Master Thesis
M.Sc. in Economics and Business
Specialisation in Behavioural Economics
Erasmus School of Economics
Erasmus University Rotterdam

When are experts wiser than the crowd?

The existence of expertise for simple and complex questions

17 October 2017

Author: Gust van der Meeren
Supervisor: Prof. Dr. Aurélien Baillon

Abstract

This paper replicates the study of Prelec, Seung, and McCoy (2013) and extents it with a more complex question. Their expert (LST) answer is based on how well individuals predict other people's answers. In their study, it outperforms the wisdom of the crowd (group) answer on questions about US state capitals. In mine, the LST and group answer perform similarly. Besides, for a more complex question (based on a bean-jar experiment), the LST answer underperforms the group answer. This corresponds to literature that finds less expertise with more complexity. Subsequently, I identify experts with another method to control for expertise and test whether the performance difference is the result of the LST answer's failure to properly identify experts (based on experts' ability to predict other people's answers). This is not the case. In fact, experts seem more able to predict other people's answers for the more complex questions. I confirm this finding by directly comparing the predictions of subjects I identified as experts and non-experts.

Keywords

Wisdom of the crowd · Information · Expertise · Predicting · Experiment

JEL classifications

C53 · C83 · C9 · D8

"A public-opinion poll is no substitute for thought."

— Warren E. Buffett[1]

1. Introduction

The idea of the wisdom of crowds, i.e. that the average of all individuals' answers produces a group answer that outperforms any individual answer, has become well-known (Surowiecki 2005; Sunstein 2006). Examples abound, from bean-jar experiments to online customer ratings, as well as political and economic applications through democracy or financial and betting markets. Nevertheless, crowds are not always wise, in which case an expert answer is sometimes preferred. This poses a difficult problem: how to identify experts? Confidence does not signal expertise, since both the correct and incorrect are similarly confident (Slovic, Fischhoff and Lichtenstein 1985). Past performance is a better indicator of expertise, but prior data are not always available (Lock 1987).

Prelec, Seung, and McCoy (2013) propose to identify experts as those who are least surprised by the truth. It identifies them by asking a second question; individuals' predictions of other people's answers to the first question. Once all questions are answered, the answers to the first question provide the correct answer to the second question. Those with the best answers to the second question are considered Least Surprised by the Truth and hence are expected to also have the best answer to the first question (which is hereafter called the LST answer). Prelec, Seung, and McCoy (2013) test this theory and find that the LST answer strongly outperforms the group answer for "True"-or-"False" questions about a statement that says the capital of a US state is its largest city.

I replicate their study and examine if the LST answer also outperforms the group answer for another kind of questions. The relationship between accurate answers and accurate predictions of others' answers contrasts to literature that finds experts are unable to predict non-expert performance (Hinds 1999; Cho 2004). The LST answer's success may therefore very well be particular to the topographical statements of the original study. It is thus interesting to see how well the LST answer performs with a more complex question. In turn, Shanteau (1992; 2002) finds that expertise decreases with complexity, which he attributes to characteristics such as the dynamism of stimuli, the repetitiveness of the task, and the availability of feedback. The US state-capital questions are simple according to these characteristics. The more complex questions are based on the bean-jar experiment, where subjects estimate the number of beans in a jar, which is common in literature that tests the wisdom of crowds (Treynor 1987; Krause, et al. 2011). A

[1] This quote is from Buffett and Clark (2006).

comparison of the group and LST answer in the two kinds of questions can thus indicate the latter's wider applicability.

I collect data from 53 subjects who each answer 24 pairs of questions, 12 about US state capitals and 12 about the number of beans in a jar. The result show that the LST answer hardly outperforms the group answer for the US state-capital questions. For the bean-jar questions, the LST answer even underperforms the group answer. Based on these data, I conclude two things. First, the performance of the LST answer is not necessarily better than the group answer, unlike Prelec, Seung, and McCoy's (2013) findings. And, second, the performance of the LST answer is worse for more complex questions. I next analyse what causes this second conclusion.

Since the LST answer relies on both the existence of expertise and experts' ability to predict other people's answers, it is not immediately clear to which of the two any difference in performance should be attributed. To address this issue, I also compute an expert answer that only relies on the existence of expertise. It is developed by Budescu and Chen (2015) and based on individuals' contributions to the group answer through a Contribution-Weighted Model (and it is hereafter called the CWM answer). If the LST answer performs worse for the more complex question due to less-existent expertise, the CWM answer should also show a worse performance for these questions. My computations show it does. Thus, at least some of the LST answer's underperformance on the more complex questions can be attributed to less-existent expertise.

It nevertheless remains possible that experts' ability to predict other people's answers also differs between the two kinds of questions, which would solely affect LST answer's performance. I therefore directly compare the LST and CWM answer for both kinds of questions. The latter outperforms the former for the original questions, but performs similarly for the more complex questions. This suggests experts' ability to predict other people's answers increases with complexity. I test this suggestion by comparing the predictions of other people's answers by the CWM experts, the CWM non-experts, and the group as a whole. It confirms the earlier suggestion: whereas predictions for the original questions do not differ, the CWM experts outperform the non-experts and group in predicting other people's answers to the more complex questions. While expertise thus decreases with more complexity, experts' ability to predict other people's knowledge actually increases.

This study has its limitation. Most importantly, the sample size, which is small with only 24 questions and just 12 of each kind. Mine and the original study's sample are also rather different, which is especially important with regards to the topography questions. Furthermore, the experiment only employs two kinds of questions (about US state capitals and bean jars) with the particular framing of "True"-or-"False" statements, which limits broad generalisations. These limitations pose several new lines of enquiry. Larger sample sizes will increase statistical power.

More variety in the questions is necessary to draw broader conclusions on the group and LST answers' efficacy over different kinds of questions. As is similarly the case for the variety in samples. Finally, the determinants of experts' (in)ability to predict other people's answers also deserve further research.

This paper proceeds as follows. The literature review (Section 2) elaborates on the applicability of the wisdom of the crowd (Subsection 2.1) and the possibility to base answers on experts within crowds (2.2). In the methods section (3), I present the research questions and the hypotheses (3.1) and I explain the methodology (3.2) and experimental design (3.3). The results (4) of the experiment show the comparisons between the group and LST answer (4.1) and between these and the CWM method's (4.2) answers (4.2.1) and (non-)experts' predictions of other people's answers. In my concluding remarks (5), I discuss the findings (5.1), the limitations and possibilities for further research (5.2), and the conclusions (5.3). The appendices provide detailed information about the questions' subjects (I) and formulation (II), and show additional and more-detailed results (III).

2. Literature review

2.1.The wisdom of the crowd: evidence, theory, and limitations

2.1.1. Evidence of the wisdom of the crowd: experiments and real-life cases

Following Galton's (1907) ox-weighing exercise in an English country fair, there have been many more examples of the wisdom of crowds in both experiments and in real-life.[2] There are several famous examples that illustrate the wisdom of crowds in experimental settings. Knight (1921) asks students in her class to estimate the room temperature (Lorge, et al. 1958; Surowiecki 2005). Estimates range from 60 to 85 degrees Fahrenheit, but the group's average answer (72.4 degrees) approximates the true room temperature (72 degrees). The group outperforms eighty percent of the individuals, while twenty percent of individuals performs similarly or better. Treynor (1987) presents his students with bean jars and asks them to estimate the number of beans in the jar. In two experiments, the groups' average answers (841 and 871) approximates the true values (810 and 850). The first experiment involves 46 students and only two outperform the group (or 4.3% of the sample) and in the second experiment this is the case for only one out of 56 students (1.8%).

[2] For these examples, I draw heavily on a number of sources, to whom I also refer for an overview: Lorge et al. (1958), Surowiecki (2005), and Sunstein (2006).

Krause et al. (2011) also perform a bean-jar experiment, but with marbles instead of beans.[3] Two groups of visitors to a science exhibition produce average answers of 516 and 554 for a jar with 562 marbles.

There are also more elaborate questions where the wisdom of the crowd is shown to work. Gordon (1924) asks 200 students to rank ten glass bottles that are the same in everything but their weights, which ranges from 16 to 17.6 grams in equal increments. Whereas individuals' ranks have a correlation with the true order of +0.41, in groups of fifty individuals, average ranks have a correlation of +0.94 with the true order. This accuracy is matched by only 5 individuals (representing 2.5% of the sample) and better than the remaining 195 individuals (97.5%). Bruce (1935) repeats Gordon's (1924) experiment with slightly different weights (ranging from 20 to 21.8 grams in ten equal increments) and adds a visual experiment. The latter shows ten groups of small balls on a cardboard, which look similar but differ in number, ranging from 51 to 60 balls in one ball increments. Both experiments include the same 120 students, which rank the glass bottles according to weight and the groups of small balls according to size. Their individual ranks have average correlations with the true order of +0.50 and +0.82, respectively. In groups of sixty individuals, the average ranks have correlations of +0.88 and +0.95 with the true order. Their accuracy is matched by only 12 individuals (10%) in case of the glass bottle weights and 34 individuals (28%) in case of the ball group sizes.

Real-life also knows many applications of the wisdom of crowds. Although there is less opportunity to check the accuracy of the crowd's wisdom, application of it is nevertheless widespread. The TV show *Who Want to Be a Millionaire?* provides an interesting example of the wisdom of crowds. Participants in the show face multiple choice questions on a wide variety of topics (and win a prize based on their performance). When they find a question particularly difficult, they can call upon assistance from, among others, a personal contact appointed by themselves beforehand as an "expert" or the audience by polling their preferred answers. It turns out that the experts are correct 65 percent of the time, while the groups preferred answer is correct 91 percent of the time (Surowiecki 2005). Armstrong (2001) looks at the previous literature and collects thirty earlier comparisons between individual and group answers of experts to forecasting questions. He finds that expert groups make 12.5 percent less errors than the same experts individually in forecasts on subjects as diverse as the survival of patients, gross national products, a company's earnings, and livestock prices.

[3] The authors are biologists, who refer to the wisdom of the crowd as "swarm intelligence."

Other real-life examples of the wisdom of crowds are found in financial and betting markets.[4] The former aggregate investors' opinions on financial instruments, such as stocks and bonds, to determine the market price of said instruments. An interesting case involves the crash of the *Challenger* space shuttle in 1986.[5] Just after it launched on January 28, it blew up and investors sold off the four most important contractors of the endeavour. At the end of that day, one contractor's stock was down by nearly 12 percent, while the other three's stocks were down only 3 percent. The hardest hit firm was Morton Thiokol, which built the O-ring seals on the booster rockets that were responsible for the explosion. But this only came to light six months after the stock market "decided" Thiokol was responsible. Maloney and Mulherin (2003) investigate how the stock market came to this decision, but they do not find any indication that some investors were "in the know" and hence moved the stock price. Although this is only anecdotal evidence, the aggregation of individual investors' decisions seems another example of the wisdom of crowds.

Betting markets aggregate gamblers' opinions on all sorts of events, such as sporting matches and political contests, to determine betting odds to gamble against. Researching the accuracy of betting markets is easy, because it is clear whether the relevant events materialise.[6] For example, Hoerl and Fallin (1974) study all horseraces run at the Aqueduct and Belmont Park tracks in 1970. They find that over all 1,825 races, the favourite horses won most often, the least favourite horses lost most often, and the same relations between betting predictions and race results also held for all other positions and horses. Results of betting markets for political elections are also strong. The Iowa Political Markets, which are small-scale, real-money future markets for election outcomes, are, on average, considerable accurate and outperform polls (Forsythe, et al. 1992; Berg, et al. 2008).

