Judgment and Decision Making, Vol. 6, No. 4, June 2011, pp. 283–294

Repeated judgment sampling: Boundaries

Johannes Müller-Trede*

Abstract

This paper investigates the boundaries of the recent result that eliciting more than one estimate from the same person and averaging these can lead to accuracy gains in judgment tasks. It first examines its generality, analysing whether the kind of question being asked has an effect on the size of potential gains. Experimental results show that the question type matters. Previous results reporting potential accuracy gains are reproduced for year-estimation questions, and extended to questions about percentage shares. On the other hand, no gains are found for general numerical questions. The second part of the paper tests repeated judgment sampling's practical applicability by asking judges to provide a third and final answer on the basis of their first two estimates. In an experiment, the majority of judges do not consistently average their first two answers. As a result, they do not realise the potential accuracy gains from averaging.

Keywords: repeated judgments, judgment accuracy, averaging.

1 Introduction

Imagine you have been asked to make a quantitative judgment, say, somebody wants to know when Shakespeare's Romeo and Juliet was first performed, or you might be planning a holiday in the Alps and are wondering about the elevation of Mont Blanc. An effective strategy to answer such questions is to make an estimate and average it with that of a second judge: a friend, a colleague or just about anybody else (see, for example, Stewart, 2001, or Yaniv, 2004). What, though, if your colleague or friend is unavailable and cannot give you that second opinion? Recent research suggests that you could improve your answer by bringing yourself to make a second estimate and applying the averaging principle to your own two estimates (Herzog & Hertwig, 2009; Vul & Pashler, 2008).

The effectiveness of this suggestion, however, will depend on both the degree to which you are able to elicit two independent estimates from yourself and your willingness to average them. Previous research has focused on the method used to elicit the second estimate. The focus here lies on the type of question being asked, and its interaction with how successive estimates are generated. I report experimental results for different sets of questions which aim to be more representative of quantitative judgments (Brunswik, 1956). I first reproduce previous results which establish the existence of accuracy gains for year-estimation questions such as "In what year were bacteria discovered?" (Herzog & Hertwig, 2009). While I find similar gains for questions about percentage shares (e.g., "Which percentage of Spanish homes have access to the Internet?"), I do not find evidence of accuracy gains for general numerical questions such as "What is the distance in kilometers between Barcelona and the city of Hamburg, in Germany?" or "What is the average depth of the Mediterranean Sea?". I then investigate whether this difference can be explained by the degree to which answers to the various question types are implicitly bounded, but this hypothesis is not supported by the data.

A second factor is whether judges actually recognise the potential gains from averaging and behave accordingly. Larrick and Soll (2006) argue that people often do not understand the properties and benefits of averaging procedures. My experimental data provide further evidence: only a small minority of judges consistently average their estimates. Often, judges settle for one of their first two judgments as the final answer instead or even extrapolate, providing a final answer that lies outside of the range spanned by their first two estimates. They consequently fail to realise the potential gains from averaging.

1.1 Repeated Judgment Sampling

Efficiency gains from averaging are pervasive in different contexts and have been discussed extensively in the literatures on forecasting (Armstrong, 2001), opinion revision (Larrick & Soll, 2006) and group judgment (Gigone & Hastie, 1997). The phenomenon is well-understood: averaging leads to accuracy gains as long as the errors

I would like to thank Shoham Choshem-Hillel, Robin Hogarth and Ilan Yaniv as well as Jonathan Baron, four anonymous referees and seminar participants at Universitat Pompeu Fabra and Hebrew University, Jerusalem, for their suggestions and comments. All remaining errors are my own.

*Graduate Program of Economics, Finance and Management, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005, Barcelona, Spain. Email: johannes.mueller-trede@upf.edu.

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

inherent in the estimates are at least partly independent (Surowiecki, 2004). Vul and Pashler (2008) and Herzog and Hertwig (2009), using different methods to sample multiple judgments from the same judge, found that averaging these also leads to accuracy gains.

In both of these studies, participants were not aware that they would have to answer the same question multiple times and were asked for their first judgment as a best guess. Vul and Pashler (2008) then simply asked the same person to make the same judgment again. They found an accuracy gain when the second judgment followed immediately, but reported a considerable increase in effectiveness if it was delayed for three weeks. Herzog and Hertwig (2009), on the other hand, proposed a method they called dialectical bootstrapping, which presents judges with instructions on how to make the second judgment, asking them to (i) re-consider their first judgment, (ii) analyse what could have been wrong, and specifically, whether it was likely too low or too high, and (iii) make a second estimate based on these considerations (p. 234). Using this method, they obtained larger accuracy gains than without instructions.

Finally, Rauhut and Lorenz (2011) used yet another elicitation method. In their experiment, participants had to provide five answers to the same question and they were informed about this at the outset. They confirmed Vul and Pashler's (2008) and Herzog and Hertwig's (2009) findings of positive accuracy gains from averaging two estimates for four of the six questions they analysed. Furthermore, they found that repeated judgment sampling had diminishing returns: accuracy gains decreased substantially when averaging more than two estimates from the same judge.

