A JOURNAL OF THE ASSOCIATION FOR PSYCHOLOGICAL SCIENCE aps

Research Article

The Wisdom of Many in One Mind

Improving Individual Judgments With Dialectical Bootstrapping

Stefan M. Herzog and Ralph Hertwig

University of Basel

ABSTRACT—*The "wisdom of crowds" in making judgments about the future or other unknown events is well established. The average quantitative estimate of a group of individuals is consistently more accurate than the typical estimate, and is sometimes even the best estimate. Although individuals' estimates may be riddled with errors, averaging them boosts accuracy because both systematic and random errors tend to cancel out across individuals. We propose exploiting the power of averaging to improve estimates generated by a single person by using an approach we call* dialectical bootstrapping. *Specifically, it should be possible to reduce a person's error by averaging his or her first estimate with a second one that harks back to somewhat different knowledge. We derive conditions under which dialectical bootstrapping fosters accuracy and provide an empirical demonstration that its benefits go beyond reliability gains. A single mind can thus simulate the wisdom of many.*

Forecasting the future has long been believed to be the prerogative of a select few, such as the Oracle of Delphi, Roman augurs, and modern investment gurus such as Warren Buffett. When pooled, however, ordinary people's forecasts about everything from election and sports-event outcomes to the revenues of the next Hollywood blockbuster have enormous predictive accuracy (Wolfers & Zitzewitz, 2004). Moreover, when averaged, forecasts made by experts or by forecasting models about, say, macroeconomic indicators are consistently more accurate than the typical estimate, and are sometimes even the best estimate (e.g., Armstrong, 2001; Clemen, 1989; Timmermann, 2006). Pooling just a few estimates is often sufficient to tap into the power of averaging (e.g., Hogarth, 1978; Johnson, Budescu, & Wallsten, 2001). This phenomenon is known as the *wisdom of crowds* (Surowiecki, 2004).

Thus, the simple prescription for making good forecasts and accurate estimates is as follows: Gather a few predictions or estimates from sources that are likely to differ in their errors (an issue to which we return shortly) and average them (Armstrong, 2001). Sometimes, however, an individual cannot exploit the wisdom of the crowd—for instance, because other people are not available, there is no time for consultation, or rules prohibit communication. Think of the television game show "Who Wants to Be a Millionaire?" which offers contestants increasingly large cash prizes for correctly answering successive and increasingly difficult general-knowledge questions. If a contestant is unsure of an answer, he or she can use one or more "lifelines." One such lifeline is to ask audience members to choose which answer they believe is correct, and the answer receiving the most votes nearly always proves to be correct (Surowiecki, 2004). According to the rules of the game, however, the contestant can use this lifeline only once. How else can he or she exploit the wisdom of the many?

We propose that people can enhance the quality of their quantitative judgments by averaging their first estimate with a second, *dialectical estimate*. Originating from the same person, a dialectical estimate has a different error than the first estimate to the extent that it is based on different knowledge and assumptions. We call this approach to boosting accuracy in quantitative estimation *dialectical bootstrapping*. "Bootstrapping" alludes to Baron Münchhausen, who claimed to have escaped from a swamp by pulling himself up by, depending on who tells the story, his own hair or bootstraps. "Dialectical" refers to the Hegelian process of development, which has three stages: thesis (first estimate), antithesis (dialectical estimate), and synthesis (aggregation). By means of dialectical bootstrapping, the wisdom of crowds can be simulated by a single mind that averages its own conflicting opinions. We now review research on averaging estimates, then outline the dialectical-bootstrapping

Address correspondence to Stefan M. Herzog, Department of Psychology, University of Basel, Missionsstrasse 64A, CH-4055 Basel, Switzerland, e-mail: stefan.herzog@unibas.ch.

Psychological Science RESEARCH, THEORY, & APPLICATION IN PSYCHOLOGY AND RELATED SCIENCES

approach, and finally report an empirical demonstration that dialectical bootstrapping works.

