration without training: the role of task information. *Applied Cognitive Psychology*, *17*(4), 419–441.

Seaver, D. A., von Winterfeldt, D., & Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior & Human Performance*, *21*(3), 379-391. doi:10.1016/0030-5073(78)90061-2

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making*, *18*(1), 29-53. doi:10.1002/bdm.486

Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: choosing between intuitive and nonintuitive alternatives. *Journal of experimental psychology: General*, *135*(3), 409-28. doi:10.1037/0096-3445.135.3.409

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

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., & Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, *30*(3), 512-23. doi:10.1111/j.1539-6924.2009.01337.x

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, *47*(2), 143-148. doi:10.1016/0001-6918(81)90005-6

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. doi:10.1002/acp.1085

Tiedens, L. Z., & Linton, S. (2001). Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing. *Journal of personality and social psychology*, *81*(6), 973-88.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124-31. doi:10.1126/science.185.4157.1124

Tversky, A., & Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547-567. doi:10.1037/0033-295X.101.4.547

Weary, G., & Jacobson, J. A. (1997). Causal uncertainty beliefs and diagnostic information seeking. *Journal of Personality and Social Psychology*, *73*(4), 839-848. doi:10.1037/0022-3514.73.4.839

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective Probability Intervals: How to Reduce Overconfidence by Interval Evaluation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*(6), 1167-1175. doi:10.1037/0278-7393.30.6.1167

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*(4), 424-432.

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, *10*(1), 21-32. doi:10.1002/1099-0771

Yaniv, I., & Schul, Y. (1997). Elimination and Inclusion Procedures in Judgment. *Journal of Behavioral Decision Making*, *10*(3), 211-220.

Zickfeld, K., Morgan, M. G., Frame, D. J., & Keith, D. W. (2010). Expert judgments about transient climate response to alternative future trajectories of radiative forcing. *Proceedings of the National Academy of Sciences*, 1-6. doi:10.1073/pnas.0908906107

Tables

Table 1. Range Width and Grain Size Condition Assignment in Experiment 2

Range Width	Number of Intervals	Grain Size	Extreme Intervals
Narrow (100ºF)	20 + 2 extreme intervals	Fine (5ºF)	-16°F or lower 85°F or higher
Narrow (100ºF)	10 + 2 extreme intervals	Medium (10ºF)	-16°F or lower 85°F or higher
Wide (200ºF)	20 + 2 extreme intervals	Medium (10ºF)	-66°F or lower 135°F or higher
Wide (200ºF)	10 + 2 extreme intervals	Coarse (20ºF)	-66°F or lower 135°F or higher

Table 2. 90% SPIES hit-rates by range width and grain size in Experiment 2

		Range Width	
		Narrow	Wide
Number of	10	80.95% (40.24%)	83.33% (38.07%)
Intervals	20	61.90% (49.76%)	69.23% (47.07%)

Table 3. 90% SPIES width by range width and grain size in Experiment 2

		Range Width (SD)	
		Narrow	Wide
Number of	10	33.40 (16.58)	44.95 (11.80)
Intervals	20	25.48 (11.12)	33.50 (12.93)

Table 4. The order of estimates made by the four groups in Experiment 4.

Group	First estimate population range (method)	Second estimate population range (method)
Same population, SPIES-first	341 – 440 (SPIES)	341 – 440 (90% confidence interval)
Same population, SPIES-second	341 – 440 (90% confidence interval)	341 – 440 (SPIES)
Different populations, SPIES-first	731 – 830 (SPIES)	341 – 440 (90% confidence interval)
Different populations, SPIES-second	731 – 830 (90% confidence interval)	341 – 440 (SPIES)

Table 5. The type of distribution of the estimated value population and elicitation method of each estimate, by group, in Experiment 5.

Practice Estimate		Test Estimate	
Distribution	Elicitation method	Distribution	Elicitation method
Narrow	SPIES	Moderate	90% confidence interval
Narrow	90% confidence interval		
Flat	SPIES		
Flat	90% confidence interval		

Table 6. Absolute width, inclusiveness and density of 90% SPIES by group in Experiment 6. Standard deviations are in parentheses.

