Running head: QUANTIFYING ACCURACY IMPROVEMENT

Quantifying Accuracy Improvement in Sets of Pooled Judgments: Does Dialectical Bootstrapping Work?

Chris M. White & John Antonakis

University of Lausanne

Author note: Address correspondence to John Antonakis, Faculty of Business and Economics, University of Lausanne, CH-1015 Lausanne, Switzerland; e-mail: john.antonakis@unil.ch

Quantifying Accuracy Improvement in Sets of Pooled Judgments: Does Dialectical Bootstrapping work?

Galton (1907) first demonstrated the "wisdom of crowds" phenomenon by averaging independent estimates of unknown quantities given by many individuals. Herzog and Hertwig (2009; hereafter H&H) showed that individuals' own estimates can be improved by asking them to make two estimates at separate times and averaging them.

H&H claimed to observe far greater improvement in accuracy when participants received "dialectical" instructions to consider why their first estimate might be wrong before making their second estimates than when they received standard instructions. We reanalyzed H&H's data using measures of accuracy that are unrelated to the frequency of identical first and second responses and found that participants in both conditions improved their accuracy to an equal degree.

Method

Participants estimated the date of 40 historical events, R_1. In the control condition, they made second estimates, R_2, without special instructions. In the dialectical bootstrapping (DB) condition, they were instructed to think about why R_1 might have been wrong before giving R_2.

Results

For each participant and item, i, H&H subtracted the absolute difference between the average of the participant's two responses, $\bar{R}$, and the true response, T, from the absolute difference between R_1 and T. The median of these 40 difference scores was the participant's accuracy change score, A_{diff}:

$$A_{diff} = Mdn_{i=1}^{i=40}\left(\frac{\left|R_{1,i} - T_i\right| - \left|\overline{R}_i - T_i\right|}{\left|R_{1,i} - T_i\right|}\right) \quad (1)$$

The mean value of A_{diff} was significantly higher in the DB condition, 0.046 (SE = 0.008), than in the control condition, 0.010 (0.008), t (99) = 3.12, p = .002.

H&H did not report a further difference between the conditions that was confounded with A_{diff}. In the DB condition, almost none of the second responses matched the participant's first response to the item, $P(R_1{=}R_2)$ = 0.7% (1.4%). In the control condition, participants' responses matched 20.2% (1.3%) of the time, significantly more than in the DB condition, t (99) = 10.3, p < .001. When R_1 equals R_2 then A_{diff} is 0 and in fact the median A_{diff} in the control condition was 0 for 29 of the 51 participants, causing the mean A_{diff} across all participants to be close to 0. In the DB condition, only two of the 50 participants had a median A_{diff} of 0. The confounding of these two measures is reflected in the significant correlation between $P(R_1{=}R_2)$ and A_{diff}, r (99) = -.307, p = .002.

We prefer to measure accuracy change independently of the proportion of identical responses. Instead of using a median value of the differences in accuracy, like A_{diff}, we analyzed a pair of median accuracy values for each participant. In addition, we used absolute measures of accuracy instead of relative measures because relative measures that have been used in hindsight bias research that have similarities to A_{diff} have "awkward statistical properties" (Pohl, 2007, p. 22).

For each participant, we took a pair of values: The median absolute error of R_1 across the 40 items, A_1, and the median absolute error of $\overline{R}$, A_{avg}:

$$A_1 = Mdn_{i=1}^{i=40}\left(\left|R_{1,i} - T_i\right|\right) \quad (2)$$

$$A_{avg} = Mdn_{i=1}^{i=40}\left(\left|\overline{R}_i - T_i\right|\right) \quad (3)$$

We analyzed this data using a mixed-design ANOVA including the independent variables of response (R_1 vs. $\overline{R}$, within-subject) and condition (DB vs. control, between-subjects). $\overline{R}$ was significantly more accurate than R_1; F (1, 99) = 14.8, $p < .001$, M = 125.6 (4.8) and 130.7 (4.6), respectively. More importantly, the response by condition interaction was not significant, F (1, 99) = 0.00, p = .98: The magnitude of the effect in the DB condition, d = 0.38, was not significantly different from that in the control condition, d = 0.39.

