Commentary

Psychological Science
24(1) 117–119

DOI: 10.1177/0956797612457399
http://pss.sagepub.com

The Crowd Within and the Benefits of Dialectical Bootstrapping: A Reply to White and Antonakis (2013)

Stefan M. Herzog and Ralph Hertwig
Max Planck Institute for Human Development, Berlin, Germany

Received 7/3/12; Accepted 7/10/12

Can the "wisdom of crowds" (Surowiecki, 2004) be exploited within a single mind? Yes, one can increase accuracy by averaging multiple estimates from the same person (Herzog & Hertwig, 2009; Hourihan & Benjamin, 2010; Müller-Trede, 2011; Rauhut & Lorenz, 2011; Stroop, 1932; Vul & Pashler, 2008; White & Antonakis, 2013; Winkler & Clemen, 2004). We proposed boosting this *crowd-within* effect with what we called *dialectical bootstrapping* (Herzog & Hertwig, 2009; hereafter, H&H): averaging a person's first estimate with his or her second, "dialectical" estimate, derived from knowledge and assumptions different from those motivating the first estimate. A dialectical estimate ideally has an error with a different sign relative to the first estimate—which fosters the chance of error cancellation. There are different ways to elicit a dialectical estimate. We tested one, the consider-the-opposite strategy (Lord, Lepper, & Preston, 1984), and found that averaging first and dialectical estimates improved accuracy more than simply asking people to make an estimate anew and averaging the two estimates (i.e., reliability condition).

White and Antonakis (2013; hereafter, W&A) reanalyzed our data using a different accuracy measure, concluding that "dialectical instructions are not needed to achieve the wisdom of many in one mind" (p. 116). Here, we delineate where we agree and disagree with W&A.

We concur with W&A that the crowd within works. W&A observed (as have we and other researchers) that averaging two estimates from the same person improves accuracy. Moreover, they obtained this result across different measures of accuracy. We also agree with W&A that "dialectical instructions are not needed to achieve the wisdom of many in one mind" (p. 116); in our previous article, we pointed out (H&H, p. 236) that passage of time appears to be enough to boost the gains obtained by averaging (Vul & Pashler, 2008). Additionally, we highlighted that "accuracy in [our] reliability condition increased as a result of aggregation" (p. 234). Our disagreement with W&A concerns the following question: Can dialectical bootstrapping boost the crowd-within effect beyond the gains observed in the reliability condition (i.e., gains expected to occur when averaging any noisy estimates)?

Dialectical Bootstrapping: Does It Have Surplus Value?

We defined the gain obtained by averaging the responses of a given participant as the "median decrease in error of the average of the two estimates relative to the first estimate" (H&H, p. 234). W&A criticized this accuracy change measure. First, they reported that participants' first and second estimates in our reliability condition[1] were, on average, identical in 20% of cases, but that first and second estimates were identical in merely 1% of cases in our dialectical condition. Furthermore, W&A reported that our accuracy change measure was confounded with the proportion of identical first and second estimates. Second, W&A noted that they "prefer to measure accuracy change independently of the proportion of identical first and second responses" (p. 115). Third, they noted that our measure has "awkward statistical properties" (p. 115). Finally, when they used a measure that decoupled accuracy change and the proportion of identical responses, accuracy gains did not differ between our dialectical and reliability conditions. From this, W&A concluded that there is no evidence that "encouraging people to alter their responses more often than they would if not given special instructions yields more accurate average responses" (p. 116).

There are various accuracy (and accuracy change) measures, and opinions about their respective merits differ—because of statistical considerations (Armstrong & Collopy, 1992) or because different measures imply different loss functions (Winkler, 2003). Using a robust measure of accuracy change on the item level, we found that dialectical bootstrapping results in accuracy gains that go beyond reliability gains; using a measure of accuracy change at the participant level, W&A found no such advantage. But how persuasive are W&A's two key reasons to prefer their measure?