WHEN MORE IS SMARTER

How can a set of individually mediocre estimates become superior when averaged? The secret is a statistical fact that, although well known in measurement theory, has implications that are often not intuitively evident (Larrick & Soll, 2006; Soll, 1999). A subjective quantitative estimate can be expressed as an additive function of three components: the *truth* (the true value of the estimated quantity), *random error* (random fluctuations in the judge's performance), and *systematic error* (i.e., the judge's systematic tendency to over- or underestimate the true value). Averaging estimates increases accuracy in two ways: It cancels out random error, and it can reduce systematic error. This can be illustrated using the concept of *bracketing* (Larrick & Soll, 2006). If two estimates are on the same side of the truth (i.e., do not "bracket" the true value), averaging them will be as accurate, on average, as randomly choosing one estimate. But if two estimates bracket the true value (i.e., one overestimates it and the other underestimates it), averaging the two will yield a smaller absolute error than randomly choosing one of the estimates.

Assume that the true value is 100, and two judges estimate it to be 110 and 120, erring by 10 and 20 units, respectively. Randomly choosing between their estimates gives an expected error of 15, whereas averaging the estimates results in 115, which is also off by 15. Now assume that the second judge's estimate is 80, rather than 120. In this case, although the two judges' estimates have the same absolute errors as before, they lie on opposite sides of the true value. Because the second estimate still errs by 20 units, one can again expect an absolute error of 15 when choosing randomly between the two estimates. Averaging them, however, gives 95, an error of only 5 units!

Averaging, therefore, dominates the strategy of choosing randomly between two estimates: Without bracketing, averaging is as accurate as random choice, and with bracketing, averaging beats random choice. More generally, averaging several estimates reduces overall error as soon as at least one estimate falls on the opposite side of the true value as the others. Bracketing can arise from random error or different systematic errors (i.e., when some judges systematically overestimate and other judges systematically underestimate the true value). Consequently, a low correlation among the errors of a set of judges virtually guarantees bracketing, making the average estimate more accurate than an estimate by a randomly selected judge (Larrick & Soll, 2006). Aggregating a few people's estimates usually suffices to boost accuracy, especially if the people have only modestly correlated errors (e.g., Hogarth, 1978). But what if estimates from other people cannot be used, as when the contestant on "Who Wants to Be a Millionaire?" has already used the audience lifeline?

DIALECTICAL BOOTSTRAPPING

Our thesis is that it is possible to reduce estimation error within one person by averaging his or her first estimate with a dialectical second estimate that is at odds with the first one. As we show shortly, this reduction is larger in magnitude than can be expected by solely reducing random error.

When Does Dialectical Bootstrapping Work?

When is the average of two estimates more accurate than the first estimate? An answer requires comparing errors.[1] For the average of the first and dialectical estimates to beat the first estimate, the dialectical estimate must lie within a *gain range* that asymmetrically extends from truth (see Fig. 1a). The upper boundary of this range is defined by the distance between the first estimate and the true value. If the dialectical estimate lies on the same side of the true value as the first estimate and is farther from the true value, the average of the two estimates will be less accurate than the first estimate. The lower boundary of the gain range is the point on the opposite side of the true value that is 3 times as far from the truth as the first estimate. If a dialectical estimate is located exactly on this boundary, the average of the two estimates will lie exactly as far from the true value as the first estimate, giving equal accuracy. For example, assume that the true value is 100. If the first estimate is 110 and the dialectical estimate is 70 (i.e., identical with the lower boundary), the average will be 90. Both the first estimate and the average have an absolute error of 10 (see Fig. 1b). If the dialectical estimate is 80 (above the lower boundary), however, the average is 95 (with an absolute error of 5) and thus more accurate than the first estimate (see Fig. 1c).

Three analytical observations about the gain range merit attention. First, if the two estimates bracket the true value, the error of the dialectical estimate can be almost 3 times as large as the error of the first estimate and the average will still beat the first estimate. If the estimates do not bracket the true value, the dialectical estimate must be more accurate than the first estimate for the average to win. Consequently, the more probable bracketing is, the larger the error of the dialectical estimate can become before the expected accuracy gain due to averaging becomes negative. Second, regardless of bracketing, if the dialectical estimate has a smaller error than the first estimate, averaging will always improve accuracy. Third, the width of the gain range decreases as the error of the first estimate decreases. Note that this analysis is agnostic regarding the extent to which the errors of the two estimates reflect random versus systematic error. The critical range simply specifies the values of the dialectical estimate for which averaging will improve judgment, given the first estimate. In sum, as long as the errors of the first

[1]This is different from the question of when averaging two estimates is superior to choosing between them (Larrick & Soll, 2006; Soll & Larrick, in press). Our analysis deals with the question of when averaging two estimates is superior to sticking with the first estimate.

and dialectical estimate are nonredundant (i.e., they have nonidentical random or systematic errors) and the dialectical estimate is not too far off the mark, the dialectical average will likely be more accurate than the first estimate (Herzog & Hertwig, 2008).