There are many more examples where the wisdom of a crowd is called upon, but which lack verification. The Copenhagen Consensus asked experts to rank a number of ways to promote

[4] There are two caveats to interpreting accurate financial and betting markets as indicators of wise crowds. First, these markets are not democratic in the strictest sense, because more wealthy traders can trade more and hence have more weight. Nevertheless, with a sufficient number of traders and dispersion of wealth, they are a good approximation of democracy. Second, most researches attribute the accuracy of financial and betting markets to smart "marginal traders" rather than the wisdom of the crowd. They suggest a small group of smarter traders corrects prices from the incorrect tendencies of a stupid crowd. Whether or not "marginal traders" exist, they are often constrained by limits to arbitrage, which means market prices remain dependent on the crowd. For example, Maloney and Mulherin (2003) do not find better-informed traders in a case-study of the stock market and Forsythe et al. (1992) do find more-able traders in political betting markets, but these are unable to fully correct prices. See for a discussion on the existence of "marginal traders" and the wisdom of crowds Surowiecki (2005). See for an overview of the limits to arbitrage Barberis and Thaler (2003).

[5] A similarly interesting but contrasting case involves the crash of the *Columbia* space shuttle in 2003, where the stock market wrongly "decided" Alliant Techsystems was responsible for the explosion (Surowiecki 2005).

[6] See for an overview of studies on the wisdom of crowds in betting markets Sauer (1998).

global welfare on their effectiveness, which were subsequently averaged for an overall ranking of projects (Lomborg 2004). The *Rotten Tomatoes* website rates each film with its "Tomatometer," which measures the percentage of professional critics that wrote a positive review. Other websites aggregate consumer reviews on a wide variety of things, such as films (*IMDb*), books (*Goodreads*), local businesses (*Yelp*), and travels (*TripAdvisor*). Finally, the democratic political model itself relies on the wisdom of crowds. If the electorate is generally wise, this is an important argument for democratic rule (Grofman and Feld 1988; Goodin 2005).

2.1.2. Theoretical basis for the wisdom of the crowd: the Condorcet Jury Theorem

The theoretical basis for the wisdom of the crowd dates from well before most of the above examples. The Condorcet Jury Theorem sets out conditions for individuals to produce a correct group answer, i.e. when the average of a groups' individual answers is correct (Condorcet 1785).[7] Suppose a group of people is asked a binary question with one correct and one incorrect answer. When group members are more likely to answer correctly than incorrectly, the average answer of the group becomes increasingly correct with more members. That is, if the probability of an individual answering correctly is over 50 percent, the probability of the group's average answer being correct approaches 100 percent as the size of the group increases. Under these conditions, groups do better than individuals and larger groups do better than smaller groups.

Sunstein shows how this works in his book *Infotopia: How Many Minds Produce Knowledge* (2006). Suppose an individual gives the correct answer 67 percent of the time (and the incorrect answer 33 percent of the time). Then a three-person group produces the correct average answer 74 percent of the time. In case an individual gives the correct answer 80 percent of the time, a ten-person group produces the correct average answer nearly 100 percent of the time. The Condorcet Jury Theorem works the other way around as well. When group members are less likely to answer correctly than incorrectly, the average answer of the group becomes increasingly incorrect as the size of the group increases. Under these conditions, individuals do better than groups and smaller groups do better than larger groups.

The above also applies to groups with different compositions, i.e. with two (or more) "sorts" of people with their own likelihoods of answering correctly. Sunstein (2006), again, gives several examples where the group answer becomes increasingly correct as the size of the group

[7] It is useful to clarify what I mean by a *group* answer, since terminology differs across the literature. My *group* answer is the average answer of the group members' individual answers. As such, the group answer is produced by an aggregation of individuals rather than a group. This distinction is important, because a group answer that is produced by a group in the latter sense, with, for instance, the possibility of interaction between group members, is fundamentally different from the *group* answer I refer to here. My definition corresponds to the *statisticised group* answer of Lorge et al. (1958), the *statistical group* answer of Sunstein (2006), and the *democratic* or *majority* answer of Prelec, Seung, and McCoy (2013).

increases. For instance, when 60 percent of a group answers correctly 51 percent of the time and 40 percent answers correctly 50 percent of the time; when 55 percent of a group answers correctly 65 percent of the time and 45 percent answers correctly 40 percent of the time; or, when 51 percent of a group answers correctly 51 percent of the time and 49 percent answers correctly 50 percent of the time. All these groups produce a group answer which' likelihood of being correct approaches 100 percent as the size of the groups increase. Possible group compositions that exhibit this characteristic abound, as long as the different "sorts" of people together have an average probability of answering correctly of more than 50 percent.[8]

One group composition is of considerable practical interest since it consists of a group with both "experts" and "laymen." Consider a group with two sorts of people: one is more likely to answer correctly and one is as likely to answer correctly as incorrectly, i.e. they answer randomly. The former are "experts" because they give better than random answers and the latter are "laymen" because they give random answers. Because the laymen divide their answers equally between the alternatives, experts can move the group answer towards the correct alternative. Dependent on the ratio of experts to laymen, it requires a smaller or larger group to approach high certainties of a correct group answer. Furthermore, the degree of expertise, i.e. how likely experts are to answer correctly, also affects the required group size.

The Condorcet Jury Theorem relies on three assumptions (Condorcet 1785; Sunstein 2006). Firstly, an individual's answer is not affected by whether their answer is decisive. Secondly, an individual's answer is not affected by other people's answers. Thirdly, an individual's answer is not statistically related to other people's answers. The first and second assumption are easily met. If an individual does not know other people's answers, they are not affected by these answers or the decisiveness of their own answer. The third assumption is harder to meet. People with similar work or education, for instance, probably approach questions similarly and are thus not independent of each other. But only statistical independence, not causal independence, is required for this assumption, which is a more realistic threshold (Estlund 2009).

2.1.3. When is the crowd unwise? Convention, bias, and worse-than-random errors

Crowds are not always wise and sometimes in fact produce worse answers. As touched upon above, the Condorcet Jury Theorem also works the other way around (Sunstein 2006). Under the wrong conditions a larger group produces an increasingly incorrect answer and individuals are in

[8] The Condorcet Jury Theorem is not only applicable to binary questions, but similarly applies to multiple choice and open questions. See for an elaborate explanation and examples Surowiecki (2005) and Sunstein (2006).

in fact more likely to answers correctly. Three mechanisms affect individuals' answers such that crowds become unwise: conventional wisdom, biases, and worse-than-random answers.

Firstly, when the conventional wisdom is incorrect, it leads to unwise crowds (Sunstein 2006). Consider the binary question again, with a correct and incorrect answer. If conventional wisdom favours the incorrect answer, which is often the case (Henrich, et al. 2001), the probability of an individual answering correctly is less than 50 percent. Consequently, the probability of a correct group answer approaches 0 percent as the size of the group increases. The same holds for more elaborate group compositions, with different "sorts" of people and likelihoods of answering correctly. As long as their average probability of answering correctly is less than 50 percent, groups have increasingly incorrect group answers as they increase in size.

Secondly, when individuals are systematically biased they combine to unwise crowds (Sunstein 2006). Individuals that are biased produce skewed answers. Everybody has biases and when a bias is not systematic, it is not a problem for the wisdom of a crowd. But when a bias is systematic, i.e. individuals are biased in a similar manner, a group of such individuals produce a skewed group answer, in line with the systematic bias. An example of a systematic bias is "anchoring." People's susceptibility to anchors, i.e. their tendency to use a given starting point for their judgement, is well documented (Chapman and Johnson 2002). For instance, recall how Galton (1907) asked country fair visitors for an estimate of an ox's weight. If he had added the following example to the question: "Your answer could, for instance, be 950 pounds,"[9] it would have functioned as an anchor. Estimations would subsequently have been biased towards this number. Because the anchor is lower than the true weight of the ox, individuals would have underestimated the ox's weight and the group answer would have been too low (and less accurate) because of the anchor. Anchoring also occurs with anchors that are irrelevant to the question (Chapman and Johnson 2002). For instance, if Galton had added the following to the question: "You receive a price of 950 pound if you are correct," it would also have functioned as an anchor, with a similar result as with the exemplary weight.

Thirdly, when individuals answer randomly or worse they combine to unwise crowds (Sunstein 2006). For very difficult or complex questions, people may lack any useful knowledge and, as such, their group answer is as uninformative as their random individual answers.[10] Recall that the random answers of such a group of unknowledgeable laymen can be overwritten by only a few experts. Experts themselves, in turn, can be overwritten by laymen if the latter's answers are worse than random. In theory, laymen know they do not know the correct answer and

[9] N.B. the group answer was 1,197 pounds and the actual weight of the ox was 1,198 pounds.
[10] For example, Sunstein (2006) asks a number of law faculty colleagues on their estimation of the weight of the rocket fuel used for a space shuttle. The group answer is 56 million pounds and the median answer is 200,000 pounds, while the true weight is in fact 4 million pounds.

consequently answer randomly. But sometimes they may think they know more than they in fact do and answer worse than random. The group answer may thus be incorrect "if confusion and ignorance are so widespread that individuals' answers are worse than random." (Sunstein 2006)

2.2.The wisdom of experts: evidence, domain-dependency, and identification

An alternative to the wisdom of crowds is the identification of experts. It follows a long-standing debate in the literature between the aggregation of a group answer and the identification of an expert answers (Rowe 1992). Larreche and Moinpour (1983) suggest group answers only substitutes expert answers because of an inability to identify experts, which implies expert answers are inherently better than group answers.[11] Although this may be true, it is limited by two essential issues: the existence of expertise and the identification of experts. The former is not always the case and the latter is not always possible.

2.2.1. Mixed evidence on the wisdom of experts

The existence of expertise is widely studied by assessing the performance of those defined as experts in a specific field. But the evidence is mixed, with studies finding considerable, limited, or no expertise at all. Many of the earliest studies do find expertise, albeit limited.[12] Hughes (1917) and Wallace (1923) find that agricultural judges do not always give the highest ratings to the highest yielding crops. Trumbo et al. (1962) find wheat judges misgrade around one-third of the wheat and give around one-third of the wheat a different grade when they grade it for a second time. Shanteau and Gaeth (1981) find soil judges differ from laboratory results more often than not and only half the repeated judgements are the same. Similar results are found for other sorts of judges: e.g. psychologists (Einhorn 1974; Oskamp 1962), physicians (Hoffman, Slovic and Rorer 1968), parole officers (Carrol and Payne 1976), and court judges (Ebbesen and Konecni 1975).

Later studies do find considerable expertise in a number of fields. An important example is the game of chess. Expert chess players perceive playing patterns more effectively (Groot 1965) and remember positions better (Chase en Simon 1973). Kundel and Nodine (1975) flashed X-ray pictures in front of experienced radiologists for one-fifth of a second and found them able to determine abnormalities with 70 percent accuracy. Furthermore, Christensen et al. (1981) find

[11] Following this line of argument, one ideally identifies the individual with the most expertise and uses their answer rather than aggregate the answers of multiple experts.

[12] For these examples, I draw heavily on a number of sources, to whom I also refer for an overview: Chi, Glaser, and Farr (1988), Mayer (1991), Shanteau (1992), and Surowiecki (2005). (Shanteau (1992) actually divides the proponents and opponents of the existence of expertise along the lines of decision-making and cognitive science research, respectively, from which he draws his examples.)

novice radiologists need more time to assess X-ray pictures than expert radiologists. Numerous other studies also find experts in fields, such as, physics (Larkin, et al. 1980), cardiology (Patel and Groen 1986), and computer programming (Mayer 1991).

Others argue against the existence of any expertise, especially in more complex fields. In a survey of expert forecasts on various subjects, Armstrong (1980) concludes: "I could find no studies that showed an important advantage of expertise." Similarly, Surowiecki writes in his book *The Wisdom of Crowds* (2005):

> "[…] there's no real evidence that one can become expert in something as broad as 'decision making' or 'policy' or 'strategy.' Auto repair, piloting, skiing, perhaps even management: these skills that yield to application, hard work, and native talent. But forecasting an uncertain future and deciding the best course of action in the face of that future are much less likely to do so."

More concrete examples abound. In investing, Bogle (2001) finds that around 90 percent of mutual-fund managers underperformed the Wilshire 5000 Index, which is a relatively low benchmark, in the period from 1984 to 1999. Other tasks also show that experts do not do better than laymen, such as evaluating applicants for medical internship (Johnson 1988) and predicting graduate school success (Dawes 1971). Expertise thus seems to exist for some questions and not for others.