1.2 Process and Environment

Vul and Pashler (2008) interpreted their initial finding as evidence for probabilistic representations of concepts in people's minds, but nobody has argued that the mechanism underlying repeated judgment sampling is the same as that leading to accuracy gains when averaging different judges' answers. So far, little is known about how judges generate their different judgments, although some suggestions have been made. Both Vul and Pashler (2008) and Herzog and Hertwig (2009) pointed out the possible role of anchoring-and-adjustment processes, and Rauhut and Lorenz (2011) conjectured that additional judgments may sometimes reflect people becoming emotional or talking themselves into taking wilder and wilder guesses.

A first step toward investigating the processes underlying repeated judgment sampling is to compare its performance in different environments. The experimental study reported below includes different types of questions, including a subset of the year-estimation questions used in Herzog and Hertwig (2009), percentage-share questions, and general numerical questions. I chose the latter two question types because they capture two common types of quantitative judgments judges could face in naturally occurring environments in accordance with representative design (Dhami et al., 2004). In addition, questions about percentage shares are on a response scale which is implicitly bounded between 0 and 100. This allows me to investigate whether the existence of such bounds affects the potential accuracy gains from repeated judgment sampling, as it has been shown to affect performance in other judgment tasks (Huttenlocher et al., 1991; Lee & Brown, 2004).

1.3 Potential and realised gains

A second issue is what judges actually do when asked to provide a third answer on the basis of their first two. This is an interesting question given people's reluctance to employ averaging strategies when combining their own opinion with somebody else's (Soll and Larrick, 2009), and neither Vul and Pashler (2008) nor Herzog and Hertwig (2009) asked judges to actually give a third estimate. In my analysis, I will distinguish between *potential* gains from averaging which I compute by taking the average of the judges' first two answers, and *realised* gains from their third and final estimates. Whether judges are more likely to average when both judgments are their own than when taking advice from somebody else is important for anyone who thinks of using repeated judgment sampling in actual decisions. In addition, how judges manipulate their previous answers in order to arrive at a third one may enable us to infer something about the processes that underlie the generation of estimates.

2 Experimental method and results

I report the results of an experimental study based on a judgment task with two stages. The first stage assesses repeated judgment sampling's performance in the context of different types of questions. It includes three different question types (within-subject) and either provides explicit bounds for the judges or does not (Bounds vs. No-bounds conditions, between-subject). In the second stage, judges are asked to provide a final estimate on the basis of their first two estimates (Self condition). Judges in a control condition are also given the two answers of a different judge, chosen at random from the participants of the experiment (Other condition). Participants were 82 undergraduate students from the subject pool of the Laboratory for Economic Experiments at Universitat Pompeu Fabra, Barcelona. They received an average payment of 8.70 Euro based on the accuracy (median percentage er-

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

ror) of their answers. Participants came from 16 different academic fields of study, and 58% were female.

2.1 Part I: Question type

The first part of the experiment analyses the effect of the question type on potential accuracy gains from repeated judgment sampling. All gains discussed in this section are like those reported in Herzog and Hertwig (2009), computed by taking the average of participants' two estimates, and comparing this average to their first answer. They are not "real" gains, since judges were not asked to provide a third answer themselves until the second part of the experiment. The results reported in this section aim to answer the question whether judges could potentially benefit from the method in different environments.

Method All participants first answered three blocks of twenty questions each (shown in Appendix B). The first block included a sub-sample of the year-estimation questions used in Herzog and Hertwig (2009). It was followed by questions about percentage shares, and the final set of questions consisted of twenty general numerical questions, the answers to which vary by many orders of magnitude. General numerical and percentage share questions were general-knowledge questions, partly sampled from local newspapers.

After completing an unrelated choice task, all participants had to answer the same questions again, in the same order. The elicitation method was adopted from Herzog and Hertwig (2009), and I provided "consider-the-opposite"-type instructions as described above. To further ensure comparability, I also adopted their payment scheme and participants were paid on the basis of the more accurate of the two answers.

Throughout the experiment, subjects in the Bounds condition were also given explicit lower and upper bounds for their answer with each question. For year-estimation items they were told the answer was between 1500 and 1900 and for percentages between 0 and 100. For general numerical questions, the ranges depended on the true unknown value.[1] Subjects in the No-bounds condition did not receive this additional information.

Before the analysis, the data were screened for anomalies. The answers of eight participants, five from the Bounds condition and three from the No-bounds condition, were dropped because they were missing a substantial number of answers. The analyses reported below are based on the answers of the remaining 74 (bounds: 28, no-bounds: 46) participants.