How to Elicit Dialectical Estimates

How can one elicit a dialectical estimate that is likely to fall in the gain range? We propose that any technique that prompts people to generate the dialectical estimate using knowledge that is at least partly different from the knowledge they used to generate the first estimate can suffice. Retrieving different but plausible information makes it likely that the second estimate will be sufficiently accurate to fall inside the gain range and that its error will be different from that of the first estimate, perhaps even causing the second estimate to fall on the opposite side of the true value, producing bracketing.

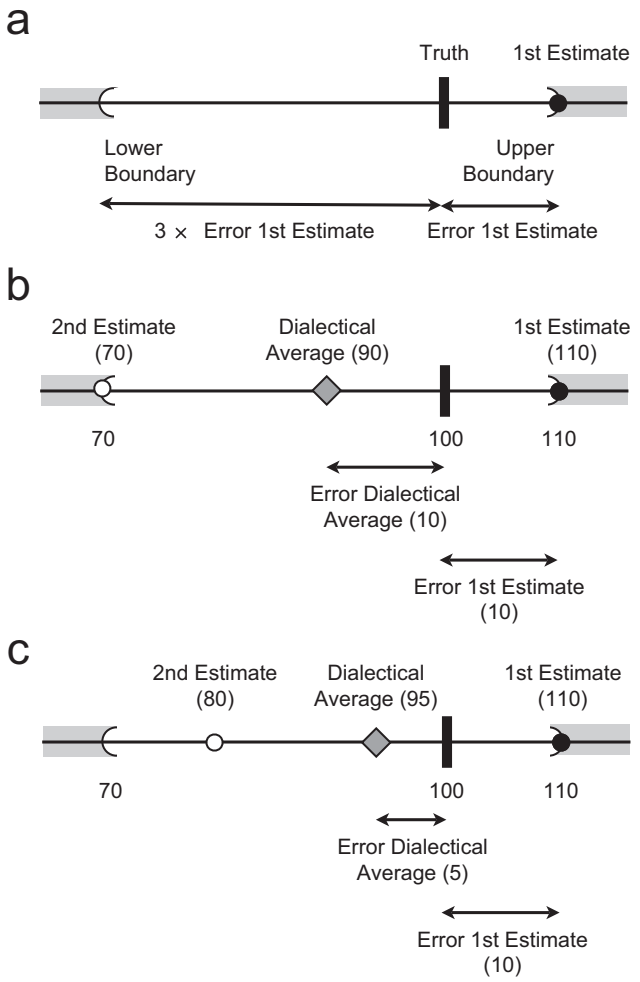

Fig. 1. Gain range: the range of dialectical estimates that will yield an increase in accuracy if averaged with the first estimate. The error of the first estimate defines the upper and lower bound of the range (a). When the dialectical estimate is identical to the lower bound (b), the dialectical average and the first estimate are equally accurate. When the dialectical estimate is within the gain range (c), the dialectical average is more accurate than the first estimate.

This proposal builds upon insights from debiasing research (e.g., Arkes, 1991; Larrick, 2004). Take, for example, the "consider the opposite" technique (Lord, Lepper, & Preston, 1984) and related techniques (e.g., Hirt & Markman, 1995; Hoch, 1985; Koriat, Lichtenstein, & Fischhoff, 1980), which prompt people to consider knowledge that was previously overlooked, ignored, or deemed inconsistent with current beliefs by, for example, asking them to think of reasons why their first judgment might be wrong. Such procedures have been shown to successfully reduce, for instance, overconfidence in confidence intervals (e.g., Soll & Klayman, 2004): Participants who were asked to estimate one boundary ("I am 90% sure that Oscar Wilde was born *after* . . .") and only then the other ("I am 90% sure that Oscar Wilde was born *before* . . .") generated confidence intervals that were better calibrated and more closely centered on the truth than the intervals generated by participants who were asked to produce the intervals in one step. Soll and Klayman (2004) suggested that the stepwise procedure encourages "people to sample their knowledge twice, once for a low estimate and again for a high estimate" (p. 300). Viewed from the perspective of dialectical bootstrapping, the second and first estimates are likely to have a different error. Soll and Klayman argued that their finding "can be thought of as analogous to the improvement in accuracy that is obtained by averaging the opinions of two judges who have nonredundant sources of information" (p. 300). Dialectical bootstrapping offers a potential gain in accuracy by averaging two estimates based on nonredundant knowledge from the *same* judge.