2.2.2. The domain-dependency of expertise: when does expertise exist?

In order to resort to expertise, it is important to know whether it exists at all. The mixed evidence provides a strong reason to believe expertise's existence is dependent on the domain of the question. In line with this theory, Shanteau (1992; 2002) proposes a distinction between different domains with varying levels of expert performance. Whether experts perform well or poorly is dependent on a domain's characteristics (Shanteau 1992). On the one hand, there are domains with good expert performance, which are characterised by static stimuli, decisions about things (rather than behaviour), and repetitive tasks. Examples of such domains are weather forecasts, test piloting, and accountancy. On the other hand, there are domains with poor expert performance, which are characterised by dynamic stimuli, decisions about behaviour (rather than things), and unique tasks. Examples of such domains are clinical psychology, court judgements, and intelligence analysis. Shanteau (2002) divides several examples among four levels of expert performance (from very good to poor expert performance). Table 2.1 gives an overview of the expert performance domain characteristics and examples of four expert performance levels.

TABLE 2.1 Expert performance by domain characteristics and examples of the categories

Domain characteristics†		Examples‡			
Expert performance level					
Good	Poor	Very good	Good	Normal	Poor
Static stimuli	Dynamic stimuli	Weather forecasters	Chess masters	Clinical psychologists	Polygraphers
Things	Behaviour	Astronomers	Livestock judges	Parole officers	Managers
Repetitive task	Unique task	Test pilots	Grain inspectors	Psychiatrists	Stock forecasters
Feedback available	Feedback unavailable	Insurance analysists	Photo interpreters	Students admissions	Parole officers
Objective analysis	Subjective analysis	Physicists	Soil judges	Intelligence analysts	Court judges

† The domain characteristics come from Shanteau (1992) and are a selection.
‡ The categories examples come from Shanteau (2002). He also supplies three examples with varied evidence for attribution to the aided decisions, competent, and restricted category: i.e. nurses, physicians, and auditors.

2.2.3. How to identify experts?

When experts exist, one still has to identify them to use their expertise. It is important to know the experts and their degree of expertise. Subsequently, answers of individuals with more expertise can receive a higher weight to create an expert answer (rather than a group answer based on equal weighting). Although there is no definitive standard for the identification of expertise, it is traditionally based on things like experience, certification, peer reviews, consistency, consensus, behavioural characteristics, or knowledge tests (Shanteau, Weis, et al. 2002). These methods all have serious limitations and their application is often complicated and rarely feasible (Shanteau, Weis, et al. 2002). But there are other methods to identify expertise, which are almost as simple and generally applicable as the wisdom of the crowd. Such methods' expert answers are therefore more comparable to group answers.

The two most common methods use either self-rated confidence or past performance (Rowe 1992). The former method lets each individual rate their own confidence and uses this as an indicator of expertise, weighting individuals' answers accordingly. A problem with this method is people's overconfidence makes people who are incorrect practically as confident as those who are correct (Slovic, Fischhoff and Lichtenstein 1985).[13] Another method looks at past performance as an indicator of expertise, also weighting individuals' answers accordingly. This method is only possible with adequate prior performance data and it ignores learning effects (Lock

[13] People's overconfidence in their own judgement is widely documented and extents to experts (Cooke 1991; Shanteau 1992). Studies show overconfident experts in fields such as wheat judgement (Trumbo, et al. 1962), clinical psychology (Oskamp 1965), physics, and music (Glenberg and Esptein 1987). See for an overview of examples, for instance, Odean (1998).

1987). Furthermore, numerous studies compare the wisdom of the crowd to these two methods, but find little advantage in the methods with differentiated weights (Rowe 1992). Rose, Gustafson, and Ludke (1974) find no statistical difference between the wisdom of the crowd and the confidence-weighted method. As do Flores and White (1989), although they, as well as Ashton and Ashton (1985), find that the performance-based method does outperform the other two methods slightly.

A more recently developed method also uses past performance to identify experts, but it looks at contributed performance rather than absolute performance (Budescu and Chen 2015). That is, it looks at how much an individual's answers contribute to correct group answers and not how many of an individual's answers are correct. Individuals' contributions are based on the difference between group answers' performance with and without the individual.[14] The intuition behind this is that judgements are often highly correlated (Broomell and Budescu 2009) and correct answer in such circumstances should not count as much towards expert identification as those in which an individual's give correct contrarian answers. Budescu and Chen (2015) find that this Contribution-Weighted Model (CWM) answer moderately outperforms the group answer, which, in turn, moderately outperforms an absolute performance-based answer, in forecasts of current general and economic events. The outperformance is attributed to both accurate identification of experts (for 60%) and accurate weighting of expertise (40%). Even though this method's results are more promising than those based on past performance, it still requires adequate prior performance data, which is not always available.

2.2.4. A new method to identify experts

Another recent method does not need prior performance data, but uses a follow-up question that asks individuals to predict how others answer the initial question (Prelec, Seung and McCoy 2013).[15] This method assumes that those who answer this second question more accurately are better informed about the first question as well. Accurate prediction in the second question thus identifies expertise on the first question. Prelec, Seung, and McCoy (2013) therefore define experts as those who are "Least Surprised by the Truth" and refer to theirs as the LST method.[16]

The intuition behind this method is best explained by an example. First, consider the question of whether Philadelphia is the capital of Pennsylvania. This is a trick question.

[14] See for the numerical definition Subsection 3.2.3 on the computation of the CWM answer.
[15] This question is also used to induce honest responding through what is called the Bayesian Truth Serum (Prelec 2004) as shown in experiments measuring brand recognition (Weaver and Prelec 2013) and questionable scientific practices (John, Loewenstein and Prelec 2012).
[16] See for the numerical definition Subsection 3.2.2 on the computation of the LST answer.

Philadelphia is the largest city of Pennsylvania; however, it is not its capital.[17] Because of this feature, those who are correct and incorrect in the first question have different predictions of what others answer in the second question. Those who are incorrect, think theirs is the obvious answer and predict most other people answer similarly. Those who are correct, know it is a trick question and predict a considerable amount of other people know this as well. Thus, the latter people answer the first question correctly and are therefore also better at predicting what other people answer in the second question. The authors reverse this relationship and base their answer on the assumption that it is likely that those who are more accurate in the second question are also more accurate in the first question. In Prelec, Seung, and McCoy's (2013) own words: "[…] prediction of others' answers could provide a signal that is strong enough to override majority opinion."

They test their LST method in a series of three similar experiments. They present subjects with fifty statements (one for each US state) that say the state's largest city (by population) is its capital. Subjects answer "True" or "False" and predict what percentage of the other subjects answers "True." Their results show that the LST answer significantly outperforms the group answer: overall, errors were more than halved from 19 to 9, from 12 to 6, and from 19 to 4 in the three respective studies. They thus provide a promising new method to identify experts in a crowd.[18]

The question is whether the assumption behind the LST method extends to other questions than those about US state capitals. That is, holds the relationship between answering a question correctly and correctly predicting others' answer to that question more broadly? The relationship requires two things. First, the existence of expertise, which, as discussed in Subsections 2.2.1 and 2.2.2, is not necessarily the case. And, second, a correlation between two different sorts of expertise: about the subject in question and about other people's knowledge about that subject. This correlation may only exist in specific kinds of questions. The topographical questions about US state capitals – using the states' largest cities as the default in each statement – may be one of this kind. They are particularly likely to create a type of trick question for which expertise on the subject and expertise on other people's knowledge about that subject coincide. In contrast, evidence in the literature shows that experts are bad at predicting other people's expertise, particularly if those other people are not experts. Hinds (1999) finds that as expertise increases, one's ability to predict how long a novice takes to perform a complex task, such as using a mobile

[17] The capital of Pennsylvania is Harrisburg.

[18] More recently, Prelec, Seung, and McCoy (2017) introduces an alternative answering method to the LST method. It uses the same follow-up question as the LST method, but now to find the answer that is "more popular than people predict," rather than the answer by those who "are least surprised by the truth" (Prelec, Seung and McCoy 2013). Based on several experiments, they find this new answer also outperforms the group answer and confidence-weighted answers. See also Subsection 5.2.

phone, decreases.[19] Cho (2004) finds a similar inability of experts to empathise with non-experts: in a writing tasks, students are more able to incorporate feedback from their peers than from their instructors. This suggests that the correlation between expertise on a subject and on other people's knowledge about that subject is not always present. As such, the LST method may perform less well in other kinds of questions.

3. Methods

The recent LST method of Prelec, Seung, and McCoy (2013) looks promising. But the literature also suggests it may face problems in other kinds of questions than those tested by the authors. By adding a more complex question, I design an experiment to test the dynamics behind the LST method in a broader context.

3.1.Research question and hypotheses

The aim of this thesis is to test the LST method of Prelec, Seung, and McCoy (2013) in a more complex domain. Its research question is whether the LST answer also outperforms the group answer for a more complex question than that of the original experiments. The literature provides two reasons why this should not be the case: firstly, the nonexistence of expertise (Shanteau 2002) and, secondly, an inability of experts to predict other people's knowledge (Hinds 1999). Therefore, I hypothesise that (1) *the LST answer does not outperform the group answer for the more complex questions*. As a robustness check for the comparison, I also try to replicate the original results (Prelec, Seung and McCoy 2013) and thus hypothesise that (2) *the LST answer outperforms the group answer for the original questions*.

Poorer performance of the LST answer can be attributed to two factors: either non or less-existent expertise or experts' inability to predict other people's answers. This makes it hard to draw clear conclusions from hypotheses one and two. Therefore, I remedy this attribution problem by introducing the CWM method of Budescu and Chen (2015). The CWM method identifies experts based on subjects' performance data rather than follow-up question like the LST method. It thus merely relies on the existence of expertise and not on experts' ability to predict other people's answers. Consequently, I can use it to determine the (degree of) existence of expertise for different kinds of questions.

[19] Hinds (1999) suggests three heuristics – availability of their own experiences, anchoring on and insufficient adjustment from their own expertise, and oversimplification of the task – make experts overestimate non-experts' ability in performing a task.

The CWM method identifies experts based on (past) performance data. Because I use the CWM method as a controlling mechanism, I use the performance data over the entire experiment to identify experts and retrospectively predict their answer. This makes for an unrealistic answering method,[20] but incorporates the entire experiment's data in the CWM answer and hence makes the identification of experts most accurate. (Because the CWM method is inherently different from the group and LST method, a direct comparison is unfair. However, it is a good way to determine the underlying causes of performance differences in the LST method.)

Firstly, I use the CWM answer to determine the (degree of) existence of expertise in the different kinds of questions. The literature finds that the CWM answer outperforms the group answer by accurately identifying experts (Budescu and Chen 2015). But, as discussed above, there are reasons to believe that there exist no or few experts for the more complex questions (Shanteau 2002). I thus hypothesise that (3) *the CWM answer performs worse, relative to the group answer, for the more complex questions than for the original questions*.

Secondly, I compare the performance of the LST and CWM answers over the different kinds of questions. This comparison provides information about the underlying causes of performance differences for the two kinds of questions. (Note that performance differences between the LST and CWM methods as such are irrelevant because they entail an unfair comparison.) The LST method requires the existence of expertise and an ability of experts to predict other people's answers, while the CWM method only requires the existence of expertise. Decreasing expertise in the more complex questions thus affects both methods similarly (Shanteau 2002). In contrast, differences in the ability of experts to predict other people's answers only affect the LST method. Given that the literature suggests this ability is limited, but does not address different kinds of questions (Hinds 1999; Cho 2004), I assume the LST answer performs worse than the CWM answer for both kinds. I therefore hypotheses that (4) *the CWM answer outperforms the LST answer for both* (a) *the more complex questions and* (b) *the original questions*. Note that an important caveat to the former hypothesis is that both expert answers perform equally bad when expertise does not exist at all.

[20] Such retrospective usage is only possible in research. In real-world applications, the laws of time entail that only performance data over past questions can be used to predict future answers.