Results Because the distributions of the answers were skewed, the data were transformed to logarithms. Despite this normalisation, the *size* of the effect depends on the response range for each question. Since these differ considerably across question types, I refrain from estimating general models which include a variable for the question type and its interaction with the condition (Bounds vs. No-bounds). Instead, I compute separate regressions according to Equation 1 for each of the three question types.

$$y_{iq} = \alpha + \beta b_i + \delta_i + \theta_q + \epsilon_{iq} \qquad (1)$$

Equation 1 describes a linear regression model with crossed random effects. In this framework, y_{iq} denotes the dependent variable (for the ith individual on the qth question), α is the main effect for gains, β the effect of the explicit bounds provided in the Bounds condition, and δ_i and θ_q denote random effects for individuals and questions, respectively. For each of the three question sets, I estimate five such regressions using different dependent variables, measuring the accuracy of the judges' two estimates and the potential gains judges could obtain from averaging their answers. All of these measures are based on the logarithms of mean absolute deviations of the various estimates from the true value; their algebraic formulae are presented in Table 1.

In Table 1, $x_{1,iq}$ and $x_{2,iq}$ refer to the first and second estimates of judge i for question q, respectively, and x_{tq} refers to the true value for that question. The first two entries in Table 1 are simply logarithms of absolute deviations from the true value.

The bottom three rows in Table 1 describe the different measures for accuracy gains. All three are computed as simple differences in absolute value with respect to the error of the first estimate. A positive coefficient therefore implies an accuracy gain over the first estimate.[2] Second, they are all based on geometric means because of the skew of the answers. For repeated judgment sampling (G_{RJS}) the geometric mean is simply the square root of the product of a judge's two estimates. "Dyadic gains" (G_{Dyad}) can be thought of as the expected accuracy gains from averaging with the estimate of a second participant drawn at random. They are computed as the average of the geometric mean of a judge's first estimate with the first estimate of a second judge. Finally, the estimate from averaging with all other judges at once—the "Wisdom-of-Crowds gain" (G_{WoC}) —is calculated on the basis of the geometric mean across all participants' first answers. It reflects the accuracy gain a participant could achieve by replacing his own estimate by the (geometric) mean of all participants' estimates.

[1] See Appendix B; bounds were constructed so that the distribution of true values with respect to the bounds resembled those of the other two categories.

[2] One could also define analogous measures for accuracy gains with respect to the second estimate. Since there is no difference in accuracy between the two estimates (see Table 2 below), however, I chose to conduct the analyses in comparison to the first estimate only.

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

Table 1: Measures of accuracy and accuracy gain.

Explanation	Formula
Accuracy of 1st estimate	$MAD_{1st} = \left\lvert ln(\frac{x_{1,iq}}{x_{tq}}) \right\rvert$
Accuracy of 2nd estimate	$MAD_{2nd} = \left\lvert ln(\frac{x_{2,iq}}{x_{tq}}) \right\rvert$
Gain from Repeated Judgment Sampling	$G_{RJS} = \left\lvert ln(\frac{x_{1,iq}}{x_{tq}}) \right\rvert - \left\lvert ln(\frac{\sqrt{x_{1,iq}x_{2,iq}}}{x_{tq}}) \right\rvert$
Dyadic Gain	$G_{Dyad} = \left\lvert ln(\frac{x_{1,iq}}{x_{tq}}) \right\rvert - \frac{1}{N-1} \sum_{j \neq i} \left\lvert ln(\frac{\sqrt{x_{1,iq}x_{1,jq}}}{x_{tq}}) \right\rvert$
Gain from Average over all judges	$G_{WoC} = \left\lvert ln(\frac{x_{1,iq}}{x_{tq}}) \right\rvert - \left\lvert ln(\frac{(\prod_i x_{1,iq})^{\frac{1}{N}}}{x_{tq}}) \right\rvert$

Table 2 summarises the results of the analysis. For all three question types, and for both conditions, it provides coefficient estimates for the various accuracy measures discussed. All coefficients reported in Table 2 are significantly different from zero at the one per cent level except for the coefficient for accuracy gains from repeated judgment sampling for general numerical questions (marked by a dagger†), which is not statistically significant.[3]

The results in Table 2 suggest that repeated judgment sampling may not lead to accuracy gains for all types of questions. The first two columns replicate Herzog and Hertwig's (2009) findings: repeated judgment sampling leads to accuracy gains for year-estimation questions, albeit smaller ones than those which can be expected from averaging one's estimate with that of another judge, or other judges. These results are confirmed for questions about percentage shares, and the effect is of similar size: accuracy gains from repeated judgment sampling are between a quarter and a third of the size of Dyadic gains, and between an eighth and a tenth of the size of the accuracy gains obtained from averaging all participants' estimates. For general numerical questions, on the other hand, the picture is different. Averaging with other judges' answers improves accuracy, but there is no evidence of accuracy gains from repeated judgment sampling for these questions. The coefficient estimate for G_{RJS} is .01, which is 24 times smaller than the estimated coefficient for Dyadic gains and is not significantly different from 0 (p=.67).