DIALECTICAL BOOTSTRAPPING: DOES IT WORK?

Is dialectical bootstrapping more than a theoretical possibility, and if so, how well does it work? We examined these questions in an empirical study in which participants first gave estimates in response to a set of questions and then generated dialectical estimates. To evaluate whether and to what extent dialectical bootstrapping improves accuracy, we compared the accuracy gains with an upper and a lower benchmark. The lower benchmark was the accuracy gain achieved by averaging the first estimate with a second estimate elicited simply by asking the person to make another estimate (without instruction to generate a dialectical estimate). The average of the two would tend to be more accurate than the first estimate because part of the random error would cancel out (e.g., Stewart, 2001; Vul & Pashler, 2008).[2] If the accuracy gain afforded by dialectical bootstrapping does not surpass this *reliability gain*, there is no reason to

[2]We assume that when asked the same question twice, a person will—unless prompted otherwise—draw on roughly the same knowledge for both estimates. Therefore, the systematic error inherent in his or her estimates will be approximately the same; only random error will vary. In this view, a second estimate is not just a degenerated copy of the first estimate, but rather represents a second draw from an internal probability distribution (cf. Vul & Pashler, 2008).

use a dialectical elicitation strategy for the second estimate. Just asking people to estimate again would be enough.

The upper benchmark was the gain in accuracy achieved by averaging the first estimate with an estimate from another person, as in research on quantitative advice taking (Soll & Larrick, in press; Yaniv, 2004). Because no two individuals are likely to have identical knowledge, it is reasonable to assume that the errors of their estimates are less correlated than the errors of first and dialectical estimates provided by the same person (see Ariely et al., 2000). Averaging a person's first estimate with that of a random person is thus likely to be superior to averaging a person's first estimate with its dialectical counterpart. To what extent can dialectical bootstrapping surpass the mere reliability gain by simulating the process behind this *dyadic gain*?

Method

Participants

Participants ($N = 101$) were students at the University of Basel. For their participation, they received a flat fee of 10 Swiss francs (ca. $9.50 at the time) or course credits, as well as the chance to win one of two iPods in a lottery. Seventy-seven (76%) participants were female; 2 participants failed to report their gender.

Procedure

Using the on-line encyclopedia *Wikipedia*, we created a date-estimation task by selecting 40 historical events (e.g., the discovery of electricity), 10 each from the 16th, 17th, 18th, and 19th centuries. Date-estimation tasks have often been employed in research on estimation and judgment aggregation (e.g., Soll & Klayman, 2004; Yaniv & Milyavsky, 2007). Each participant was randomly assigned to one of two conditions. In both conditions, participants first generated their estimates without knowing that they would be asked later to generate a second estimate. In the *dialectical-bootstrapping* condition, participants ($n = 50$) were then asked to give dialectical estimates (while their first estimates were displayed in front of them) using a technique inspired by the consider-the-opposite strategy:

> First, assume that your first estimate is off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, what do these new considerations imply? Was the first estimate rather too high or too low? Fourth, based on this new perspective, make a second, alternative estimate.

Before rendering their dialectical estimates, participants were informed that the more accurate of the two estimates for each question would be selected and that the chances of winning an iPod would increase as the absolute errors of these "better" estimates decreased. Thus, participants could dare to make a different estimate because only the better of the two estimates would count. This incentive scheme should foster bold second estimates.

