Short Report

Measuring the Crowd Within

Probabilistic Representations Within Individuals

Edward Vul[1] and Harold Pashler[2]

[1]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, and* [2]*Department of Psychology, University of California, San Diego*

A crowd often possesses better information than do the individuals it comprises. For example, if people are asked to guess the weight of a prize-winning ox (Galton, 1907), the error of the average response is substantially smaller than the average error of individual estimates. This fact, which Galton interpreted as support for democratic governance, is responsible for the success of polling the audience in the television program "Who Wants to be a Millionaire" (Surowiecki, 2004) and for the superiority of combined over individual financial forecasts (Clemen, 1989). Researchers agree that this *wisdom-of-crowds* effect depends on a statistical fact: The crowd's average will be more accurate as long as some of the error of one individual is statistically independent of the error of other individuals—as seems almost guaranteed to be the case.

Whether a similar improvement can be obtained by averaging two estimates from a single individual is not, a priori, obvious. If one estimate represents the best information available to the person, as common intuition suggests, then a second guess will simply add noise, and averaging the two will only decrease accuracy. Researchers have previously assumed this view and focused on improving the best estimate (Hirt & Markman, 1995; Mussweiler, Strack, & Pfeiffer, 2000; Stewart, 2001).

Alternatively, initial estimates may represent samples drawn from an internal probability distribution, rather than deterministic best guesses. According to this account, the average of two estimates from one person will be more accurate than a single estimate, so long as the noise contained in the two estimates is at least somewhat independent. Ariely et al. (2000) predicted that such a benefit would accrue from averaging probability judgments within one individual, but did not find evidence of such an effect. However, probability judgments are known to be biased toward extreme values (0 or 1), and averaging should not reduce the bias of estimates; if guesses are sampled from an unbiased distribution, however, averaging should reduce error (variance; Laplace, 1812/1878; Wallsten, Budescu, Erev, & Diederich, 1997).

Probabilistic representations have been postulated in recent models of memory (Steyvers, Griffiths, & Dennis, 2006), perception (Kersten & Yuille, 2003), and neural coding (Ma, Beck, Latham, & Pouget, 2006). It is consistent with such models that responses of many people are distributed probabilistically, as shown by the wisdom-of-crowds effect. However, despite the theoretical appeal of these models, there has been scant evidence that, within a given person, knowledge is represented as a probability distribution. Finding any benefit of averaging two responses from one person would yield support for this hypothesis.

METHOD

We recruited 428 participants from an Internet-based subject pool and asked them eight questions probing their real-world knowledge (derived from *The World Factbook*, Central Intelligence Agency, 2007; e.g., "What percentage of the world's airports are in the United States?"). Participants were instructed to guess the correct answers. Half the participants were unexpectedly asked to make a second, different guess for each question immediately after completing the questionnaire (immediate condition); the other half made a second guess 3 weeks later (delayed condition), also without being given advance notice that they would be answering the questions a second time. It is important that neither group knew they would be required to furnish a second guess, as this precluded subjects from misinterpreting their task as being to specify the two endpoints of a range.

RESULTS

The average of two guesses from one individual (within-person average) was more accurate (lower mean squared error) than either guess alone (see Fig. 1a). In the immediate condition, the error of the average was smaller than the error of the first guess, $t(254) = 2.25$, $p < .05$, and of the second guess, $t(254) = 6.08$, $p < .01$. In the delayed condition, the error of the average was also smaller than the error of the first guess, $t(172) = 3.94$,

Address correspondence to Edward Vul, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave. 46-4141, Cambridge, MA 02139, e-mail: evul@mit.edu.