	90% SPIES absolute width	90% SPIES inclusiveness	SPIES average error	90% SPIES density
Short exposure	73.77 (23.23)	87.42 (27.61)	11.15 (3.22)	1.17 (0.44)
Medium exposure	83.66 (10.61)	99.47 (1.23)	10.10 (2.08)	1.21 (0.21)
Long exposure	68.37 (19.64)	90.44 (18.94)	8.49 (3.43)	1.38 (0.45)

Figures

Figure 1. Illustration of hypothetical estimates using SPIES and a 90% confidence interval for the daily high temperature in Washington, DC, one month in the future. The 90% SPIES ranges from 15ºF to 54ºF, whereas a 90% confidence interval ranges from 25ºF to 40ºF.

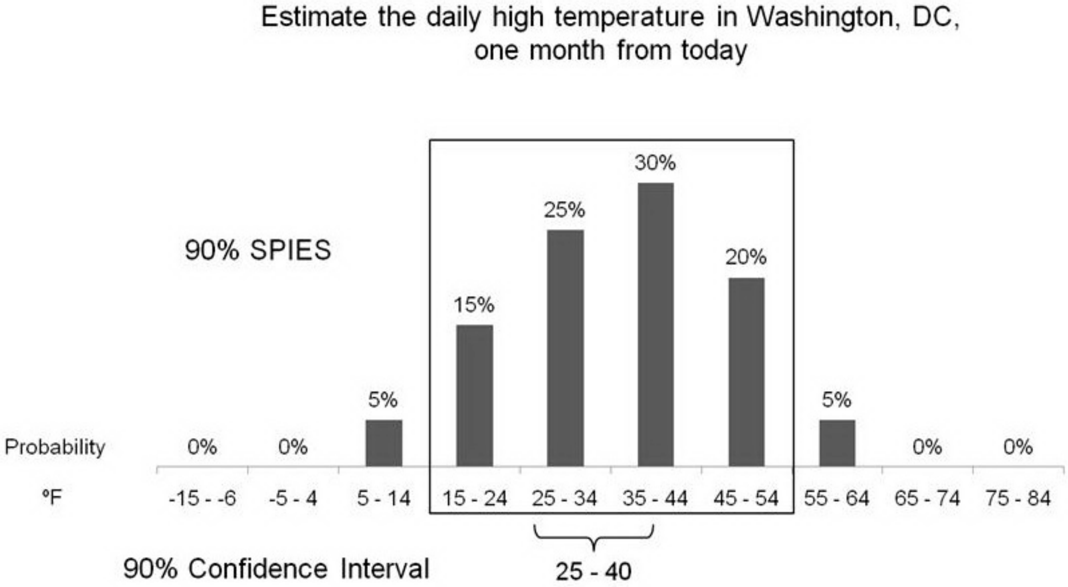


Figure 2. Hit-rates displayed by 90% confidence intervals, fractiles and 90% SPIES in Experiment 1. Error bars indicate ±1 *SE*.