In the *reliability* condition ($n = 51$), participants simply made a second estimate. No consider-the-opposite instruction was provided, and first estimates were not displayed while participants made their second estimates. Before making their second estimates, participants were informed that one of the two estimates (first or second) for each question would be randomly selected and that the absolute errors of these selected estimates would determine the chance of winning an iPod. This incentive scheme embodies the aim of the reliability condition, namely, to elicit a participant's "best" estimate on both occasions in order to quantify reliability gains.

We employed two orderings of the 40 items (a random ordering and its reverse). Order of items had no influence on any of the analyses reported here; we therefore pooled the estimates across task orders.[3]

Results

For each participant, we calculated the median absolute deviation between his or her estimates and the actual dates; then, we averaged this measure across participants. In terms of this accuracy measure, first estimates were off by 130.8 years ($SD = 30.7$, $Mdn = 132.5$), repeated estimates were off by 126.5 years ($SD = 32.4$, $Mdn = 131.0$), and dialectical estimates were off by 123.2 years ($SD = 26.8$, $Mdn = 122.5$). In the reliability condition, the first and second estimates for each question were nearly identically accurate, with a mean within-participants difference of 0.4 ($SD = 6.7$; $Mdn = 0.0$; confidence interval, or CI = 0.0–+1.4; $d = 0.06$). In the dialectical-bootstrapping condition, the second estimates were somewhat, but not reliably, more accurate than their respective first estimates (within-participants difference: $M = 4.5$, $SD = 19.6$; $Mdn = 3.0$; $CI = -1.0$–+10.4; $d = 0.23$).

To see whether dialectical bootstrapping pays, we compared accuracy gain in the dialectical-bootstrapping and reliability conditions. The accuracy gain for a participant was defined as the median decrease in error of the average of the two estimates relative to the first estimate, across items. As expected, accuracy in the reliability condition increased as a result of aggregation. However, as Figure 2 shows, this reliability gain was very small, averaging 0.3 percentage point ($SD = 2.3\%$; $Mdn = 0\%$; $CI = 0.0\%$–0.8%; $d = 0.12$). In comparison, dialectical bootstrapping improved accuracy by an order of magnitude: 4.1 percentage points ($SD = 7.8\%$; $Mdn = 3.6\%$; $CI = 2.0\%$–6.4%)—an effect of medium size ($d = 0.53$). Figure 3 shows the

[3]We used robust statistical methods for the statistical analyses (e.g., Wilcox, 2001). We report 20%-trimmed means, the corresponding 95% confidence intervals (percentile bootstrap method; Wilcox & Keselman, 2003), medians, and a robust estimator for the standard deviation (S_n; Rousseeuw & Croux, 1993). All averages of two estimates were rounded, so any superiority of the averages cannot be explained by their being more fine grained than the raw estimates. We report effect sizes using Cohen's d (1988).

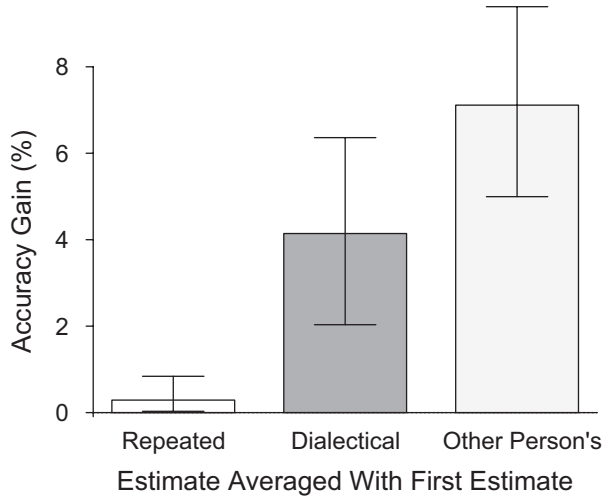

Fig. 2. Accuracy gain obtained by averaging two estimates. The graph shows the mean gain obtained when the original estimate was averaged with the repeated estimate in the reliability condition, when the original estimate was averaged with the dialectical estimate in the dialectical-bootstrapping condition, and when the original estimate of a participant in the dialectical-bootstrapping condition was averaged with the original estimate of a randomly selected other person. Ninety-five percent confidence intervals are shown.