Next, consider the effect of the bounds. I hypothesized that the difference between year-estimation, percentage share, and general numerical questions was the degree to which answers to these questions were implicitly bounded. The spectrum ranged from percentage share questions with their implicit bounds between 0 and 100 to general numerical questions, which had no obvious bounds associated with them. Year-estimation questions can be thought of as in between the two extremes, given judges' familiarity with the Gregorian calendar. The results in Table 2 suggest that the provision of bounds indeed affects judges' performance differently depending on the question type. They do not support the hypothesis that bounds on the range of possible answers are a sufficient condition for the existence of accuracy gains from repeated judgment sampling, however. As hypothesized, judges' performance on percentage share questions is not affected by the provision of bounds at all. Bounds slightly improve accuracy for year-estimation questions, but do not effect the potential accuracy gains from the different averaging methods. They have a stronger effect on general numerical questions, with a more pronounced improvement in terms of accuracy, and effects on both Dyadic and Wisdom-of-Crowds gains. Note that these latter effects are negative: bounds reduce the accuracy gain which can be expected from averaging (although first answers are more accurate when bounds are provided, so while the improvement is smaller, it is an improvement over a more accurate first answer). Potential accuracy gains from repeated judgment sampling, on the other hand, are not affected either positively or negatively by the provision of bounds.

2.2 Part II: Third Estimates

Method Having completed the first part of the experiment, all participants were asked to make a final judgment for a subset of fifteen questions, five for each question type. Participants in the treatment or Self condition had to make this estimate on the basis of their previous two estimates, while these were displayed on screen. The exact instructions they were given were the following: "For the last time, we would like to present you some of the questions which you have answered during this experiment. On the basis of your previous responses, we would like to ask you for a third answer. For this part of the experiment, you will be paid up to 8 Euros, based

[3] A description of the statistical methods employed to assess the significance of the α and β coefficients is included in Appendix A.

Table 2: Accuracy and potential gains by question type and condition.

	Year-Estimation		Percentage		Numerical	
	No-bounds	Bounds	No-bounds	Bounds	No-bounds	Bounds
MAD_1	.09	.07	.65	*.65*	1.8	.50
MAD_2	.09	.06	.63	*.63*	1.8	.51
G_{RJS}	.003	*.003*	.03	*.03*	.01†	*.01*†
G_{Dyad}	.008	*.008*	.10	*.10*	.24	.05
G_{WoC}	.019	*.019*	.23	*.23*	.61	.16

† All coefficient estimates are significantly different from 0 at the 1% level, apart from the one marked by the dagger which is not significantly different from 0. A coefficient in italics in the bounds condition indicates the absence of a treatment effect, resulting in the same coefficient estimate as in the no-bounds condition.

only on the accuracy of this third and final answer”[4]. The wording of the instructions was chosen so that participants would have no reason to believe that the subset of 15 answers was selected depending on the accuracy of their previous estimates, and to make clear that only accuracy mattered. In order to avoid priming subjects in a mindset which would make them average less, they were told to give the final answer “on the basis of their previous answers”.

Participants in a control condition (Other) had to make the final judgment on the basis of their own two answers as well as the two answers of a different judge chosen at random among the other participants of the experiment. They did not have any information regarding the order of the two judgments from the second judge. Their instructions were similar: “For the last time, we would like to present you some of the questions which you have answered during this experiment. Here, you can see both your own two previous answers and the two answers of another participant of this experiment, who has been chosen at random. On the basis of this information, we would like to ask you for a third answer. For this part of the experiment, you will be paid up to 8 Euros, based only on the accuracy of this third and final answer.”

Of the 82 participants in the experiment, seven were missing a substantial number of answers and had to be dropped. The analysis reported below are on the basis of the answers of 52 participants in the Self condition and 23 participants in the Other condition. Because of software issues, answers to the last question that was asked were not recorded correctly for a large number of participants, so this item was also excluded from the analysis, restricting the analysis to five year-estimation, five percentage-share and four general numerical questions.

Results The data from this part of the experiment can be used to answer two questions: How do judges arrive at their third answer?, and: Are third answers actually more accurate than first answers, as repeated judgment sampling suggests? To preview the findings of the analysis, different judges arrive at their final answers differently, but only a small minority of judges average consistently. Final answers are not significantly more accurate than first (or second) answers, and judges do not realise the potential gains from repeated judgment sampling.

As a starting point for the analyses, assume that judges in the self condition arrive at their third judgment by taking a weighted average of their first two estimates. Denoting by ψ the weight placed on the first estimate, their final estimates can then be expressed as in Equation 2, a framework adopted from the literature on opinion revision (Larrick & Soll, 2009):

$$x_3 = \psi x_1 + (1 - \psi)x_2 \tag{2}$$

The value of ψ can then be calculated separately for each final answer. Note that this method cannot be applied to judges in the Other condition, where the corresponding expression is an equation in three unknowns.[5] The Other condition was included as a standard of comparison for the gains judges realise.

Figure 1 shows the distribution of ψ, aggregated over both questions and participants. From the figure, two assertions can be made about judges’ behaviour.

The first observation is that judges often extrapolate and provide a final answer outside of the range spanned by their first two answers, as indicated by the left- and

[4] Original instructions were in Spanish. They were both written and translated here by the author.