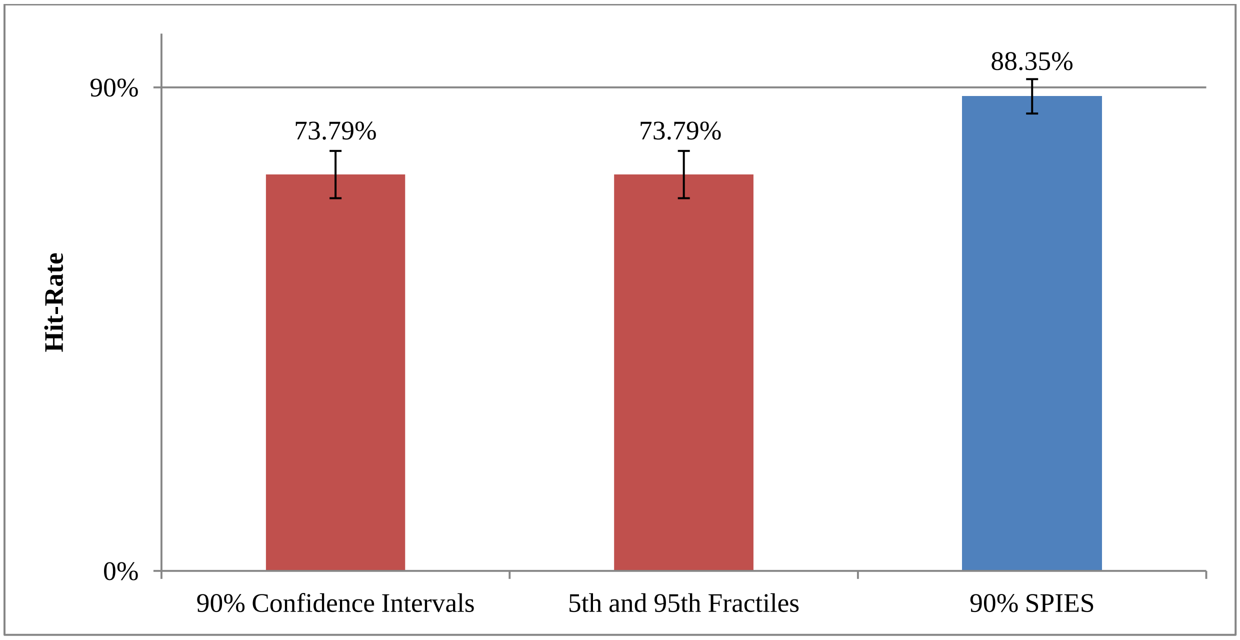


Figure 3. Hit-rates displayed by SPIES of different range widths and grain sizes and by 90% confidence intervals in Experiment 2. Error bars indicate ±1 *SE*. See Table 2 for hit-rates of the different SPIES configurations.

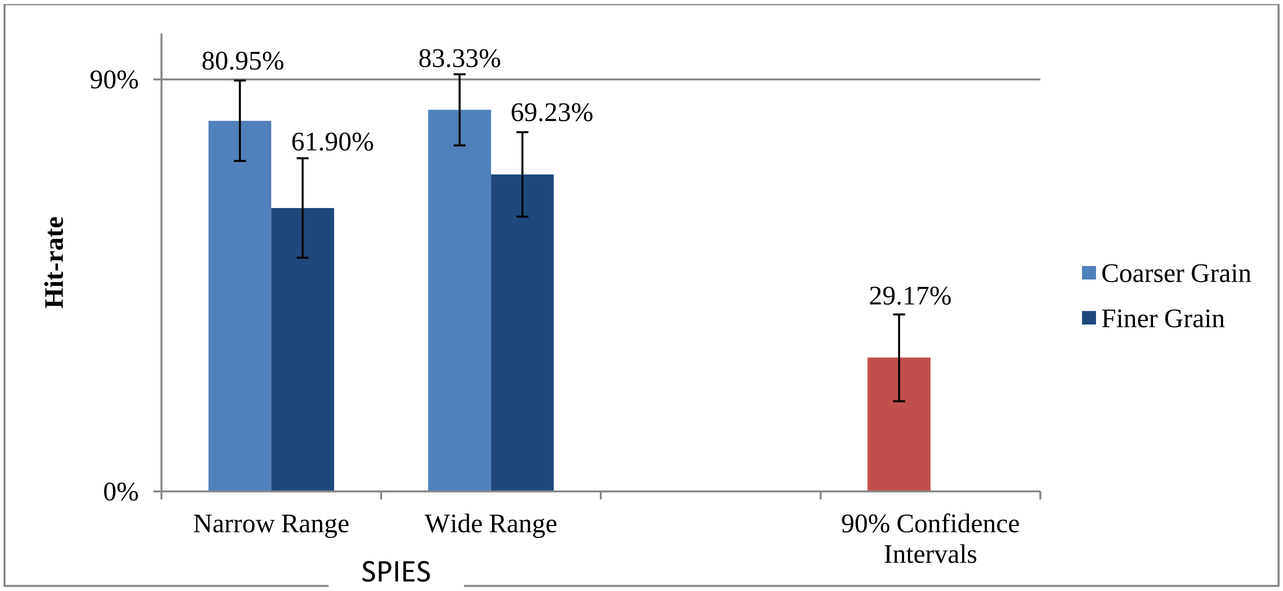


Figure 4. Estimate-by-estimate mean widths of 90% confidence intervals made in the first set of estimates (before SPIES), compared to those of 90% confidence intervals in the second set of estimates, after having made SPIES in the first set in Experiment 3.