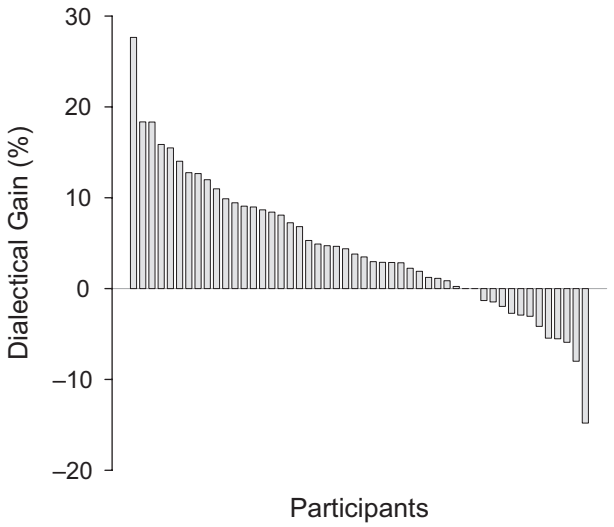

Fig. 3. Distribution of the magnitude of the dialectical gain across participants. Each bar represents a participant's gain (sorted in descending order).

distribution of the dialectical gains across participants. Nearly three fourths of participants (36 of 50, or 72%) benefited from dialectical bootstrapping. Two participants neither benefited nor suffered as a result of the technique. For about one fourth of participants (12 of 50, 24%), accuracy actually decreased.

Although dialectical bootstrapping boosts accuracy, it cannot quite emulate the wisdom of the many. As Figure 2 shows, averaging each person's first estimate for a given item with that of a random other person in the study yielded an average dyadic gain of 7.1 percentage points (SD = 8.3%; Mdn = 6.7%; CI = 5.0%–9.4%)—an effect of large size (d = 0.86).[4] Thus, a person could have achieved a higher accuracy by asking another person for estimates than by using his or her own dialectical estimates (within-participants differences: M = 3.4%, SD = 10.3%; Mdn = 3.5%; CI = −0.2%–+6.2%; d = 0.33). Nevertheless, one's own second, dialectical opinion is worth half the opinion of another judge.

The differences in reliability, dialectical, and dyadic gains were mirrored in the bracketing rates. Merely repeating the estimation process had the lowest bracketing rate, 7.9% (CI = 6.1%–10.2%). In contrast, the first and the dialectical estimates bracketed the true value in 13.6% (CI = 11.1%–16.9%) of cases. Finally, one participant's first estimate and that of another person bracketed the true value in 17.6% (CI = 16.8%–18.5%) of cases.

[4]We simulated the dyadic gain for a question by pairing a given participant's first estimate with the respective first estimates of all other participants, one at a time, thus calculating the expected accuracy gain across the simulated dyads.

What caused the larger gain in the dialectical-bootstrapping condition relative to the reliability condition: the consider-the-opposite strategy or the different incentive scheme? In the dialectical-bootstrapping condition, the better of the two estimates determined the chance to win an iPod, whereas in the reliability condition, a randomly selected estimate did so. To rule out the possibility that the use of the different incentive scheme can account for the key difference found in this study, we ran a new reliability condition that was identical to the previous one except that participants' (N = 50) chance to win was determined by the better of the two estimates (as in the dialectical-bootstrapping condition). The averaging gains were close to those in the original reliability condition and much smaller than the dialectical gains (M = 0.7%, SD = 3.6%; Mdn = 0%; CI = 0.3%–1.7%; d = 0.20). Thus, the method of determining the chance of winning in the dialectical-bootstrapping condition per se did not produce the larger accuracy gains in that condition.

DIALECTICAL BOOTSTRAPPING: A MENTAL TOOL TO FOSTER ACCURACY

Part of the wisdom of the many resides in an individual mind. Our study provides an empirical demonstration that averaging two estimates from the same person—one of which is a dialectical estimate—can improve accuracy beyond mere reliability gains. Although averaging repeated estimates from the same person is known to foster accuracy by reducing random error (e.g., Vul & Pashler, 2008), the notion of averaging two somewhat contradictory estimates from the same person is, to the best of our knowledge, novel. We also outlined the conditions under which averaging two estimates by one person reduces error relative to the person's first estimate alone (thus going beyond the work by, for instance, Einhorn, Hogarth, & Klempner, 1977,

and Soll & Larrick, in press). Moreover, the current approach interprets debiasing strategies such as "consider the opposite" as potential tools to produce judgments with different errors, thus connecting research on debiasing (e.g., Larrick, 2004) with research on judgment aggregation (e.g., Einhorn et al., 1977; Gigone & Hastie, 1997; Soll & Larrick, in press; Yaniv, 2004).