[5] Equation 2 is derived from $x_3 = \psi_1 x_1 + \psi_2 x_2$, under the assumption that $\psi_1 + \psi_2 = 1$. Judges in the Other condition had access to four pieces of information, so that their final answer should be a function of all four of them: $x_3 = \psi_1 x_{1,self} + \psi_2 x_{2,self} + \psi_3 x_{1,Other} + \psi_4 x_{2,Other}$. The assumption $\psi_1 + \psi_2 + \psi_3 + \psi_4 = 1$ is not sufficient to be able to calculate these weights for each item separately.

Figure 1: Aggregate distribution of weight on the first estimate.

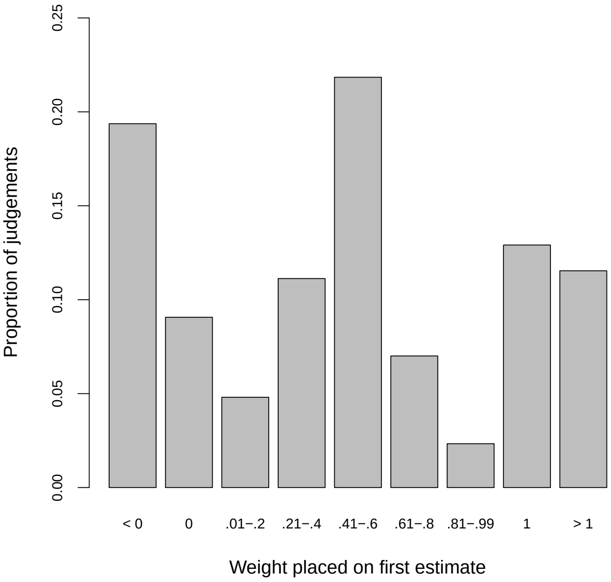
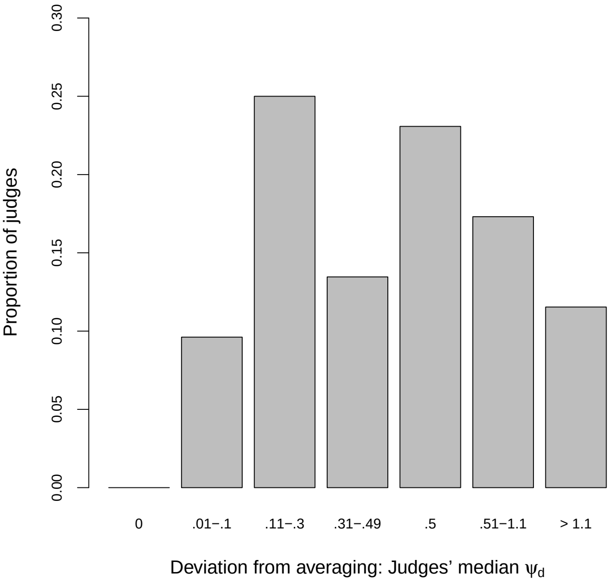


Figure 2: Distribution of judges' tendency to average.



rightmost columns in Figure 1. This constitutes a marked difference from the literature on advice-taking and opinion revision, in which estimates outside of the bounds spanned by one's own estimate and that of the advisor tend to account for less than 5% of answers (Soll & Larrick, 2009; Yaniv & Kleinberger, 2000). In comparison, in the present study such answers account for over 30% of all answers. While in opinion revision, they may be attributed to error and hence disregarded (Yaniv & Kleinberger, 2000), it seems hard to make such an argument in the present case.

A second observation concerns the skew to the left evident in Figure 1. Judges tend to lean more toward their second answer than their first when giving a final answer: 44% of the aggregated judgments lie to the left of the central column in Figure 1, compared to only 33% to its right. This effect is not as strong as the self/other effect in advice-taking (Yaniv, 2004), but in the present context, both answers are one's own. The skew can also be detected in judges' behaviour at the individual level. Comparing the number of questions on which a particular judge uses weights with $\psi < .4$ with the number of questions on which he uses weights with $\psi > .6$, 60% of judges lean more toward their second estimate, and only 33% lean more toward their first.

What else can be said about how individual judges arrive at their final answers? Do they all behave similarly or are there individual differences? In particular, are there judges who consistently average their first two answers? In order to answer these questions, I compute a second measure which is closely related to ψ. For each answer, I calculate how far a judge deviates from taking an average:

$$\psi_d = |\psi - .5|$$

A reliability analysis shows that ψ_d is a reliable measure of individual differences, with a standardised Cronbach's alpha of .85. For each judge, I then calculate the median[6] ψ_d across the 14 questions, the distribution of which is shown in Figure 2. This median characterises judges in terms of how far they deviate from averaging. A median ψ_d smaller than .1 implies that a judge averages on at least 50% of answers; a median larger than .5 implies extrapolation for at least 50% of answers.

Figure 2 shows that around 10% of judges average consistently, resulting in a median ψ_d lower than or equal to .1. It also confirms the importance of extrapolations: 28% of judges exhibit a median ψ_d larger than .5, and therefore extrapolate on more than half of the 14 questions, providing final answers which lie outside of the bounds spanned by their first two estimates. Finally, for almost 25% of judges, the median ψ_d is exactly .5. This does not imply that 25% of judges consistently settle for either of their first answers as their finale estimate, however, as this figure also includes judges who mix strategies and in addition to sometimes providing one of their previous answers as their third answer, average roughly as often as they extrapolate.