When using these paired accuracy measures (A_1 and A_{avg}), there is no support for the effectiveness of the DB instructions beyond that of the control instructions. Tellingly, the accuracy gain shown by these paired accuracy measures (measured as A_1 - A_{avg}) is unrelated to $P(R_1=R_2)$, r (99) = 0.003, p = .97. A robust Wald test showed that this correlation is significantly lower than that reported above between $P(R_1=R_2)$ and A_{diff}, $\chi^2(1, n=101) = 4.85$, p = .03.

[Insert Table 1 approximately here]

The results for all of these measures are shown in Table 1. Several alternative accuracy measures are included in the online Supplemental Material. Every variation that is similar to A_{diff} in that the median of a set of difference scores is used (whether the value is normalized for item difficulty, as in A_{diff}, or not) is significantly correlated with $P(R_1=R_2)$ and every variation that uses a set of paired accuracy scores for each person, similar to A_1 and A_{avg} (again, whether the values are normalized or not), is not significantly correlated with $P(R_1=R_2)$. Importantly, it is

only the accuracy measures that are correlated with $P(R_1=R_2)$ that show significant differences between the conditions.

Discussion

H&H concluded that the accuracy gained by making a second response was significantly greater for participants in the dialectical bootstrapping than in the control condition. This conclusion was based on an accuracy change measure that is confounded with the proportion of identical first and second responses. Participants in the DB condition were instructed to "assume that your first estimate is off the mark ... make a second, alternative estimate" (H&H, p. 234). Observing a difference between the conditions when using a measure that is confounded with the difference in the proportion of identical responses therefore only serves as a manipulation check.

Using measures that are independent of each other is important in many fields of research. In hindsight bias research, measures of the percentage of perfect recall must be separated from measures of retrieval bias (Pohl, 2007).

When using measures of accuracy that are uncorrelated with the proportion of identical first and second responses the difference between the conditions disappears. People may have some awareness of when they cannot improve upon their first response and in these cases they will only change their response if explicitly instructed to do so. There is no evidence in H&H's data that encouraging people to alter their responses more often than they would do without special instructions yields more accurate average responses. Dialectical instructions are not needed to achieve the wisdom of many in one mind.

References

Galton, F. (1907). Vox populi. *Nature, 75*, 450-51.

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

Pohl, R. F. (2007). Ways to assess hindsight bias. *Social Cognition, 25*, 14-31.

Table 1. Descriptive and Inferential Statistics for Various Dependent Measures.

Condition / Type of Analysis	Measure				
	A_1	A_{avg}	A_1 - A_{avg}	A_{diff}	$P(R_1=R_2)$
Dialectical bootstrapping condition, *M*	130.9	125.8	5.1	0.046	.007
Control condition, *M*	130.5	125.4	5.1	0.010	.202
Response x condition, *F* (1,99)	0.00, $p = .98$				
Condition, *t* (99)				3.12, $p = .002$	10.3, $p < .001$
Correlation with $P(R_1=R_2)$, *r* (99)			-.003, $p = .97$		-.307, $p < .001$

Note: R_1 = first response; R_2 = second response; $\bar{R}$ = average of R_1 and R_2; A_1 = median absolute error of R_1; A_{avg} = median absolute error of $\bar{R}$; A_{diff} = median relative difference between accuracy of R_1 and $\bar{R}$; $P(R_1=R_2)$ = proportion of items for which R_1 and R_2 were identical. Refer to online Supplemental Material for results from one more control condition and additional accuracy measures.

Quantifying Accuracy Improvement in Sets of Pooled Judgments: Does Dialectical Bootstrapping Work?

Supplemantal Material

Chris M. White & John Antonakis
University of Lausanne

To test the robustness of the conclusions in our commentary, we present a more complete set of analyses as Supplemental Material. We include five different accuracy measures (instead of two) and the data from all three of Herzog and Hertwig's (2009; hereafter referred to as H&H) experimental conditions (instead of two).