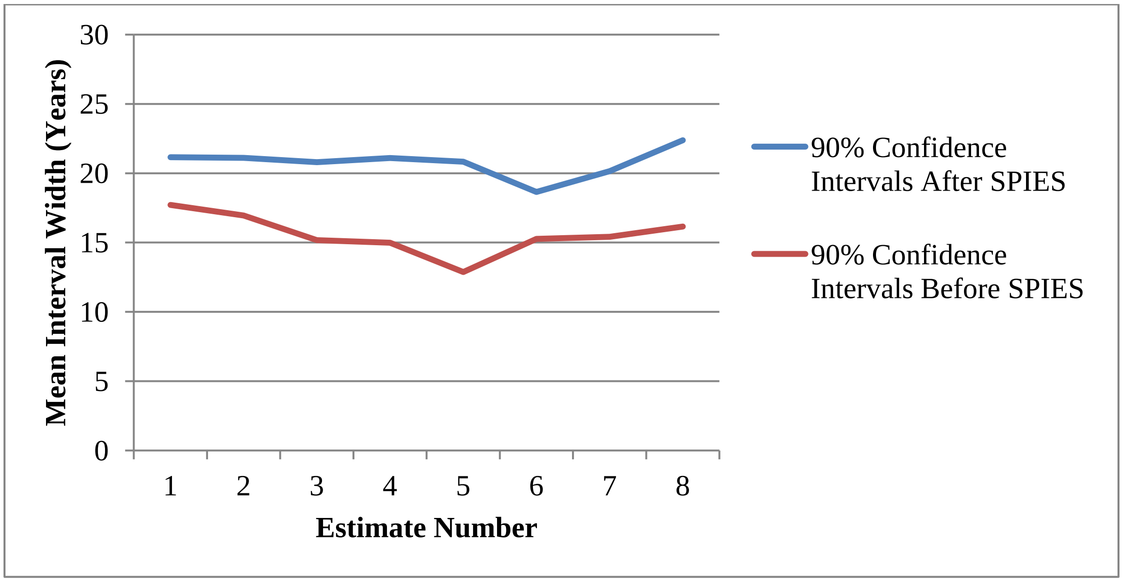


Figure 5. A hypothetical subjective probability density function for a time estimate. Intervals J and K represent opposite 10-year errors in estimating the interval I, which is perfectly calibrated with the requested probability of the estimate. This graph appears in Soll & Klayman, (2004).