What are the implications for the actual accuracy gains judges were able to realise when making their final es-

[6] Possible values of ψ_d range from 0 to infinity, so the median seems to be the more sensible measure of central tendency than the mean.

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

Table 3: Realised- and optimal gains by condition

Condition	$\text{MAD}_{\text{Final}}$	Realised gain	Potential gain
Self	$.32^{***}$	.008	$.03^{**}$
Other	$.30^{***}$	.027	$-.20$

Significance levels: *** 1%, ** 5% level; H_0: coeff. estimate = 0.

timates? Table 3 shows both the *potential* gains from averaging, computed as before by taking the average of the judges' two estimates, and the *realised* gains, that is, the accuracy gain of the third and final answer over the first answer. Since I show above that there are no potential gains for general numerical questions, I conduct this analysis on the basis of the 10 year-estimation and percentage-share questions only.

Table 3 shows that judges in the Self condition were unable to realise significant accuracy gains. Potential gains, on the other hand, are positive for judges in the Self condition, who could have improved their judgment accuracy by simply averaging their previous estimates. Judges in the Other condition were not able to realise any gains, either, but unlike judges in the Self condition, they would not have reliably benefited from averaging. The coefficient estimate for potential gains is estimated at $-.2$, and is not statistically different from zero.[7] This explains why judges in the Other condition are not significantly more accurate in their final judgments than judges in the Self condition as can be seen in the first column of Table 3.

Finally, do realised gains differ between individuals? Maybe judges who average consistently improve in accuracy, while those who extrapolate do not. To answer this question, I correlate the judges' median ψ_d with the average gains they were able to realise for the 10 questions. If judges who average consistently outperform their fellow participants, this correlation should be significantly negative. The analysis does not yield significant evidence that "averagers" do better, however: Spearman's rank correlation coefficient is estimated at $-.13$ (p=.34). A more complex analysis could aim to answer the question at the more disaggregated level of individual answers instead, but a regression analysis with crossed random effects finds no significant effect of ψ_d on the final accuracy gain achieved on a particular question, either. The estimate for the coefficient associated with ψ_d is $-.01$ and fails to reach significance (p=.45).

3 Discussion

In this paper, I provided new evidence that sampling more than one judgment from the same judge and averaging them can lead to accuracy gains in judgment tasks. For a sub-sample of the questions used in Herzog and Hertwig (2009) which ask judges to estimate the year in which a particular event happened, I replicated their finding of potential gains from repeated judgment sampling. I then confirmed this result for a second set of questions in which judges estimate percentage shares. On the other hand, I showed that repeated judgment sampling does not lead to accuracy gains for a third set of general numerical questions. Finally, I reported experimental data on how judges combine their two estimates when asked to do so, a question not previously addressed in the literature. The majority of judges did not consistently average their answers. In the experiment, they failed to realise the potential accuracy gains from repeated judgment sampling.

The finding that accuracy gains from repeated judgment sampling depend on the question being asked constitutes a challenge to the analogy drawn by Vul and Pashler (2008) and Herzog and Hertwig (2009) between repeated judgment sampling and the so-called "Wisdom of Crowds". Accuracy gains from repeated judgment sampling behave like those from averaging different people's estimates for two of the three question sets I examine, but not for the third set of questions. While the source of the accuracy gains in repeated judgment sampling—the averaging principle—is doubtlessly the same as when averaging with somebody else's estimate, how judges generate their successive estimates remains unclear.

While my data fall short of answering this question, they reveal cues about what might be going on in the judges' minds. When asked for a third answer, judges often exhibit a reluctance toward averaging their first two answers, and many of them extrapolate outside of the range spanned by their first two answers. This suggests that they may have thought of more information which could be relevant for the question and which they had not considered when giving their previous estimates. Successive answers could then reflect how judges mentally integrate this cumulative information retrieved from memory to make their judgment. This account of repeated judgment sampling is also consistent with the findings that third estimates lean more toward the second, rather than the first estimates, and that the method does not always emulate the "Wisdom of Crowds". If the variability in the estimates is caused by different pieces of information,

[7] Note that this finding does not contradict the results in Part I according to which potential gains from dyadic averaging are *on average* higher than those from repeated judgment sampling. In Part I, gains from dyadic averaging are expected values over all other participants in the experiment; here, judges were paired at random with another participant. The pairings were such that potential gains were not significant. This in itself is an interesting observation that suggests that gains from repeated judgment sampling may be less variable than gains from dyadic averaging. In terms of comparing realised- and potential gains, however, it defeats the purpose of using the results from the Other condition as a comparative standard for those from the Self condition.

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

judges need at least some knowledge about a question for them to be able to benefit from repeated judgment sampling. On the other hand, even an ignorant judge would benefit from the "Wisdom of Crowds".