Measures

In the equations below, $R_{1,p,i}$ is participant p's first response to item i. $\bar{R}_{,p,i}$ is participant p's mean response to item i, and T_i is the correct value for item i.

In our commentary, we referred to H&H's main measure of accuracy improvement as A_{diff}. It is a median *Diff*erence score and each value is *Norm*alized using a measure of item difficulty that is specific to that *Ind*ividual. Because of the number of similar accuracy measures presented in this Supplemental Material, we must alter the labeling system, and so we refer to this accuracy measure as *Diff-Norm-Ind* or A^{DNI}:

$$A_p^{DNI} = Mdn_{i=1}^{i=40}\left(\frac{\left|R_{1,p,i} - T_i\right| - \left|\overline{R}_{p,i} - T_i\right|}{\left|R_{1,p,i} - T_i\right|}\right) \quad \text{(S1)}$$

We referred to the other main accuracy measure in the commentary as the pair of values A_1 and A_{avg}. It is a *P*air*ed* value with each value being *Non-Norm*alized, and so we refer to it as *Prd-Non-Norm*, or A^{PNN}:

$$A_{p,1}^{PNN} = Mdn_{i=1}^{i=40}\left(\left|R_{1,p,i} - T_i\right|\right) \quad \text{(S2a)}$$

$$A_{p,1\&2}^{PNN} = Mdn_{i=1}^{i=40}\left(\left|\overline{R}_{p,i} - T_i\right|\right) \quad \text{(S2b)}$$

Diff-Norm-Ind not only differs from *Prd-Non-Norm* because *Diff-Norm-Ind* is based on difference scores instead of paired accuracy scores, but also because the values are normalized. To investigate whether normalization was important, we used a variation of *Prd-Non-Norm* in which the values were normalized.

To normalize the values, *Diff-Norm-Ind* uses the absolute error observed in the individual's first response, $|R_{1,p,i}-T|$, which we refer to as $D_{p,i}$. It does not make sense to use $D_{p,i}$ when normalizing the paired accuracy data because the first value in the pair would always be 1 and the second value would only differ from *Diff-Norm-Ind* by a constant. We therefore used a more stable measure of item difficulty for each item i, which was the median (across all participants, p) of the absolute difference between each participant's average response and the

true score for that item. We refer to this as the median item difficulty or $D_{\text{mdn,i}}$, the formal definition is:

$$D_{mdn,i} = Mdn_{p=1}^{p=n}\left(\frac{R_{1,i,p} + R_{2,i,p}}{2} - T_i\right) \tag{S3}$$

We call the accuracy measure obtained by taking the *Pai*r*ed* accuracy data and *Norm*alizing using the *Med*i*an* item difficulty value the *Prd-Norm-Mdn* or A^{PNM}:

$$A_{p,1}^{PNM} = Mdn_{i=1}^{i=40}\left(\frac{\left|R_{1,p,i} - T_i\right|}{D_{mdn,i}}\right) \tag{S4a}$$

$$A_{p,1\&2}^{PNM} = Mdn_{i=1}^{i=40}\left(\frac{\left|\overline{R}_{p,i} - T_i\right|}{D_{mdn,i}}\right) \tag{S4b}$$

To determine whether it changes the results substantially when different measures of item difficulty are used, we also analyzed a variation in which a *Diff*erence score was *Norm*alized using the *Med*i*an* item difficulty, we refer to this measure as *Diff-Norm-Mdn* or A^{DNM}. This measure is very analogous to *Prd-Norm-Mdn*.

$$A_{p}^{DNM} = Mdn_{i=1}^{i=40}\left(\frac{\left|R_{1,p,i} - T_i\right| - \left|\overline{R}_{p,i} - T_i\right|}{D_{mdn,i}}\right) \tag{S5}$$

Finally, we also analyzed a *Diff*erence score that was *Non-Norm*alized, which we refer to as *Diff-Non-Norm* or A^{DNN}. This measure is very analogous to *Prd-Non-Norm*.