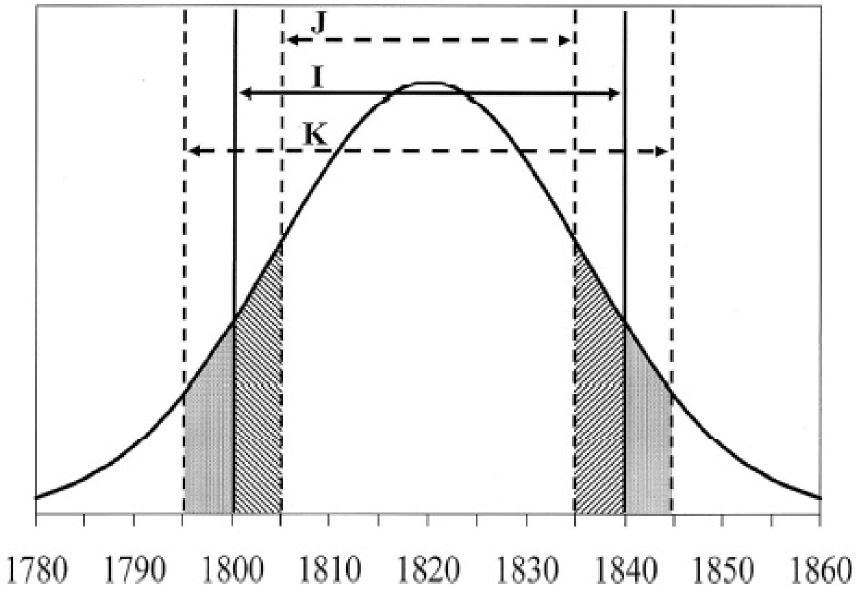
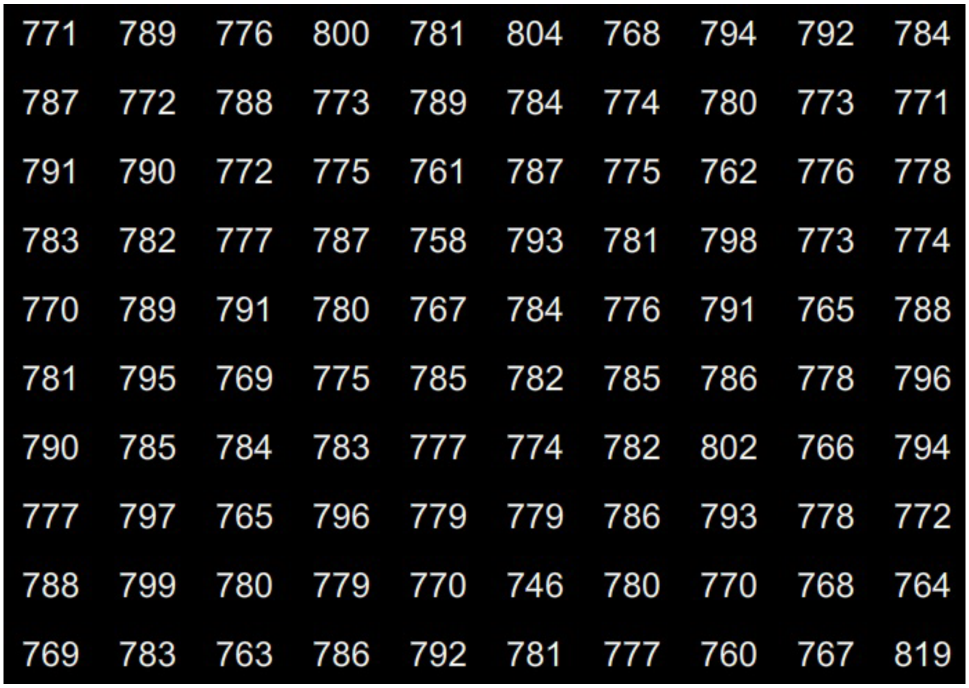
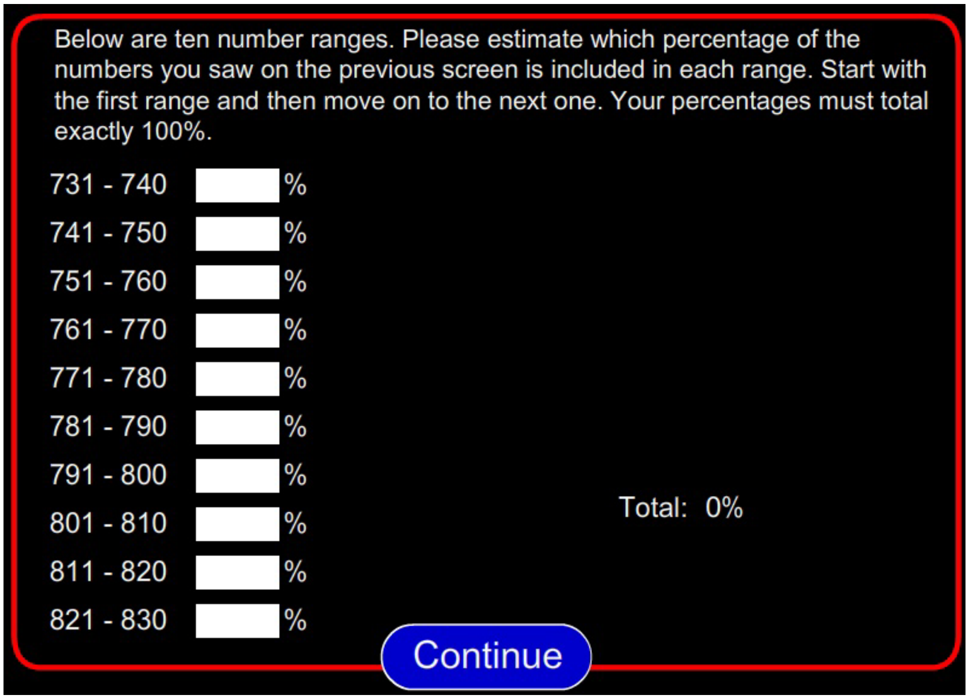


Figure 6. A value population, as presented in Experiments 4, 5, and 6. The order of the values in the population was randomized between participants in all experiments.