Finally, I use the CWM method to test experts' ability to predict other people's answers over the different kinds of questions. The LST method depends on a superior ability among experts for this, because it identifies them by this metric (Prelec, Seung and McCoy 2013). But other studies suggest experts may have inferior abilities in this regard (Hinds 1999; Cho 2004). I therefore hypothesise that (5) *the CWM experts' predictions of other people's answers do not outperform the CWM non-experts and group's predications of other people's answers for both* (a) *the more complex questions and* (b) *the original questions.*

3.2.Methodology

3.2.1. The different kinds of questions and their complexity

This thesis replicates and expands Prelec, Seung, and McCoy's (2013) study. Subjects are asked two questions about each of several statements. The first question asks whether the statement is "True" or "False." The second question asks what percentage of other subjects answers "True," i.e. it asks for a prediction of other people's answers. Statements are either based on the original questions from Prelec, Seung, and McCoy (2013), about US state capitals, or based on the more complex questions.

As the more complex questions, I adapt the bean-jar experiment, which is common in the literature (Treynor 1987; Krause, et al. 2011). It asks subjects to predict the number of beans in a jar. This kind of questions is more complex than those about US state capitals according to the characteristics identified by Shanteau (1992). As such, expertise is thus less likely to exist for it. These characteristics are, among others, the uniqueness of the task, the dynamisms of the stimuli, and the availability of the feedback. With respect to everyday life, predicting the number of beans in a jar is a more singular (i.e. unique) task than the relatively common (i.e. repetitive) questions about topography. Furthermore, the stimuli of beans in a jar are more dynamic because of changing numbers of beans and forms of jars which require different estimations. This contrast to the static topographical knowledge of US state capitals. Finally, bean-jar experiments rarely provide feedback, whereas topographical questions often do. Intuitively, this difference in complexity is also easy to see. On the one hand, my sample is unlikely to include "expert" predictors of beans in a jar. On the other hand, topography is often part of educational programmes or other studies and also more applicable in real life. As such, the difference in complexity between the two kinds of questions should be ample to test my hypotheses.

3.2.2. Computation of the group and LST answer

The group and LST answers are computed in accordance with Prelec, Seung, and McCoy's (2013) definitions. The group answer can be formulated as:

$$\underset{k}{\operatorname{argmax}} \Pr [X^r = k \mid \Omega = i^*] \tag{1}$$

where Ω is the state of the world i and X^r is the answer of respondent r. Moreover, i^* is the true state of the world. The group answer is thus that which is most likely, i.e. the answer which is most frequent.[21]

The LST answer is that of the subjects who are least surprised by the truth, which can be formulated as:

$$\underset{k}{\operatorname{argmax}} \Pr [\Omega = i^* \mid X^r = k] \tag{2}$$

The LST answer is thus that of the subjects who think the true state of the world the most likely, i.e. the answer of those who are most accurate in the second question. Unfortunately, the true state of the world (i^*) is unknown. For a binary question, the maximum (2) can fortunately be found through the comparison:

$$\frac{\Pr [\Omega = i^* \mid X^r = k]}{\Pr [\Omega = i^* \mid X^s = j]} \tag{3}$$

where the two probabilities are the inverse surprise of subjects who give either answer. Using Bayes' Rule, the division (3) can subsequently be rewritten as:

$$\frac{\Pr [X^r = k \mid \Omega = i^*] * \Pr [X^s = j]}{\Pr [X^s = j \mid \Omega = i^*] * \Pr [X^r = k]} = \frac{\Pr [X^r = k \mid \Omega = i^*] * \Pr [\mathrm{X}^s = j \mid X^r = k]}{\Pr [X^s = j \mid \Omega = i^*] * \Pr [\mathrm{X}^r = k \mid X^s = j]} \tag{4}$$

These probabilities are more easily estimated (see also Table 3.1). They are estimated by the frequencies of the two answers, k and j, for the probabilities on the left-hand sides of the nominator and denominator, respectively. The right-hand sides' probabilities are estimated by the predicted frequencies of the two answers, k and j, by those who answer the opposite, i.e. j and k,

[21] This way of computing the group answer, i.e. through the mean answer, is the most common one (Larrick, Mannes and Soll 2011). Other ways include the trimmed or winsorised mean (Jose, Grushka-Cockayne and Lichtendahl 2014) or median (Hora, et al. 2013), but these are irrelevant for the present binary question.

respectively. All four of these are directly provided by the first and second question, which therefore makes the computation of the LST answer straightforward from here.

TABLE 3.1 Estimations of the different probabilities in the LST method

Probability	Estimations
Pr [$X^r = k \mid \Omega = i^*$]	The frequency of answer k
Pr [$X^s = j \mid \Omega = i^*$]	The frequency of answer j
Pr [$X^r = k \mid X^s = j$]	The predicted frequency of answer k by those who answer j
Pr [$X^s = j \mid X^r = k$]	The predicted frequency of answer j by those who answer k

This information is based on the paper of Prelec, Seung, and McCoy (2013).

3.2.3. Computation of the CWM answer

The Contribution Weighted Model (CWM) method is developed by Budescu and Chen's (2015). It measures each subject's individual contribution to the correct group answer and bases its expert answer on this information, i.e. higher contributors receive more weight and vice versa. For every question, i, the group answer receives a score, S_i, based on the quadratic scoring formula:

$$S_i = a + b \sum_{r=1}^{R_i} (o_{ir} - m_{ir})^2 \tag{5}$$

where R_i is the total number of possible answers, r, and m_{ir} and o_{ir} are the group answer and the correct answer, respectively. The group answers' scores are aggregated for all questions with the formula:

$$S = a + b \sum_{i=1}^{N} \left(\sum_{r=1}^{R_i} (o_{ir} - m_{ir})^2 \right) \tag{6}$$

where N gives the total number of questions. By using constants $a = 100$ and $b = -50$, like Budescu and Chen (2015) do, scores range from 0 to 100, with 0 being the lowest possible score and 100 the perfect score. Each judge's, j, contribution, C_j, is calculated by comparing the group answers' performances with and without the judge for all questions, based on the formula:

$$C_j = \sum_{i=1}^{N} \left(S_i - S_i^{-j} \right) / N_j \tag{7}$$

The negative contributors are subsequently discarded and the remaining judges' contributions then produce a normalised weighted mean of their personal answer that creates the CWM answer.

I repeat the CWM method for all three question sets. That is, I independently identify and compute expert answers for the US state-capital and bean-jar questions, as well as for all questions taken together. This ensures that the CWM method always identifies the experts that have expertise with respect to the relevant kind of questions.

3.2.4. Outputs of the group, LST, and CWM answers

Based on the methodologies of the group, LST, and CWM answer, I calculate two outputs for each question. The first output is a "probability" of a correct answer, which lies somewhere between 0 and 1.[22] Before I calculate these, I adjust the "True"-or-"False" answers such that 0 and 1 always stands for respectively the incorrect and correct answer. For the group answer this is simply the mean answer of all subjects, i.e. the percentage of subjects that answers correctly. For the CWM answer, this is the weighted-mean answer of all subjects that are net positive contributors to the correctness of the group answers. The probability of a correct answer from the LST method is based on the relative size in inverse surprise of subjects that give either answer. For instance, if subjects that answer "True" are twice as surprised as subjects that answer "False", based on formula (3) the LST answer produces a probability output of 0.67 in favour of the "False" answer. The second output is the method's final answer, which can be either "True" or "False." It is based on rounding the first output to either "Correct" or "Incorrect," i.e. "1" for probabilities between 0.5 and 1 and "0" for those between 0 and 0.5.

3.2.5. Computation of the CWM experts' predictions of other people's answers

The CWM method uses each subject's individual contribution (C_j) to the correct group answer to identify experts. Experts, i.e. net positive contributors, produce a normalised weighted mean of their personal predications of other people's answers. This is what I refer to as the CWM experts' predictions of other people's answers. In the same way, "non-experts", i.e. net negative contributors, produce a normalised weighted mean of their personal predications of other people's answers. And this is what I refer to as the CWM non-experts' predictions of other people's answers. (The group's prediction of other people's answers – which I compare to those of the CWM experts and non-experts – is simply the mean of subjects' individual predictions of other people's answers.)

[22] As the description below shows, this is not a probability per se, but a value between 0 and 1 that indicates how strongly the answer point towards or away from the correct answer (1).

Because I want to compare the predictions' performance (rather than the predictions itself), I calculate each predictions' absolute difference from the true answer, i.e. its error. The error thus measures the (inverse) performance of each prediction, which I can easily compare. Again, I repeat the CWM method's (non-)expert identification for the US state-capital and bean-jar questions, as well as for all questions taken together. This ensures that the CWM method always identifies the experts that have expertise with respect to the relevant kind of questions.

3.3.Experimental design

3.3.1. Subject selection, incentives, duration, and order effects

The two kinds of questions are presented to subjects in an experiment with a within-subject design. The experiment is conducted through an online survey built with the *Qualtrics* software. My sample includes fellow students, in the M.Sc. in Economics and Business programme of the Erasmus School of Economics at the Erasmus University Rotterdam in the Netherlands (from both the Behavioural Economics and Financial Economics specialisations), as well as relatives and other social relations. This presumably creates a quite diverse subject pool with respect to gender, age, educational level, and interests. As such, my sample's expertise is sufficiently diverse for a proper functioning of the wisdom of the crowd and the LST and CWM methods. (Because I do not collect demographic data on my subjects, I cannot examine this assumption in more detail.)

The experiment is not directly incentivised, but I assume that my subjects answer honestly and carefully, as do Prelec, Seung, and McCoy (2013). Besides, Budescu and Chen (2015) rely on no other incentives than the publication of performance. Participation in the experiment is incentivised, though, by a €10 award for four randomly-selected subjects. Furthermore, I aim for a ten-minute duration of my experiment, which I deem at the higher end of reasonability for voluntary subjects. In practice, this means the experiment is limited to 24 statements, i.e. 12 about US state capitals and 12 about bean jars. Given the two-question format, the 24 statements correspond to a total of 48 questions. This equates to 25 seconds for each statements' pair of questions. (Because of its relationship to the first question, I expect the second question takes less time to answer compared to when it would be asked in isolation.)

The original and more complex questions statements are presented alternately to avoid order effects between the two. To make sure any remaining order effects are the same for all subjects, the order does not change between subjects. Remaining order effects are unlikely to be problematic because I compare the different kinds of questions statements and the relationship between the first and second question. The order within each kind of questions however, is

randomised. This avoids that the states and bean jars follow an easily recognisable pattern, such as increasing or decreasing size. Table A.1 in Appendix I gives an overview of the random order of the statements.

3.3.2. The more complex questions: bean-jar experiments

The more complex questions consist of estimating the number of beans in a jar, which are shown to subjects in a picture. Instead of a direct estimate, I frame the question such that it is comparable to the "True"-or-"False" statements about the US state capitals. The experiment uses coffee beans in different sized and shapes of glasses and jars to achieve twelve different setups. These range from a shot glass to a large jar and from 89 to 3074 coffee beans. Like the original study, this "test has some richness because problems range in difficulty, and [questions] are challenging for a variety of different reasons." (Prelec, Seung and McCoy 2013) The questions have the following format (Appendix II gives the used Dutch version of the question):

> "There are more than 1100 coffee beans in this glass. Do you think this statement is true or false?
> - o True
> - o False"

The number of coffee beans are estimates. They are calculated by dividing the total weight of the coffee beans in each jar by the average weight of a single coffee bean. This weight is 0.094 grams per coffee bean, based on a total weight of 94 grams for a thousand coffee beans. As a robustness check, I both weigh and count the number of coffee beans in the shot glass, which results in 85 and 89 coffee beans, respectively. Since this is an error of less than 5%, which should also decrease with larger jars and more coffee beans, I find this an acceptable way to determine the number of coffee beans in a jar.