This notion is closely related to Rauhut and Lorenz's (2011) hypothesis that question difficulty affects potential accuracy gains, as it predicts no accuracy gains for a hard question that a judge does not know enough about. An interesting issue is what would happen for easy questions judges know a lot about, as these could include professional or expert judgments. Would experts benefit from repeated judgment sampling? In this context, note that my findings also qualify Rauhut and Lorenz's (2011) result that sampling more than two opinions from the same judge is subject to strongly diminishing returns, since all their questions are of the general-numerical type, which are here shown to be the type of questions repeated judgment sampling performs worst on. It is conceivable that, for easy questions, accuracy gains are particularly large, and that returns from sampling more than twice diminish more slowly.

A final consideration concerns the role of the instructions. On the one hand, the effects of Herzog and Hertwig's (2009) "consider-the-opposite" technique, designed to induce judges to give two independent estimates could have persisted longer than intended and influenced final answers in the present experimental setup. This could have contributed to the judges' relutance to average their answers, and also to their tendency to extrapolate. On the other, the finding that only a relatively small minority of judges average consistently has implications for the instructions that would have to be provided, were repeated judgment sampling to be used in decision support. Since judges do no average voluntarily, for the technique to be effective, somebody has to average their judgments for them. That judges should be aware of this when asked for their judgments seems reasonable, even inevitable if a judge were to use the technique more than once. Future work should therefore examine the effects of informing judges about the benefits of averaging *before* eliciting their judgments.

References

Armstrong, J.S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*. Kluwer Academic Publishers.

Baayen, R. H., Davidson, D. J., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Baayen, R. H. (2009). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 1.2. http://CRAN.R-project.org/package=languageR

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

Dhami, M., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959–988.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgment. *Psychological Bulletin, 121*, 149–167.

Herzog, S., & Hertwig, R. (2009). The wisdom of many in one mind. *Psychological Science, 20*, 231–237.

Huttenlocher, J., Hedges, J.V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352–376.

Larrick, R., & Soll, J. (2006). Intuitions about combining opinions: Misappreciations of the averaging principle. *Management Science, 52*, 111–127.

Lee, P.J., & Brown, N.R. (2004). The role of guessing and boundaries on date estimation biases. *Psychological Review & Bulletin, 11*, 748–754.

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: how individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology, 55*, 191–197.

Soll, J. & Larrick, R. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 780–805.

Stewart, T. (2001). Improving reliability of judgmental forecasts. In Armstrong, J. (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, socities, and nations*. Random House of Canada.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*, 645–647.

Yaniv, I. (2004). The benefit of additional opinions. *Current directions in psychological science, 13*, 75-78.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behaviour and Human Decision Processes, 83*, 260–281.

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

Appendix A: Model specification and significance testing

In order to establish the significance of the effects shown in Table 2, I followed a two-step procedure. First, I established whether the experimental manipulation of providing explicit bounds to the participants had an effect on the particular dependent variable y that I was concerned with. Using the *lmer()* class in **R**, I estimated models both of the form given in Equation 1, reproduced here again for convenience, and also of the simpler form shown in Equation 3.

$$y_{iq} = \alpha + \beta b_i + \delta_i + \theta_q + \epsilon_{iq}$$

$$\tilde{y}_{iq} = \alpha + \delta_i + \theta_q + \epsilon_{iq} \tag{3}$$

I then tested for an effect of the bounds, comparing the two estimated models with an F-test. For those models for which I could reject the Null hypothesis at the five per cent level (all tests for which I could reject the Null were also significant at the one per cent level), I have reported the estimated coefficients of the more complex model described in (1). If the Null could not be rejected, the estimates in Table 2 are based on the simpler model (3) instead.

For both types of models, I then proceeded to test the estimated coefficients for significance in the second step of my analysis. Significance testing in models with crossed random effects is not trivial because the distribution of the test statistic is unclear. I employed a Monte-Carlo simulation approach as put forward by Baayen, Davidson and Bates (2008), using the *pvals.fnc()* function provided in the *languageR* package in **R** (Baayen, 2009). Coefficients which were estimated as being larger (or smaller) than zero on more than 99% of simulation runs are reported as significant (or not) at the one per cent level in Table 2. Results are generally based on 10000 simulation runs. In the one case in which a coefficient was at the border of the one per cent significance level, I increased the number of simulations to 100000.

Appendix B: Questions used in the experiment and their associated bounds

Tables 4 to 6 display all sixty questions which were used in the experiment, translated from Spanish, and their respective answers. It also includes the bounds provided to subjects in the Bounds condition. The fifteen questions for which judges had to provide third estimates in the second part of the experiment are indicated with daggers†.

The bounds for the general numerical questions were constructed so that in absolute distance to the closest bounds, the distribution of the true values would resemble those of the other two question types. The mean absolute distance to the closest bound, as a percentage of the distance between the lower- and the upper bound is 0.27 for year-estimation, 0.25 for percentage-share, and 0.28 for general numerical questions. The associated standard deviations are 0.16, 0.13 and 0.13, respectively.

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

Table 4: Year-estimation questions: In what year...