First, an alleged weakness of our measure is that it has awkward statistical properties (p. 115)—presumably because it

Corresponding Author:
Stefan M. Herzog, Max Planck Institute for Human Development, Lentzeallee 94, D-14195 Berlin, Germany
E-mail: herzog@mpib-berlin.mpg.de

includes a ratio of two variables (W&A cited p. 22 of Pohl, 2007, where Pohl discussed the "awkward statistical properties" of a "quotient of two variables"). W&A, however, also analyzed two variants of our measure that employed either a more reliable denominator or no denominator, and in both cases, the findings obtained with our original measure were confirmed (see their online Supplemental Material).

Second, W&A conjectured that it is better to measure accuracy change independently of the proportion of identical responses, but what is wrong when a genuine psychological fact—that people hesitate to alter their opinion—enters the accuracy analysis? W&A motivated their independence requirement by reference to hindsight-bias research, in which cases of perfect recall must be separated from cases of reconstruction (e.g., Hoffrage, Hertwig, & Gigerenzer, 2000)—because hindsight bias can occur only in the latter cases. This analogy with the hindsight bias, however, is misleading. The goal of dialectical bootstrapping is to increase accuracy by fostering independence between repeated estimates, and the low proportion of identical estimates indicates that this goal was met. Therefore, we disagree with W&A's stipulation that a measure that gauges accuracy change independently of the proportion of identical responses is preferable.

Conclusion

W&A and we agree that the crowd within works. The question is whether, when, and how dialectical bootstrapping can foster its potential. We are grateful for W&A's comments. Their reanalysis highlights the necessity of including different accuracy measures in future studies and analyzing whether the results obtained using them converge (and if not, why not). When there is no good reason to prefer one measure over others, aggregating them may be one solution to their plurality (Armstrong & Collopy, 1992, p. 75).

Does dialectical bootstrapping improve accuracy beyond mere reliability gains? Clearly, as W&A showed, when one employs the consider-the-opposite strategy (as in H&H), the advantage of dialectical bootstrapping depends on the accuracy measure. This, however, should not be taken as a general verdict on the dialectical-bootstrapping framework, which we explicitly did "not confine . . . to the consider-the-opposite strategy" (H&H, p. 236). There are many ways to leverage "people's capacity to construct conflicting realities" (H&H, p. 236), and thus to achieve dialectical bootstrapping. For instance, we are currently exploring the extent to which averaging different non-Bayesian strategies makes people more Bayesian and the extent to which averaging holistic and analytical judgments improves accuracy. The dialectical-bootstrapping framework poses a wealth of questions, including questions concerning how to design successful and robust dialectical techniques, the ecological conditions under which dialectical bootstrapping pays, and whether people intuitively use this strategy. The work on how and when to poll the crowd in one's head has just begun.

Acknowledgments

The authors thank Laura Wiles for editing the manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

A Swiss National Science Foundation Grant (100014_129572/1) to both authors supported their research discussed in this Commentary.

Note

1. W&A used the term "control condition" when referring to our "reliability condition."

References

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*, 69–80. doi:10.1016/0169-2070(92)90008-W

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*, 231–237. doi:10.1111/j.1467-9280.2009.02271.x

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 566–581. doi:10.1037/0278-7393.26.3.566

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1068–1074. doi:10.1037/a0019694

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. doi:10.1037/0022-3514.47.6.1231

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making*, *6*, 283–294. Retrieved from http://journal.sjdm.org/11/101217a/jdm101217a.pdf

Pohl, R. F. (2007). Ways to assess hindsight bias. *Social Cognition*, *25*, 14–31. doi:10.1521/soco.2007.25.1.14

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. doi:10.1016/j.jmp.2010.10.002

Stroop, J. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology*, *15*, 550–562. doi:10.1037/h0070482

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Garden City, NY: Doubleday.

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

Science, *19*, 645–647. doi:10.1111/j.1467-9280.2008.02136.x

White, C. M., & Antonakis, J. (2013). Quantifying accuracy improvement in sets of pooled judgments: Does dialectical bootstrapping work? *Psychological Science*, *24*, 115–116.

Winkler, R. L. (2003). *An introduction to Bayesian inference and decision* (2nd ed.). Gainesville, FL: Probabilistic.

Winkler, R. L., & Clemen, R. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis*, *1*, 167–176. doi:10.1287/deca.1030.0008