In order to produce statements that can be true or false, the estimated number of beans is adjusted by 10%. There are several reasons for this. Most importantly, to make the "True"-or-"False" question feasible, the statement has to be true or false to a certain degree. Without sufficient adjustment, estimates that are only slightly off (e.g. 10 beans) are immediately incorrect. (Theoretically, a sufficiently large group produce a group answer that is correct to the single bean, but in my experiment the number of subjects and hence the group size is limited.) Secondly, earlier bean-jar experiments produce accurate group answers that are nevertheless up to 5% off the true number. To allow for this level of accuracy, the statement requires a certain degree of "True" or "False"-ness. Thirdly, the weight-based estimation requires a margin of error, given that it is 5% off the true number in my robustness check. Together, these make a 10% adjustment

appropriate. The adjustment is either upward or downward dependent on weight-estimated number of beans: i.e. upwards for the even numbers and downwards for the uneven numbers. This is deemed practically random, given the number of confounding factors that determine whether the estimated number of beans is even or uneven.

The adjusted number of beans are rounded to hundreds for ease of comprehension. By coincidence this rounding only exaggerates the earlier adjustments and hence does not intervene with its aim. The resulting numbers are used to compose a statement which says the jar contains more beans than the respective number. Subjects are asked whether the statement is "True" or "False." For my bean jars, six statements are true and six statements are false. (Table A.2 in Appendix I gives an overview of the different jars, the weight of the beans, the estimated number of beans, and the adjusted number of beans.)

3.3.3. The original questions: US state capitals and the follow-up

The original questions come from Prelec, Seung, and McCoy (2013). They consist of "True"-or-"False" questions on statements that suggest a state´s largest city is its capital. These questions range in difficulty for different reasons, as the authors point out: "The prominence of a city is sometimes misleading (Philadelphia-Pennsylvania), and sometimes a valid cue (Boston-Massachusetts), and many less populous states have no prominent city." (Prelec, Seung and McCoy 2013) The questions have the following format (Appendix II gives the used Dutch version of the question):

> "The capital of the US state Pennsylvania is Philadelphia Do you think this statement is true or false?
> - o True
> - o False"

Unlike the original study, I only use twelve states (instead of all fifty) to limit the duration of the survey. I pick the states with the highest populations in accordance with the latest US Census Bureau (2016b) estimates. My Dutch sample is less knowledgeable on US state capitals than those of the original study (students at the Massachusetts Institute of Technology and Princeton University). By using the most populous states, the questions are easier and I (partly) mitigate this deficiency. For balance between true and false statements, I maximise the number of either at two thirds (i.e. a maximum of eight out of twelve states can be either true or false). The largest cities for each state are also based on population size and in accordance with the latest US Census Bureau (2016a) estimates. For the selected states, four statements are true and eight statements

are false. (Table A.3 in Appendix I gives an overview of the selected states, capitals, and largest cities.)

The second question, on the prediction of the other subjects' answers, follows each "True"-or-"False" question (i.e. both the US state-capital and bean-jar questions). They are shown simultaneously with the first questions, such that subjects can review the statement. The questions have the following format (Appendix II gives the used Dutch version of the question):

> "Estimate how all other respondents answer the above question. What percentage do you think will answer "True"?
> 0% ---------- 100%"
> *(A slider moves across this line to indicate the chosen percentage.)*

4. Results

During October 2016, a total of 53 subjects completed the questionnaire. Based on their answers, I compute the group, LST, and CWM answer for all questions and for the US state-capital and bean-jar questions separately. I follow the above methodology to calculate the probability of a correct answer for the group, LST, and CWM answer per question and the resulting number of correct answers. These are averaged over all questions for the total number of correct answers and probability of correct answers, for the US state-capital and bean-jar questions, separately and taken together. Table 4.1 below shows the results. (Table A.4 and A.5 in Appendix III give an overview of the results for each question.[23])

4.1. Comparison of the group answers to the LST answers

There is a clear difference between the group and LST answer in the total number of correct answers. They are 16 and 19, respectively, out of a total of 24 questions. This difference is wholly attributable to the US state-capital questions, where the group answer is correct only 8 and the LST answer 11 out of 12 times. In contrast, both the group and LST answer are correct 8 out of 12 times in the bean-jar questions. I examine the statistical significance of these differences with the two-sided matched-pairs sign test.[24] The LST answers are not statistically different from the group answers for either the US state-capital questions, the bean-jar questions, or all questions taken together.

[23] The data for the CWM method for all question taken together are available upon request.
[24] For the number of correct answers, I use the two-sided matched-pairs sign test to determine any statistically significant differences between the answering methods, because the dummy variable for correct answers provides categorical data. (The two-sided matched-pairs sign test is often referred to as the McNemar's test in case it is used for two dummy variables (McNemar 1947; Conover 1999).)

TABLE 4.1 Results of the group, LST, and CWM answers

	All questions	US state capitals	Bean jars
	Total number of correct answers		
Group answer	16	8	8
LST answer	19	11	8
CWM answer†	23**	12	11
	Average probability of a correct answer		
Group answer	0.63	0.64	0.62
LST answer	0.63	0.66	0.60**
CWM answer†	0.75***	0.87***	0.74
N	24	12	12

† The CWM method identifies experts based on their contributions to correct group answers. Because expertise may differ across the mix of all questions and the US state-capital or bean-jar questions separately, experts are identified independently for each of these kinds of questions.
The statistical differences of the LST and CWM answers from the group answers are indicated for the 10% (*), 5% (**), and 1% (***) significance levels in a two-sided matched-pairs sign test (for the total number of correct answers) and two-sided matched-pairs Student's *t* test (for the average probabilities of a correct answer).

The probabilities of a correct answer provide more-detailed information about the difference between the group and LST answers. For all question together, the group and LST answers are now the same, namely 0.63. The difference between the group and LST answers remains for the US state-capital questions, albeit less pronounced, with respectively 0.64 and 0.66. The inverse difference occurs for the bean-jar questions, where the group answer stands at 0.62 and the LST answer stands slightly lower at 0.60. I determine the significance of these differences with the two-sided matched-pairs Student's *t* test.[25] Again, the LST answers are not statistically different from the group answers for the US state-capital questions or all questions taken together. But the differences between the group and LST answers for the bean-jar questions is statistically significant at a 5% level.

The number of correct answers and the probabilities of a correct answer provide different pictures of the group and LST answers' performance. Although the probabilities differ little between the two answering methods, the corresponding numbers of correct answers differ considerably. One reason for this is that in at least two questions, a small advantage of the LST answer over the group answer pushes it just over the 0.5 mark, resulting into a rounded answer

[25] For the probabilities of a correct answer, I use the two-sided matched-pairs Student's *t* test to determine any statistically significant differences between the answering methods, because the probabilities of a correct answer provide normal data. That is, five and six (out of six) variables pass respectively the Shapiro-Wilk *W* test and skewness and kurtosis tests for normality.

that is correct (see Texas and Illinois in Table A.4 in Appendix III). Given these roundingeffects, I follow the probabilities rather than the number of correct answers in case of discrepancies.

My findings do not support Hypothesis 1 and 2. Although the bare numbers of correct answers are in line with my first and second hypotheses: the LST answer does not outperform the group answer for the more complex questions (in concurrence with Hypothesis 1); and the LST answer does outperform the group answer for the original questions (in concurrence with Hypothesis 2). These results, however, are statistically insignificant. Furthermore, the probabilities of a correct answer do not follow the hypotheses altogether. Here, the LST answer underperforms the group answer for the more complex questions (in contrast to Hypothesis 1); and the LST and group answers perform equally well for the original questions (in contrast to Hypothesis 2). And these results are statistically significant. I thus reject Hypothesis 1 and 2.

Despite the rejection of my hypotheses, the mixed findings support the dynamic behind hypotheses one and two. This dynamic entails that the LST answers' performance worsen relative to that of the group answers for the more complex questions. It is based on the idea that the LST method is at a disadvantage with the more complex questions because there exists less expertise. I nevertheless must reject hypothesis one and two because for both kinds of questions the LST answer performs worse, relative to the group answer, than I hypothesised.

4.2.Comparison of the group and LST answers based on the CWM method

4.2.1. Analysis of the CWM answers compared to the group and LST answers

Besides the group and LST answers, I compute the CWM answers as well. Table 4.1 above also shows these answers. Unsurprisingly, the CWM answer is more often correct than both the group and LST answer.[26] It is correct 23 out of 24 times for all questions taken together, compared to 16 and 19 times for respectively the group and LST answers. It is correct for all 12 US state-capital questions and 11 of the 12 bean-jar questions, compared to respectively 11 and 8 correct LST answer and 8 correct group answers for both kinds of questions. I examine the statistical significance of these differences with the two-sided matched-pairs sign test.[27] Only one of the differences between the CWM answers and the group and LST answers is statistically significant. The CWM answers and group answers for all questions taken together are statistically different at a 5% significance level. The CWM answers and LST answers for all questions taken together

[26] See for an explanation of the unfairness of directly comparing the CWM method with the group and LST answer Subsection 3.1.
[27] See footnote 24.

are not statistically different, neither are the differences between the CWM answer and group and LST answers for the US state-capital and bean-jar questions.

A more-detailed comparison is possible based upon the probabilities of a correct answer. Again, unsparingly, the CWM answer is more likely correct than the group and LST answer: for all questions taken together (0.74 compared to twice 0.63); for the US state-capital questions (0.87 compared to respectively 0.64 and 0.66); and for the bean-jar questions (0.74 compared to 0.62 and 0.60). I investigate the significance of these differences with the two-sided matched-pair Student's *t* test.[28] The CWM answers are statistically different from the group answers for all questions taken together and the US state-capital questions, both at a 1% significance interval. For these questions, the CWM and LST answers are also statistically different at a 1% significance level. In contrast, for the bean-jar questions, the CWM answers are not statistically different from the group and LST answers.

The results are predominantly in line with Hypotheses 3, on the relative performance of the group and CWM answers. For the number of correct answers, the CWM answer performs worse for the more complex questions than for the original questions, while the group answer performs similarly for both. This concurs with Hypothesis 3, but the differences are not statistically significant. With respect to the probabilities of a correct answer, both the group and CWM answers perform worse for the more complex questions, but the latter much more so. The previous concurrence with Hypothesis 3 is now also reflected in the statistical significance of the difference. That is, there is a statistically significant difference for the original questions, but there is not for the more complex questions. The CWM answer thus performs worse, relative to the group answer, for the more complex questions. I thus cannot reject Hypothesis 3.

Hypothesis 4, on the relative performance of the LST and CWM answers, is backed only half. For both a more complex and the original questions, the CWM answer outperforms the LST answer in the number of correct answers, but the outperformance is not statistically significant (providing mixed evidence for Hypothesis 4). The CWM answer's probability of a correct answer is also larger than that of the LST answer for both kinds of questions. But with respect to the more complex questions, this difference is not statistically significant (in contrast to Hypothesis 4a). For the original questions, the CWM answer's outperformance of the LST answer is significant (in concurrence with Hypothesis 4b). The CWM answer therefore outperforms the LST answer for the original questions, but not for the more complex questions. I thus reject Hypothesis 4a, but cannot reject Hypothesis 4b.

[28] Here, respectively eight and nine (out of nine) variables pass the Shapiro-Wilk *W* test and skewness and kurtosis tests for normality. See also footnote 25.

Based on these hypotheses, it is possible to further analyse the changes of expertise's existence and experts' ability to predict other people's answers. Firstly, the CWM answer performs worse, relative to the group answer, for the more complex questions than for the original questions. Because the CWM method identifies experts and uses their answer, this finding confirms that for the more complex questions there exists less expertise. Secondly, the CWM answer also performs worse, relative to the LST answer, for the more complex questions than for the original questions. The previous finding, i.e. expertise decreases when complexity increases, affects the LST and CWM answers similarly, which means any remaining differences over the two kinds of questions are due to changes in experts' ability to predict other people's answers. The LST answer's relatively strong performance for the more complex questions therefore suggest experts' ability to predict other people's answers increases with complexity.