		Bounds	
Question	Answer	Lower	Upper
... was the university of Harvard in Cambridge, MA (USA) founded?	1636	1500	1900
... was the first pocket watch built?	1510	1500	1900
... was the grammophone invented?	1887	1500	1900
... did construction work begin for the Palace of Versailles?†	1661	1500	1900
... were bacteria discovered?	1676	1500	1900
... did Benjamin Franklin invent the lightning conductor?	1752	1500	1900
... was the patent awarded for barbed wire?	1875	1500	1900
... did the plague hit the city of London?	1665	1500	1900
... was electricity discovered?†	1733	1500	1900
... were the 4 concerts for violin 'The 4 Seasons' published?†	1725	1500	1900
... was the thermometer invented?	1592	1500	1900
... was the first fan produced?	1711	1500	1900
... was dynamite invented?†	1866	1500	1900
... did the religious wars begin in France?	1562	1500	1900
... did the English fleet destroy the Spanish Armada?	1588	1500	1900
... was the last woman murdered for witchery in Europe?†	1782	1500	1900
... was Shakespeare's 'Romeo and Juliet' premiered in London?	1595	1500	1900
... was the Bill of Rights passed in England, opening the way for constitiutional monarchy?	1689	1500	1900
... did Louis Braille invent the scripture known as Braille?	1825	1500	1900
... was the first public screening of a film?	1895	1500	1900

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

Table 5: Percentage-share questions: Which (is the) percentage...

Question	Answer	Bounds	
		Lower	Upper
... of the adult population in Spain who smoke on a daily basis?†	27	0	100
... of the Masters students in the Master in Economics at Universitat Pompeu Fabra in 2010 who are foreigners?	85	0	100
... of votes in Catalunya that CiU obtained in the last general elections?	21	0	100
... of its annual income that an average household in Spain spends on alcohol and tobacco?	3	0	100
... of Spanish homes that have access to the Internet?	54	0	100
... of the adult population in Spain that has completed third-level studies?	29	0	100
... of world GDP comes from the USA and the EU combined?	55	0	100
... of the population of Spain that lives in Catalunya?	16	0	100
... of Internet users connect from China?†	21	0	100
... of the 159 'Clasicos' which have been played in the Spanish league that FC Barcelona has won?†	48	0	100
... of the people who live in Barcelona are 65 years old or older?	20	0	100
... of the Spanish population that earned a yearly income of 6000 Euros or less in 2009?†	23	0	100
... of the Spanish population who would prefer to have a business of their own to being a employee, if they had sufficient resources?	40	0	100
... of civil servants who went on strike in the general strike on June 8th 2010, according to the government?	11	0	100
... of final customers who change their tele-com provider do so primarily to save money?	75	0	100
... of women working in Catalunya who are in executive positions?†	7	0	100
... of the time they spend on-line do Spanish Internet users dedicate to social networks?	20	0	100
... of Spanish women who have suffered from domestic violence at least once in their lives?	25	0	100
... of employees in Spain who knew with certainty that they would lose their jobs during the next six months in June 2010?	13	0	100
... of the adult population in Spain who call their mum at least once when they go on a trip?	40	0	100

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press

Table 6: General numerical questions: How many / What is ...

		Bounds	
Question	Answer	Lower	Upper
... underage homeless did the Generalitat have to support in 2009?†	1481	1000	3000
... ZARA stores are there in the city of Barcelona?	12	0	100
... the height of the highest elevation in Montseny, in metres?	1712	1000	3000
... victims (injuries and deaths) did the terror attacks on the Madrid Metro claim in 2004?	2049	1000	3000
... Euros did FC Barcelona pay for new players in the season 2009/2010?	$101.5*10^6$	$50*10^6$	$200*10^6$
... minors between 14 and 17 were detained for drug-use on the street in Barcelona in 2008 and 2009?	1323	1000	3000
... modern Summer Olympics have been celebrated?	29	0	100
... Spanish soldiers are currently deployed in oversea missions?	2600	1000	3000
... Euros of public investment did the 2010 Pressupost of the Generalitat provide for?	$6.177*10^9$	$5*10^9$	$9*10^9$
... the distance in kilometres between Barcelona and the city of Hamburg in Germany?	1815	1000	3000
... calories is the recommended daily intake of an adult woman?	2000	1000	3000
... the life expectancy of a baby born in Spain in 2009?†	80	0	100
... homes will be built in Catalunya with financial support of the Spanish central government in 2012?	1850	1000	3000
... 'municipios' are there in Catalunya?	946	0	1000
... the population of Barcelona?	1615908	0	3000000
... days of rain are there in Barcelona each year on average?	72	0	100
... the average depth of the Mediterranean Sea, in metres?†	1500	1000	3000
... is the speed of sound, in kilometres per hour?	1236	1000	3000
... passengers flew in and out of Barcelona airport in 2009?†	30208134	0	40000000
... Catalunya's GDP per capita in 2008, in Euros?†	29757	0	300000

https://doi.org/10.1017/S1930297500001893 Published online by Cambridge University Press