771	789	776	800	781	804	768	794	792	784
787	772	788	773	789	784	774	780	773	771
791	790	772	775	761	787	775	762	776	778
783	782	777	787	758	793	781	798	773	774
770	789	791	780	767	784	776	791	765	788
781	795	769	775	785	782	785	786	778	796
790	785	784	783	777	774	782	802	766	794
777	797	765	796	779	779	786	793	778	772
788	799	780	779	770	746	780	770	768	764
769	783	763	786	792	781	777	760	767	819

Figure 7. A SPIES task for estimating a value population, as presented in Experiments 4, 5, and 6.



Below are ten number ranges. Please estimate which percentage of the numbers you saw on the previous screen is included in each range. Start with the first range and then move on to the next one. Your percentages must total exactly 100%.

731 - 740 %

741 - 750 %

751 - 760 %

761 - 770 %

771 - 780 %

781 - 790 %

791 - 800 %

801 - 810 %

811 - 820 %

821 - 830 %

Total: 0%

Figure 8. Percentage of values from the estimated population included in participants' 90% confidence intervals, by value population similarity and method order in Experiment 4. Error bars represent ±1 SEM.

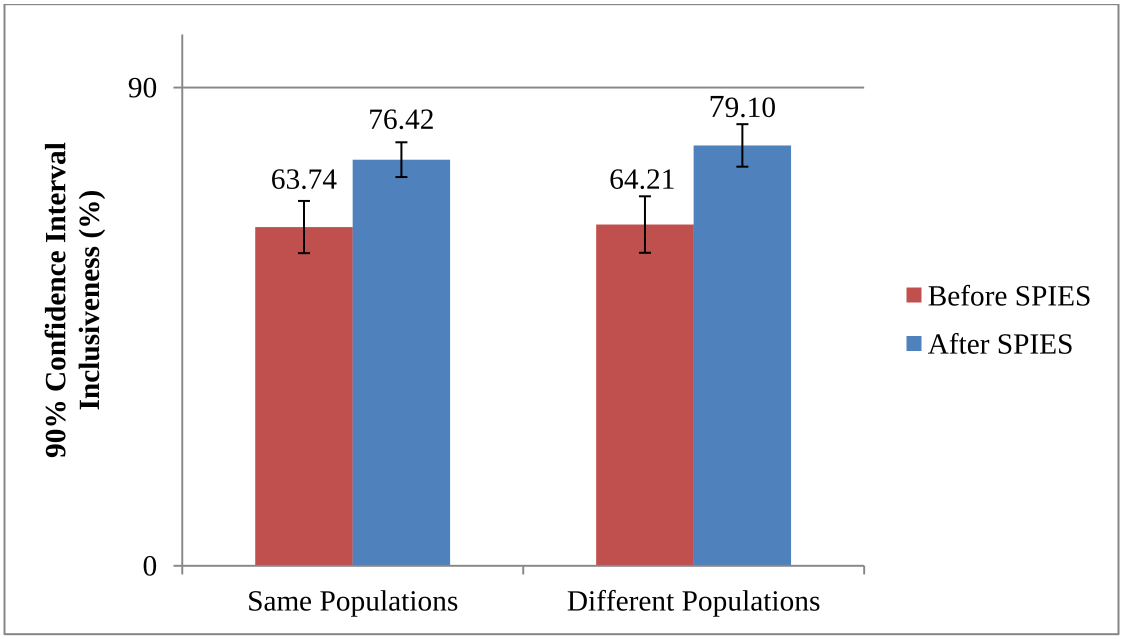


Figure 9. Inclusiveness of 90% confidence intervals in test estimates, by distribution and elicitation method used in practice estimates in Experiment 5. Error bars represent ±1 SEM.