4.2.2. Analysis of the CWM experts' predictions of other people's answers

The previous findings indirectly suggest that experts are more able to predict other people's answers for the more complex questions than the original kind. I can easily test this suggestion directly, because the LST method asks for predictions of other people's answers and the CWM method identifies experts. I compare the predictions of other people's answers by the group, the CWM experts, and CWM non-experts. For this purpose, I use their predictions' average absolute difference with respect to the true distribution of answers, i.e. their error. Table 4.2 provides these errors for the US state-capital questions, the bean-jar questions, and all questions taken together. (Table A.6 and A.7 in Appendix III give an overview of the predictions and errors for each question.[29])

The absolute errors of the CWM experts and non-experts suggest a small outperformance by the former: for the US state-capital questions they are respectively 0.23 and 0.24; for the bean-jar questions 0.22 and 0.32; and for all question taken together 0.24 and 0.26. The group's performance lies somewhere in between. I determine the significance of these differences with the two-sided matched-pairs Student's *t* test and the two-sided Wilcoxon signed-rank test.[30] The differences in errors between the CWM experts and non-experts is insignificant for the US state-capital questions and for all questions taken together. However, the CWM experts and non-experts have different errors for the bean-jar questions with a 5% significance level. The group's errors

[29] The data for the CWM method for all question taken together are available upon request.

[30] The errors of the US state-capital and bean-jar questions pass the Shapiro-Wilk *W* test and skewness and kurtosis tests for normality. The error for all questions taken together do not at a 10% significance level. For these errors, I use the two-sided Wilcoxon signed-rank test to determine any statistically significant differences from the group for the CWM experts and non-experts, because the errors do not provide normal data. See also footnote 25.

are also different from the CWM experts and non-experts for the bean-jar questions (at 10% and 1% significance levels, respectively) and insignificantly different for the US state-capital questions and all questions taken together.

Hypothesis 5 states that CWM experts do not outperform the CWM non-experts in their ability to predict other people's answers. My data show the CWM experts' predictions actually do outperform those of the CWM non-experts for the more complex questions. This is, however, not the case for the original questions, where the CWM experts and non-experts predict other people's answers no different. I thus reject Hypothesis 5a, but cannot reject Hypothesis 5b. Like my results on Hypothesis 4, this suggest experts' ability to predict other people's answers increases with complexity.

TABLE 4.2 Prediction performance of the group and CWM experts

	All questions	US state capitals	Bean jars
	Average absolute error in the predicted answers		
CWM experts†	0.24	0.23	0.22
CWM non-experts†	0.26	0.24	0.32**
Group	0.25	0.24	0.27*
N	24	12	12

† The CWM method identifies experts based on their contributions to correct group answers. Because expertise may differ across the mix of all questions and the US state-capital or bean-jar questions separately, expert weights are identified independently for each of these categories of questions.
The statistical differences from the CWM experts' predictions are indicated for the 10% (*), 5% (**), and 1% (***) significance levels in a two-sided Wilcoxon signed-rank test (for all questions taken together) and a two-sided matched-pairs Student's *t* test (for the US state-capital and bean-jar questions).

5. Concluding remarks

5.1. Discussion

The research question of this paper is whether the LST answer also outperforms the group answer for a more complex question than that of the original experiments. Prelec, Seung, and McCoy (2013) show the accuracy of the LST answer for US state-capital questions in a comparison with the group answer. I am unable to fully replicate this result. Some of my results suggest the LST answer may be better for US state-capital questions then the group answer, but most show the LST answer performs similarly or worse than the group answer.

This finding makes the research question's formulation slightly difficult, but I continue with comparing the group and LST answers in the original and more complex questions. The

literature suggests expertise decreases or even disappears with increasing complexity (Shanteau 1992; 2002). If this is the case, the LST answer, which tries to identify experts, should do worse the more complex questions. My results confirm this expectation. When I compare (simple) US state-capital questions to (complex) bean-jar questions, the LST answer's performance, compared to that of the group answer, is worse for the latter. There thus seems to exist less expertise when there is more complexity.

Because the LST answer not only depends on the existence of expertise, but also on expert identification through their ability to predict other people's answers, I cannot immediately attribute the above effect to either source. The former seems to have an effect based on the above findings. But the latter may also be problematic, since the literature suggests an inability of experts to predict other people's answers, especially that of non-experts (Hinds 1999; Cho 2004). I address this issue, by also computing the CWM answer, which does not require experts' ability to predict other people's answers, but merely requires the existence of experts. The CWM answer outperforms the group answer for both kinds of questions, but this outperformance decreases and becomes statistically insignificant for the more complex questions. This suggest there indeed exist less expertise when there is more complexity and, as such, the LST answer's worse performance in the more complex questions is at least partly driven by this difference.

The increase in complexity affects the LST and CWM answers differently. When I compare the two, the latter outperforms the former in both kinds of questions. But the outperformance is only statistically significant in the (simple) US state-capital and not in the (complex) bean-jar questions. The LST answer thus performs better, relative to the CWM answer, for the more complex questions. Since they are equally exposed to the existence of expertise, this must thus result from a difference in experts' ability to predict other people's answers over the two kinds of questions. Since the LST answers performs relatively better for the more complex questions, experts' ability to predict other people's answers seems to increase with complexity. I check this finding by directly comparing the CWM experts, non-experts, and group's predictions of other people's answers. The CWM experts' predictions outperform those of the non-experts and group for the (complex) bean-jar questions, while there is no statistically significant difference for the (simple) US state-capital questions. This confirms that experts' ability to predict other people's knowledge increases when complexity increases.

Less expertise for the more complex question is likely the result of the increased complexity of said question, in accordance with Shanteau (2002). However, it is not clear why experts' ability to predict other people's answers differs between the two kinds of questions. There are no obvious reasons for experts to be more able to predict other people's answers for the more complex questions. Hinds (1999) attributes experts' lack of ability to predict other people's

answers to three biases and heuristics. These are experts' availability of experiences, oversimplification of the task, and anchoring on and insufficient adjustment from their own expertise. This may explain the difference in experts' ability to predict other people's answers Given the likely novelty of predicting the number of beans in a jar, experts may lack available (easy) experiences. Since they also perform the task for the first time, it may also be harder to oversimplify it. Finally, the "True"-or-"False" statements themselves may suggest a bean count that serves as an anchor, rather than any self-created anchor by experts. Overall, experts' lack of experience in the bean-jar questions may thus improve their ability to predict other people's answers.

5.2.Limitations and further research

There are number of limitations to this research. Firstly, the sample size of my questionnaire is rather small. I limit its duration to ten minutes, because participation is voluntary. (Even with the current limitation, I receive much negative feedback from subjects about the length and repetitiveness of the questionnaire.) This means I have a total of 24 questions with only 12 questions of the original and 12 of the more complex kind. Such a small sample reduces the statistical power of my results. Although some of the plain numbers (e.g. 8 versus 11 correct answers out of 12 questions) suggest considerable differences, these results lack statistical significance. Further research ideally uses more questions to achieve a larger sample size.

Secondly, the replication of the original study by Prelec, Seung, and McCoy (2013) is also flawed in several other ways. The number of subjects and questions in my questionnaire are substantially lower: 53 compared to 116 subjects and 12 compared to 50 questions. Furthermore, my heterogeneous Dutch sample is considerably different from that of the original study (students at the Massachusetts Institute of Technology and Princeton University). This probably extends to their expertise on US state capitals as well. Although the share of correct and incorrect group and LST answers do not differ that much across the studies, the probability of a correct answer is probably much lower in my study (the original study does not report these data). But not only are the Dutch less knowledgeable than Americans about US state capitals, they are probably also less knowledgeable about each other's knowledge about the subject. Day-to-day conversation or quizzing about US state capitals, which creates such knowledge, is much less likely in the Netherlands than in the US. It will be interesting to see how similar samples perform on different questions (e.g. European capitals) or how other samples (e.g. Mexicans or British) perform on the same kinds of questions.

Thirdly, I limit my study to two kinds of questions, about the US state capitals and beans in a jar, to reflect a simple and complex question, respectively. These decisions are heavily

influenced by Shanteau's (2002) work on the domain-dependency of expertise. But they are also largely the result of precedents such as the original study (Prelec, Seung and McCoy 2013) and earlier research on the wisdom of crowds (Treynor 1987; Krause, et al. 2011). However, the characteristics of complexity that affect expert-performance domains (e.g. stimuli, repetitiveness, and feedback) allow for many kinds of simple and complex questions (Shanteau 1992; 2002). Ideally, further research uses a variety of simple and complex questions within in the framework of expertise's domain-dependency, to verify these findings for broader applicability.

Fourthly and more fundamentally, the current questions concerned "True" or "False" statements, which in themselves affect the characteristics of questions. For the US state capitals, this creates statements that are easily recognisable trick questions, which may be why the LST answer is particularly effective in the original study. For the bean jars, it anchors answers around the given bean count, which can also have unintended consequences. Such statements also considerably limit the sorts of questions that can be answered. For instance, questions with more than two feasible answers do not fit "True"-or-"False" statements.

Additionally, the focus of mine and much other research has been on the existence of expertise for different kinds of questions, but expert identification is also of great importance. For the LST answer, this is based on experts' ability to predict other people's answers. As the literature shows, this ability is all but certain among experts (Hinds 1999; Cho 2004). Hinds (1999) points towards several reasons that may explain this: availability of their own experiences, anchoring on and insufficient adjustment from their own expertise, and oversimplification of the task. Further research, with questions that differ in their susceptibility to these reasons, will provide more insight into what determines experts' ability to predict other people's answers.

Finally, a more recent paper by Prelec, Seung, and McCoy (2017) introduces an alternative to the LST method. Instead of the answer by those who "are least surprised by the truth" (Prelec, Seung and McCoy 2013), it proposes the answer that is "more popular than people predict." Like the LST method, it uses a follow-up question that asks individuals to predict how others answer the initial question. Prelec, Seung, and McCoy (2017) show this answering method outperforms group and confidence-based answers in several experiments. In the future, comparisons between variations of these two methods may further explore the dynamics behind the ability of both groups and experts to predict other people's knowledge for different questions and the possibility of this information to compute the correct answer.

5.3.Conclusion

Scientists and non-scientist alike apply the wisdom of the crowd to a large variety of questions. Whether this produces a group answer that is correct depends on several factors, such as incorrect

conventional wisdom, biases, and worse-than-random answers. In case the crowd is unwise, expert answers can be a good alternative. There are numerous ways to create an expert answer, of which a recent attempt by Prelec, Seung, and McCoy (2013) looks especially promising. They identify experts as those who best predict how others answer the question concerned, i.e. those who are "Least Surprised by the Truth" probably provide a good expert answer (the LST answer). In their experiment they use "True"-or-"False" statements about the capitals of US states. But the question is, whether the LST answer performs similarly well for other questions. Therefore, I replicate and extent their research with a more complex question.

In addition to the US state-capital questions, I ask subjects to estimate the number of beans in a jar. This question is more complex on characteristics such as stimuli, repetitiveness, and feedback, and it therefore allows for less expertise according to the literature (Shanteau 1992; 2002). As a result, I expect the LST answer to have no or a smaller advantage over the group answer for the more complex questions. To test this hypothesis, I conduct an experiment with 24 pairs of questions (12 of each kind) among a diverse group of subjects. It confirms that the LST answer performs worse relative to the group answer for the more complex questions. However, the absolute performance of the LST answer is lower than in the original study, i.e. it performs similar or worse than the group answer.

Although the above results suggest less-existent expertise causes the LST answer to perform worse with more complexity, another aspect of the LST answer, i.e. the ability of experts to predict other people's answers, may also cause this difference. By using an expert answer that does not rely on this feature, i.e. the CWM answer that identifies experts based on (past) performance data, I can determine which aspect is the cause of the difference. My results show that the CWM answer performs worse, relative to the group answer, for the more complex questions. This confirms that the bean-jar question's complexity allows for less expertise and thus (partly) causes the LST answer to perform worse for this kind of questions.