Are there ways of improving on mere reliability gains apart from using the consider-the-opposite strategy employed here? Vul and Pashler (2008) showed that increasing the time delay between two repeated estimates also boosts gains produced by averaging. They reasoned that "temporal separation of guesses increases the benefit of within-person averaging by increasing the independence of guesses" (p. 647), attenuating the anchoring effect of the first estimate. That is, letting time elapse between elicitation of the first and the second estimates also enables use of nonredundant knowledge. The consider-the-opposite strategy, as used here, makes it possible to instantaneously exploit the boost in accuracy gained from asking oneself the same question twice.

Dialectical bootstrapping is a simple mental tool that fosters accuracy by leveraging people's capacity to construct conflicting realities. We do not claim that people spontaneously make use of this tool. Rather, we suggest that after learning about the power of averaging and its key requirement—which can be described on a proximal level as generating good estimates based on different knowledge and on a distal level as producing valid estimates with modestly correlated errors—anyone can benefit from dialectical bootstrapping. Although limited to the domain of quantitative estimates and predictions, this mental tool has a versatility stemming from its general statistical rationale. We do not confine dialectical bootstrapping to the consider-the-opposite strategy. Rather, we suggest that any elicitation procedure that taps into somewhat nonredundant, yet plausible knowledge is potentially capable of eliciting effective dialectical estimates. In fact, we observed dialectical gains similar to the ones reported here in people's forecasts of parliamentary election outcomes. In a study in which people twice predicted the representation of the Swiss political parties that would result from the 2007 election, we found dialectical gains when participants were asked to make predictions both from their own perspective and from the perspective of a dissimilar other person.

The French poet Paul Valéry once said, "Je ne suis pas toujours de mon avis" ("I don't always agree with myself"; Valéry, 1957–1961, p. 760). Vacillating between opinions can be agonizing. But as dialectical bootstrapping illustrates, being of two minds can also work to one's advantage.

Acknowledgments—We thank our research assistants for collecting the data, the members of the Cognitive and Decision Sciences lab at the University of Basel for helpful comments, and Laura Wiles and Valerie M. Chase for editing the manuscript.

REFERENCES

Ariely, D., Au, W.T., Bender, R.H., Budescu, D.V., Dietz, C., Gu, H., et al. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*, 130–147.

Arkes, H.R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, *110*, 486–498.

Armstrong, J.S. (2001). Combining forecasts. In J.S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Norwell, MA: Kluwer Academic.

Clemen, R.T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Einhorn, H.J., Hogarth, R.M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*, 158–172.

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

Herzog, S.M., & Hertwig, R. (2008). *Dialectical bootstrapping: How contradicting yourself can improve your judgment*. Manuscript in preparation.

Hirt, E.R., & Markman, K.D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69*, 1069–1086.

Hoch, S.J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 719–731.

Hogarth, R.M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*, 40–46.

Johnson, T.R., Budescu, D.V., & Wallsten, T.S. (2001). Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making*, *14*, 123–140.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

Larrick, R.P. (2004). Debiasing. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–337). Oxford, England: Blackwell.

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

Lord, C.G., Lepper, M.R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231–1243.

Rousseeuw, P.J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*, 1273–1283.

Soll, J.B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, *38*, 317–346.

Soll, J.B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 299–314.

Soll, J.B., & Larrick, R.P. (in press). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Stewart, T.R. (2001). Improving reliability of judgmental forecasts. In J.S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 81–106). Norwell, MA: Kluwer Academic.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Doubleday.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Amsterdam: North Holland.

Valéry, P. (1957–1961). *Cahiers* (Vol. VII). Paris: Centre National de Recherche Scientifique (CNRS).

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

Wilcox, R. (2001). *Fundamentals of modern statistical methods*. New York: Springer.

Wilcox, R.R., & Keselman, H.J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254–274.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*, 107–126.

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*, 75–78.

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*, 104–120.

(Received 4/17/08; Revision accepted 7/15/08)