$$A^{DNN} = Mdn_{i=1}^{i=40}\left(\left|R_{1,p,i} - T_i\right| - \left|\overline{R}_{p,i} - T_i\right|\right) \tag{S6}$$

One might expect that if the difference was taken between any two values in the paired data, then the result should be the same as the difference score, and this would mean that some of the accuracy measures are redundant. However, this is only true if all the values are based on *means*, but because *medians* are used, none of these measures are redundant. It is necessary to use the median(s) to summarize an individual's data in this dataset because of the skewed distributions and extreme outliers that would make the mean values a poor measure of central tendency.

Results and Discussion

The results are shown in Table S1. We present the mean values in each condition, which is either a pair of scores or a single difference score. For the paired scores, we report the

result of the inferential test for the interaction of response type (R_1 vs. $\bar{R}$) and condition (DB vs. control). For the difference scores, we present the result of the main effect of condition. The first inferential test listed, with 1 and 99 degrees of freedom, involves the DB and main control condition. H&H also conducted a second control condition, which is explained in the original manuscript. There are two ways to include the second control condition – either treat the factor of condition as having three levels, or maintain just two levels and combine the data from the two control conditions to increase the power of the test. We present the results using both methods, with the former having 2 and 148 degrees of freedom and the latter method having 1 and 149.

As shown in the table, the results are consistent across all versions of the analyses. The response by condition interaction is not statistically significant for any of the paired accuracy measures, but the effect of condition is statistically significant for all of the difference measures. The conclusions that we present in our commentary therefore do not depend on which version of the analysis is considered most appropriate.

We also correlated the proportion of identical first and second responses, $P(R_1=R_2)$, with each accuracy measure, and the results are given in the bottom row of the table. For the paired accuracy scores, we computed a difference score for each participant based on their paired values; the means of these values in each condition are shown in the columns labeled R_1 - $\bar{R}$. The correlations all include the data from all participants in all three conditions. Once again, the conclusions do not depend on the exact version of the accuracy measure used – there is always a significant correlation between the difference scores and the proportion of identical responses, and no significant correlation between the difference in the paired accuracy measures and the proportion of identical responses.

It is the combination of the use of difference scores, the need to take median values, and the distribution of values within each participant that caused H&H's main measure of accuracy change to be correlated with the proportion of identical first and second responses. This is true for all measures that are based on median difference scores. It is not true for any that yield a pair of median accuracy values.

Table S1. Means and Inferential Statistical using Five Different Accuracy Measures.

	Measure									
	Prd-Non-Norm			Prd-Norm-Mdn			Diff-Non-Norm	Diff-Norm-Mdn	Diff-Norm-Ind	P (R_1=R_2)
	R_1	$\bar{R}$	R_1 - $\bar{R}$	R_1	$\bar{R}$	R_1 - $\bar{R}$				
Dialectical bootstrapping condition	130.9	125.8	5.1	1.077	1.028	0.049	7.36	0.051	0.046	.007
Original control condition	130.5	125.4	5.1	1.035	1.009	0.026	1.34	0.010	0.010	.202
Second control condition	112.5	107.4	5.1	0.953	0.908	0.045	1.78	0.018	0.022	.170
Response x condition, F (1,99)	0.00, p = .98			2.16, p = .15						
Response x condition, F (2,148)	0.00, p = 1.0			1.11, p = .33						
Response x condition, F (1,149)	0.00, p = .99			1.02, p = .31						
Condition, F (1,99)							15.2, p < .001	13.7, p < .001	9.90, p = .002	105, p < .001
Condition, F (2,148)							13.3, p < .001	10.0, p < .001	6.38, p = .002	50.3, p < .001
Condition, F (1,149)							26.6, p < .001	19.4, p < .001	11.6, p = .001	97.2, p < .001
Correlation with P(R_1=R_2), r (149)		-.097, p = .24			-.133, p = .10		-.368, p < .001	- .373, p < .001	-.354, p < .001	

Note. See text for explanation of measures and analyses.