I not only examine how complexity affects the existence of expertise, but also whether experts' ability to predict other people's knowledge changes over the different kinds of questions. The literature suggests such ability is all but evident (Hinds 1999). For this purpose, I compare the LST and CWM answers to determine if the complexity of the questions affects them differently. The CWM answer outperforms the LST answer in the original questions, but not in the more complex questions. Existence of expertise across the two kinds of questions affects both answering methods similarly, but experts' ability to predict other people's answers – unique to the LST answer – does not. This difference therefore indicates an increase in experts' ability to predict other people's answers for the more complex questions. I directly test this by looking at the CWM experts' predictions of other people's answers and indeed find that they make better

predictions, relative to the CWM non-experts and the whole group, for the more complex questions. Experts ability to predict other people's knowledge thus increases from the original to the more complex questions, while in the meantime the existence of expertise decreases.

The research on which these conclusions draw faces several important limitations. It lacks much statistical power due to the small sample size (i.e. 24 questions in total, 12 of each kind). In replicating the original study, this may be a considerable problem, as ar the number of subjects and their characteristics. Especially their origins (Dutch instead of American) is an important factor that affects expertise in US state capitals and probably one's ability to predict other people's answers as well. Also, the two kinds of questions, although based on the literature, are only a "sample of one" for their respective kinds of questions (i.e. simple and complex).

Further research can address many of these shortcomings. Other experimental setups may increase sample size to improve statistical power. Repetition of studies with other samples and similar questions or similar samples and other questions will ensure the broader validity of the results. It can also closer pinpoint the kinds of questions most suitable for either the group, LST, or CWM method. Finally, experts' ability to predict other people's answers and possible factors that influence it, can pose other interesting and more novel approaches to determine the appropriate answering method.

REFERENCES

Armstrong, J.S. "Combining forecasts." In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, by J.S. Armstrong, 417-439. Norwell, MA: Kluwer Academic Publishers, 2001.

—. "The seer-sucker theory: the value of experts in forecasting." *Technology Review* 83 (1980): 16-24.

Ashton, A.H., and R.H. Ashton. "Aggregating subjective forecasts: some empirical results." *Management Science* 31 (1985): 1499-1508.

Barberis, N., and R. Thaler. "A survey of behavioral finance." In *Handbook of the Economics of Finance, Vol. 1B*, by G.M. Constantinides, M. Harris and R. Stulz, 1053-1128. Amsterdam: North Holland, 2003.

Berg, J., R. Forsythe, F. Nelson, and T. Rietz. "Results from a dozen years of election futures markets research." In *Handbook of Experimental Economic Results, Vol. 1*, by C.R. Plott and V.L. Smith, 742-751. Amsterdam: North Holland, 2008.

Bogle, J. *John Bogle on investing.* New York, NY: McGraw Hill, 2001.

Broomell, S., and D.V. Budescu. "Why are experts correlated? Decompasing correlations between judges." *Psychometrika* 74-3 (2009): 531-553.

Bruce, R.S. "Group judgements in the fields of lifted weights and visual discrimination." *Journal of Psychology* 1 (1935): 117-121.

Budescu, D.V., and E. Chen. “Identifying expertise to extract the wisdom of crowds.” *Management Science* 61-2 (2015): 267-280.

Buffett, M., and D. Clark. *The Tao of Warren Buffett.* New York, NY: Scribner, 2006.

Carrol, J.S., and J.W. Payne. "The psychology of parole decision process: a joint application of attribution theory and information-processing psychology." In *Cognition and Social Behavior,* by J.S. Carrol and J.W. Payne, 13-32. Hillsdale, NJ: Erlbaum, 1976.

Chapman, G.B., and E.J. Johnson. “Incroporating the irrelevant: anchors in judgements of belief and value.” In *The psychology of intuitive judgment,* by T. Gilovich, D.W. Griffin and D. Kahneman, 120-138. New York, NY: Cambridge University Press, 2002.

Chase, W.G., and H.A. Simon. “Perception in chess.” *Cognitive Psychology* 4 (1973): 55-81.

Chi, M.T.H. “Knowledge structure and memory development.” In *Children's Thinking: What Develops?,* by R. Siegler, 73-96. Hillsdale, NJ: Erlbaum, 1978.

Chi, M.T.H., R. Glaser, and M.J. Farr. *The Nature of Expertise.* Hillsdale, NJ: Erlbaum, 1988.

Cho, K. “When experts give worse advice than novices: the type and impact of feedback given by students and an instructor on student writing.” *Unpublished dissertation, University of Pittsburg,* 2004.

Christensen, E.E., R.C. Murry, K. Holland, J. Reynolds, M.J. Landay, and J.G. Moore. “The effect of search time on perception.” *Radiology* 138 (1981): 361-365.

Condorcet, M.J.A.N. de Caritat, Marquis de. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité de voix.* Paris, 1785.

Conover, W.J. “Section 3.4: The Sign Test.” In *Practical Nonparametric Statistics,* by W.J. Conover, 157-176. New York: Wiley, 1999.

Cooke, R. *Experts in uncertainty: opinion and subjective probability in science.* New York, NY: Oxford University Press, 1991.

Dawes, R.M. “A case study of graduate admissions: application of three principles of human decision making.” *American Psychologist* 26 (1971): 180-188.

Ebbesen, E., and V. Konecni. "Decision making and information integration in the courts: the setting of bail." *Journal of Personality and Social Psychology* 32 (1975): 805-821.

Einhorn, H. “Expert judgment: some neccessary conditions and an example.” *Journal of APplied Psychology* 59 (1974): 562-571.

Estlund, D.M. *Democratic Authority: A Philosophical Framework.* Princeton, NJ: Princeton University Press, 2009.

Flores, B.E., and E.M. White. “Subjective vs objective combining of forecasts: an experiment.” *Journal of Forecasting* 8 (1989): 331-341.

Forsythe, R., F. Nelson, G.R. Neumann, and J. Wright. “Anatomy of an experimental political stock market.” *American Economic Review* 82-5 (1992): 1142-1161.

Galton, F. “Vox populi.” *Nature* 75 (1907): 450-451.

Glenberg, A.M., and W. Esptein. “Inexpert calibration of comprehension.” *Memory and Cognition* 15 (1987): 84-93.

Goodin, R.E. *Reflective Democracy.* Oxford: Oxford University Press, 2005.

Gordon, K.H. “Group judgements in the field of lifted weights.” *Journal of Experimental Psychology* 7 (1924): 398-400.

Grofman, B., and S. Feld. “Rousseau's general will: a Condorcetian perspective.” *American Political Science Review* 82 (1988): 567-576.

Groot, A.D. de. *Though and Choice in Chess.* The Hague: Mouton, 1965.

Guthrie, C., J.J. Rachlinski, and A.J. Wistrich. “Inside the judicial mind.” *Cornell Law Review* 86 (2001): 790-791.

Hastie, R., D.A. Schkade, and J.W. Payne. “Do plaintiffs' requests and plaintiffs' identities matter?” In *Punitive Damages: How Juries Decide*, by C.R. Sunstein, R. Hastie, J.W. Payne, D.A. Schkade and W.K. Viscusi, 62-74. Chicago, IL: University of Chicago Press, 2002.

Henrich, J., et al. “Group report: what is the role of culture in bounded rationality?” In *Bounded Rationality: The Adaptive Toolbox*, by G. Gigerenzer and R. Selten, 343-359. Cambridge, MA: The MIT Press, 2001.

Hinds, P.J. “The curse of expertise: the effects of expertise and debiasing methods on prediction of novice performance.” *Journal of Experimental Psychology: Applied* 5-2 (1999): 205-221.

Hoerl, A.E., and H.K. Fallin. “Reliability of subjective evaluations in a high incentive situation.” *Journal of the Royal Statistical Society* 137 (1974): 227-230.

Hoffman, P., P. Slovic, and L. Rorer. “An analysis of variance model for the assessment of configural cue utilization in clinical judgment.” *Psychological Bulletin* 69 (1968): 338-349.

Hora, S.C., B.R. Fransen, N. Hawkins, and I. Susel. “Median aggregation of distribution functions.” *Decision Analysis* 10-4 (2013): 279-291.

Hughes, H.D. “An interesting corn seed experiment.” *The Iowa Agriculturalist* 17 (1917): 424-425.

John, L.K., G. Loewenstein, and D. Prelec. “Measuring the prevalence of questionable research practices with incentives for truth telling.” *Psychological Science* 23 (2012): 524-532.

Johnson, E.J. “Expertise and decision under uncertainty: performance and process.” In *The Nature of Expertise*, by M.T.H. Chi, R. Glaser and M.J. Farr, 209-228. Hillsdale, NJ: Erlbaum, 1988.

Jose, V.R.R., Y. Grushka-Cockayne, and K.C., Jr. Lichtendahl. “Trimmed opinion pools and the crowd's calibration problem.” *Management Science* 60-2 (2014): 463-475.

Kerr, N.L., R.J. MacCoun, and G.P. Kramer. “Bias in judgment: comparing individuals and groups.” *Psychological Review* 103-4 (1996): 687-719.

Knight, H.C. “A comparison of the reliability of group and individual judgments.” *Unpublished master thesis. Columbia University.* 1921.

Krause, S., R. James, J.J. Faria, G.D. Ruxton, and J. Krause. “Swarm intelligence in humans: diversity can trump ability.” *Animal Behaviour* 81 (2011): 941-948.

Kundel, H.L., and C.F. Nodine. “Interpreting chest radiographs without visual search.” *Radiology* 116 (1975): 527-532.

Larkin, J.H., J. McDermott, D.P. Simon, and H.A. Simon. “Expert and novice performance in solving physics problems.” *Science* 208 (1980): 1335-1342.

Larreche, J.C., and R. Moinpour. "Managerial judgment in marketing: the concept of expertise." *Journal of Marketing Research* 20 (1983): 110-121.

Larrick, R.P., A.E. Mannes, and J.B. Soll. “The social psychology of the wisdom of crowds.” In *Frontiers of Social Psychology: Social Judgment and Decision Making*, by J.I. Krueger, 227-242. New York, NY: Psychology Press, 2011.

Lock, A. “Integrating group judgments in subjective forecasts.” In *Judgmental Forecasting*, by G. Wright and P. Ayton, 109-128. Chichester: Wiley, 1987.

Lomborg, Bjorn. *Global Crisis, Global Solutions.* Cambridge: Cambridge University Press, 2004.

Lorge, I., D. Fox, J. Davitz, and M. Brenner. “A survey of studies contrasting the quality of group performance and individual performance, 1920-1957.” *Psychological Bulletin* 55-6 (1958): 337-372.

Maloney, M.T., and J.H. Mulherin. “The complexity of price discovery in a efficient market: the stock market reaction to the Challenger crash.” *Journal of Corporate Finance* 9 (2003): 453-479.

Mayer, R.E. *Thinking, Problem Solving, Cognition.* 2. New York, NY: Freeman, 1991.

McNemar, Q. “Note on the sampling error of the difference between correlated proportions or percentages.” *Psychometrika* 12-2 (1947): 153-157.

Odean, T. “Volatility, price, and profit when all traders are above average.” *Journal of Finance* 53-6 (1998): 1887-1934.

Oskamp, S. “Overconfidence in case-study judgment.” *Journal of Consulting Psychology* 29 (1965): 261-265.

—. “The relationship of clinical experience and training methods to several criteria of clinical prediction.” *sychological Monographs* 76 (1962).

Patel, V., and G.J. Groen. “Knowledge based solution strategies in medical reasoning.” *Cognitive Science* 10 (1986): 91-116.

Prelec, D. “A bayesian truth serum for subjective data.” *Science* 306 (2004): 462-466.

Prelec, D., H.S. Seung, and J. McCoy. “A solution to the single-question crowd wisdom problem.” *Nature* 541 (2017): 532-535.

—. “Finding truth even if the crowd is wrong.” 2013.

Rowe, G. “Perspectives on expertise in the aggregation of judgement.” In *Expertise and Decision Support*, by G. Wright and F. Bolger, 155-180. New York, NY: Plenum Press, 1992.