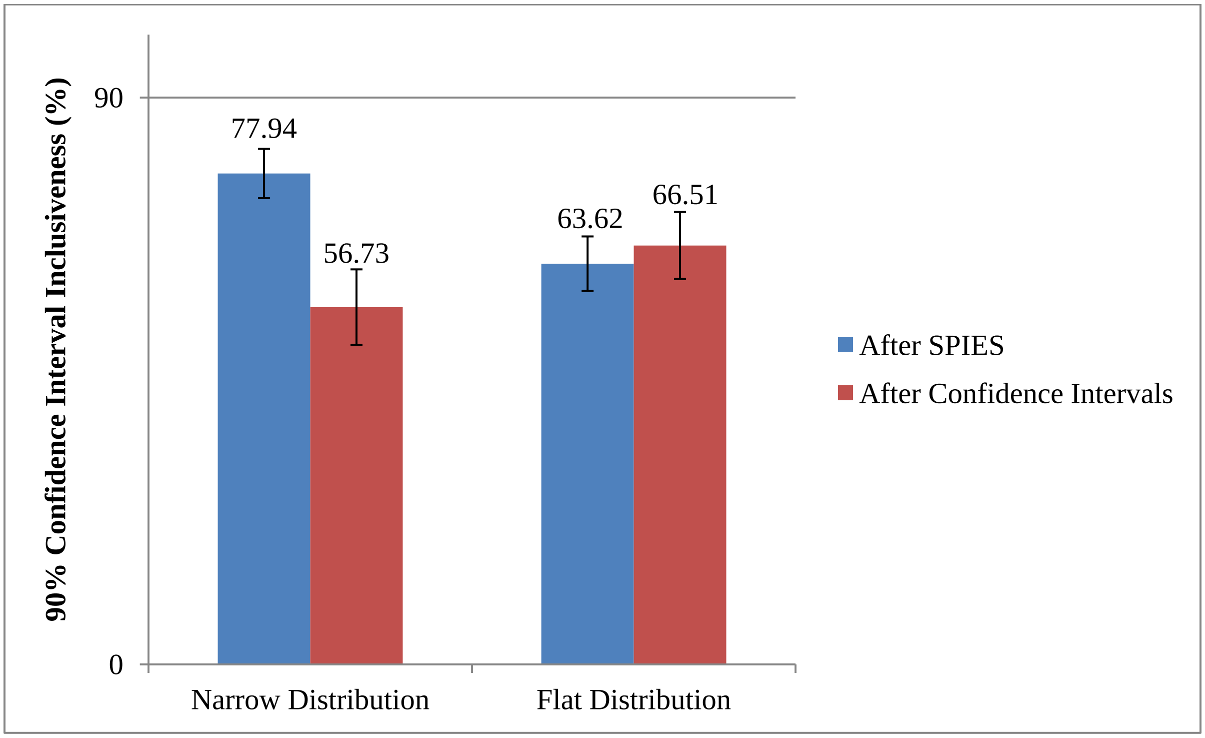
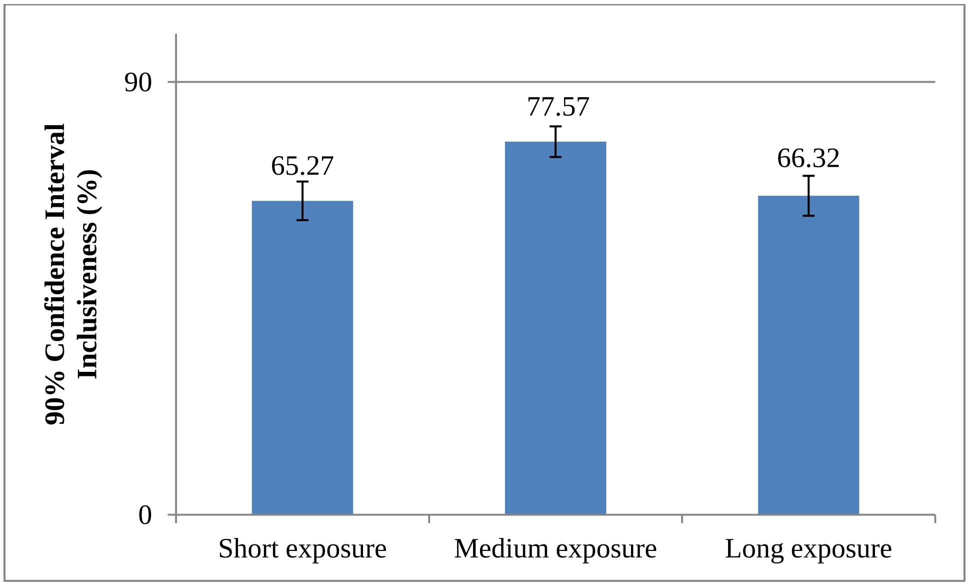


Figure 10. Inclusiveness of 90% confidence intervals by group in Experiment 6. Error bars represent ±1 SEM.


90% Confidence Interval Inclusiveness (%)
Short exposure
Medium exposure
Long exposure