Rowse, G.L., D.H. Gustafson, and R.L. Ludke. “Comparison of rules of aggregating subjective likelihood ratios.” *Organizational Behavior and Human Performance* 12 (1974): 274-285.

Sauer, R.D. “The economics of wagering markets.” *Journa of Economic Literature* 36 (1998): 2021-2064.

Shanteau, J. “Competence in experts: the role of task characteristics.” *Organizational Behavior and Human Decision Processes* 53 (1992): 252-266.

—. “Domain differences in expertise.” *Unpublished manuscript.* 2002.

—. “The psychology of experts: an alternative view.” In *Expertise and Decision Support*, by G. Wright and F. Bolger, 11-23. New York, NY: Blenum Press, 1992.

Shanteau, J., and G.J. Gaeth. *Evaluation of the field method of soil texture classification: a psychological analysis of accuracy and consistency.* Technical Report 79-1, Kansas State University, Department of Psychology, 1981.

Shanteau, J., D.J. Weis, R.P. Thomas, and J.C. Pounds. “Performance-based assessment of expertise: how to decide if someone is an expert or not.” *European Journal of Operational Research* 136 (2002): 253-263.

Slovic, P., B. Fischhoff, and S. Lichtenstein. “Regulation of risk: a psychological perspective.” In *Regulatory Policy and the Social Sciences*, by R.G. Noll, 241-277. Berkeley, CA: University of California Press, 1985.

Sunstein, C.R. *Infotopia: How Many Minds Produce Knowledge.* New York, NY: Oxford University Press, 2006.

Surowiecki, J. *The Wisdom of Crowds.* London: Little, Brown, 2005.

Treynor, J.L. “Market efficiency and the bean jar experiment.” *Financial Analysts Journal* 43-3 (1987): 50-53.

Trumbo, D., C. Adams, M. Milnerl, and L. Schipper. “Reliability and accuracy in the inspection of hard red winter wheat.” *Cereal Science Today* 7 (1962).

US Census Bureau. *Annual Estimates of the Resident Population for Incorporated Places of 50,000 or More.* United States Census Bureau, Population Division, 2016a.

—. *Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto RIco.* United States Census Bureau, Population Division, 2016b.

Wallace, H.A. “What is in the corn judge's mind?” *Journal of the American Society of Agronomy* 15 (1923): 300-324.

Weaver, R., and D. Prelec. “Creating truth-telling incentives with the Bayesian Truth Serum.” *Journal of Marketing Research* 50 (2013): 289-302.

APPENDICES

Appendix I: Question order and overview for the US states and bean jars

TABLE A.1 Randomised order of the US state-capital and bean-jar questions

Order	#	State	Order	#	Jar
1	7	New York	2	4	Long drink
3	11	Texas	4	10	Jar 1
5	5	Illinois	6	3	Dessert
7	3	Colorado	8	11	Jar 2
9	8	North Carolina	10	1	Shot
11	4	Florida	12	12	Jar 3
13	9	Ohio	14	5	Wine 1
15	2	California	16	7	Whisky 1
17	1	Arizona	18	8	Whisky 2
19	6	Indiana	20	6	Wine 2
21	12	Washington	22	9	Cognac
23	10	Pennsylvania	24	2	Coffee

TABLE A.2 Overview of the jars and bean weights, numbers, and adjustments

#	Jar / Glass	Bean weight (gr)	Bean number†	Adjusted number	Rounded number	Correct answer
1	Shot	8	89‡	98	100	“False”
2	Coffee	24	255	281	300	“False”
3	Dessert	35	372	335	300	“True”
4	Long drink	93	989	1088	1100	“False”
5	Wine 1	88	936	843	800	“True”
6	Wine 2	187	1989	2188	2200	“False”
7	Whisky 1	103	1096	986	900	“True”
8	Whisky 2	126	1340	1206	1200	“True”
9	Cognac	246	2617	2879	3000	“False”
10	Jar 1	106	1128	1015	1000	“True”
11	Jar 2	118	1255	1381	1400	“False”
12	Jar 3	289	3074	2767	2700	“True”

† The bean numbers are estimated based on the beans’ weight and an average weight of 0.094 grams per beans (based on a weighted sample of one thousand beans).
‡ The bean number of the shot glass is counted instead of estimated. Estimations came to 85 beans.

TABLE A.3 Overview of selected US states, capitals, and largest cities

#	State†	Capital	Largest city‡	Correct answer
1	Arizona	Phoenix	Phoenix	"True"
2	California	Sacramento	Los Angeles	"False"
3	Colorado	Denver	Denver	"True"
4	Florida	Tallahassee	Jacksonville	"False"
5	Illinois	Springfield	Chicago	"False"
6	Indiana	Indianapolis	Indianapolis	"True"
7	New York	Albany	New York	"False"
8	North Carolina	Raleigh	Charlotte	"False"
9	Ohio	Columbus	Columbus	"True"
10	Pennsylvania	Harrisburg	Philadelphia	"False"
11	Texas	Austin	Houston	"False"
12	Washington	Olympia	Seattle	"False"

† States are selected based on population size (US Census Bureau 2016b). Because either true or false statements are maximised at two thirds of the total, three states are discarded despite having larger populations: New Jersey, Virginia, and Washington.
‡ Largest cities are selected based on population size (US Census Bureau 2016a).

Appendix II: Dutch versions of the questionnaire

The questionnaire has the following introduction (in English and Dutch, respectively):

> "Welcome, this questionnaire consists of 24 sections of each 2 questions which test your knowledge and insights. At the end, you can leave your e-mail address for a chance to win one of the four prices of €10. It is important you answer in the questions as well and honestly as possible. Thanks in advance for your participation!"

> "Welkom, deze vragenlijst bestaat uit 24 onderdelen van 2 vragen die je kennis en inzicht testen. Aan het einde kun je je e-mailadres achter laten om kans te maken op een van de vier prijzen van €10. Het is belangrijk dat je de vragen zo goed en eerlijk mogelijk probeert te beantwoorden. Alvast bedankt voor het invullen!"

The bean-jar questions have the following format (in Dutch):

> "Er zitten meer dan 1100 koffiebonen in dit glas. Denk je dat deze uitspraak waar of niet waar is?
> - o Waar
> - o Niet waar"

The US state-capital questions have the following format (in Dutch):

> "De hoofdstad van de Amerikaanse staat Pennsylvania is Philadelphia Denk je dat deze uitspraak waar of niet waar is?
> - o Waar
> - o Niet waar"

The follow-up questions have the following format (in Dutch):

> "Schat in hoe de andere deelnemers bovenstaande vraag beantwoorden. Hoeveel procent denk je dat "Waar" zal antwoorden?
> 0% ---------- 100%"
> *(A slider moves across this line to indicate the chosen percentage.)*

Appendix III: Results of the answers per question and error scores

TABLE A.4 Answers of the group, LST, and CWM methods for US state-capital questions

State	Capital	Largest city	Group answer		LST answer		CWM answer	
			Correct	Pred.	Correct	Pred.	Correct	Pred.
New York	Albany	N.Y. City	Yes	0.64	Yes	0.82	Yes	0.87
Texas	Austin	Houston	No	0.47	Yes	0.51	Yes	0.90
Illinois	Springfield	Chicago	No	0.47	Yes	0.52	Yes	0.93
Colorado	Denver	Denver	Yes	0.87	Yes	0.83	Yes	0.98
N. Carolina	Raleigh	Charlotte	Yes	0.64	Yes	0.63	Yes	0.73
Florida	Tallahassee	Jacksonville	Yes	0.74	Yes	0.64	Yes	0.97
Ohio	Columbus	Columbus	Yes	0.68	Yes	0.71	Yes	0.90
California	Sacramento	Los Angeles	No	0.42	Yes	0.62	Yes	0.79
Arizona	Phoenix	Phoenix	Yes	0.96	Yes	0.94	Yes	1.00
Indiana	Indianapolis	Indianapolis	Yes	0.81	Yes	0.85	Yes	0.86
Washington	Olympia	Seattle	Yes	0.55	Yes	0.56	Yes	0.77
Pennsylvan.	Harrisburg	Philadelphia	No	0.42	No	0.44	Yes	0.71
Total (number or average)			8	0.64	11	0.66	12	0.87

TABLE A.5 Answers of the group, LST, and CWM methods for the bean-jar questions

Jar	Actual beans	Stated beans	Group answer		LST answer		CWM answer	
			Correct	Pred.	Correct	Pred.	Correct	Pred.
Long drink	989	1100	Yes	0.64	Yes	0.65	Yes	0.60
Jar 1	1128	1000	Yes	0.58	Yes	0.60	No	0.45
Dessert	372	300	No	0.32	No	0.28	Yes	0.73
Jar 2	1255	1400	Yes	0.79	Yes	0.77	Yes	0.80
Shot	89	100	Yes	0.91	Yes	0.87	Yes	0.90
Jar 3	3074	2700	Yes	0.72	Yes	0.65	Yes	0.95
Wine 1	936	800	No	0.34	No	0.34	Yes	0.75
Whisky 1	1096	900	No	0.32	No	0.29	Yes	0.77
Whisky 2	1340	1200	No	0.43	No	0.37	Yes	0.93
Wine 2	1989	2200	Yes	0.72	Yes	0.71	Yes	0.54
Cognac	2617	3000	Yes	0.81	Yes	0.82	Yes	0.66
Coffee	255	300	Yes	0.87	Yes	0.83	Yes	0.76
Total (number or average)			8	0.62	8	0.60	11	0.74

TABLE A.6 Predictions of other people's answers for the US state-capital questions

Jar	Group			CWM experts		CWM non-experts	
	Prob.	Pred.	Error	Pred.	Error	Pred.	Error
New York	0.64	0.62	0.03	0.55	0.09	0.69	0.05
Texas	0.58	0.54	0.07	0.51	0.04	0.58	0.11
Illinois	0.32	0.55	0.07	0.55	0.08	0.56	0.09
Colorado	0.79	0.62	0.49	0.64	0.51	0.60	0.47
N. Carolina	0.91	0.48	0.16	0.44	0.20	0.51	0.13
Florida	0.72	0.36	0.37	0.35	0.39	0.42	0.31
Ohio	0.34	0.51	0.19	0.58	0.26	0.51	0.19
California	0.32	0.68	0.26	0.59	0.18	0.75	0.33
Arizona	0.43	0.63	0.59	0.59	0.56	0.65	0.62
Indiana	0.72	0.61	0.42	0.59	0.40	0.61	0.42
Washington	0.81	0.50	0.05	0.50	0.05	0.56	0.01
Pennsylvania	0.87	0.54	0.13	0.45	0.03	0.60	0.18
Average	0.64	0.55	0.24	0.53	0.23	0.59	0.24

TABLE A.7 Predictions of other people's answers for the bean-jar questions

Jar	Group			CWM experts		CWM non-experts	
	Prob.	Pred.	Error	Pred.	Error	Pred.	Error
Long drink	0.64	0.47	0.17	0.47	0.17	0.42	0.23
Jar 1	0.58	0.50	0.09	0.49	0.10	0.48	0.10
Dessert	0.32	0.49	0.19	0.62	0.06	0.39	0.29
Jar 2	0.79	0.43	0.36	0.44	0.35	0.43	0.36
Shot	0.91	0.35	0.56	0.34	0.56	0.30	0.61
Jar 3	0.72	0.61	0.32	0.69	0.41	0.55	0.27
Wine 1	0.34	0.47	0.19	0.61	0.05	0.40	0.26
Whisky 1	0.32	0.50	0.18	0.62	0.06	0.45	0.23
Whisky 2	0.43	0.54	0.03	0.66	0.09	0.44	0.12
Wine 2	0.72	0.46	0.25	0.54	0.18	0.42	0.30
Cognac	0.81	0.42	0.39	0.56	0.25	0.35	0.46
Coffee	0.87	0.34	0.53	0.47	0.39	0.24	0.63
Average	0.64	0.46	0.27	0.54	0.22	0.41	0.32