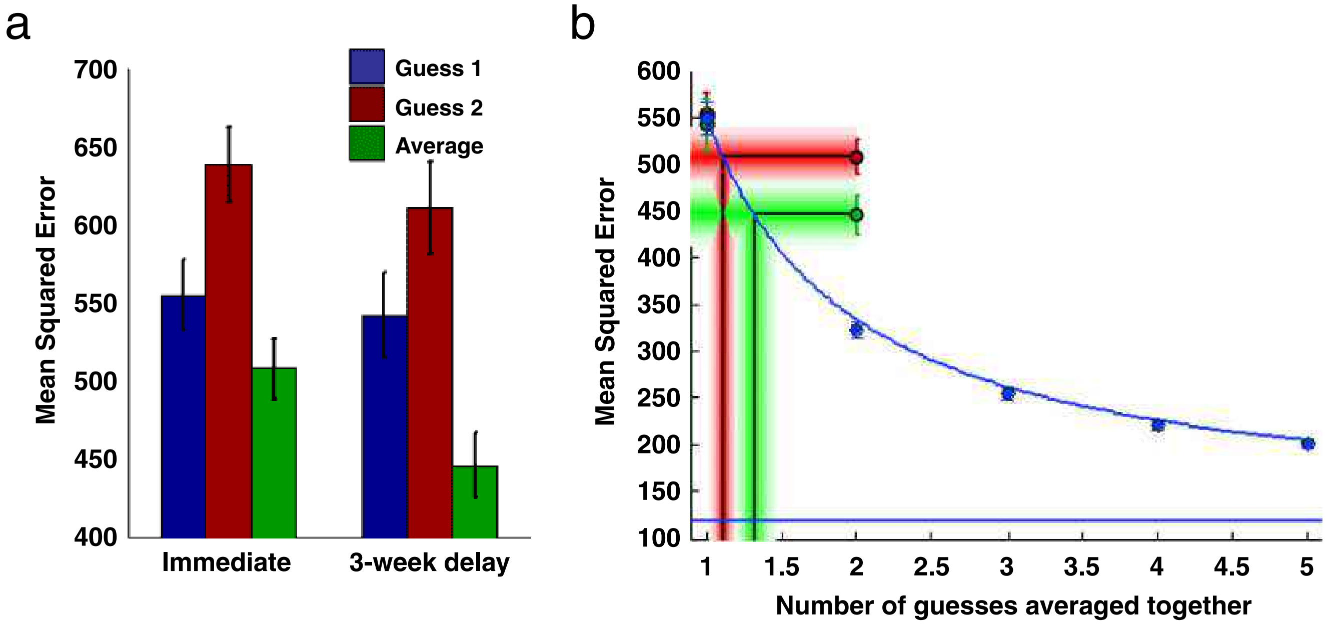

Fig. 1. Experimental results. The bar graph (a) presents mean squared error for the first and second guesses and their average, as a function of condition (immediate vs. 3-week delay). The line graph (b) shows mean squared error as a function of number of guesses averaged together. The data points show results for guesses from independent subjects (blue), a single subject in the immediate condition (red), and a single subject in the delayed condition (green). The blue curve shows convergence to the population bias, which is indicated by the horizontal blue line (the error of the guess averaged across all people). Through interpolation (black lines), we computed the value of two guesses from one person relative to two guesses from independent people, for both the immediate and the delayed conditions. The shaded regions are bootstrapped 90% confidence intervals. Error bars represent standard errors of the means.

$p < .01$, and of the second guess, $t(172) = 6.59$, $p < .01$. This result indicates that subjects did not produce a second guess by simply perturbing the first; rather, the error of the two guesses was somewhat independent. This benefit of averaging cannot be attributed to subjects' finding more information between guesses, because second guesses were less accurate than first guesses (see Fig. 1a) in both the immediate condition, $t(254) = 3.6$, $p < .01$, and the delayed condition, $t(172) = 2.8$, $p < .01$. Moreover, the benefit of averaging was greater when the second guess was delayed by 3 weeks than when it was immediate; that is, the difference in error between the first guess and the average was greater in the delayed condition than in the immediate condition, $t(426) = 2.12$, $p < .05$. The 95% confidence intervals for percentage of error reduced relative to the first guess were [2.5%, 10.4%] in the immediate condition and [11.6%, 20.4%] in the delayed condition. Thus, one benefits from polling the "crowd" within, and the inner crowd grows more effective (independent) when more time elapses between guesses.

We compared the efficacy of within-person averaging and across-person averaging via hyperbolic interpolation (see Fig. 1b). The error of the average guess across all people corresponds to the bias of the distribution of beliefs in the population. According to the central limit theorem, if different subjects' deviations from the group bias are independent, the mean squared error of the average of N guesses from N people should be a hyperbola that converges to the group bias as N goes to infinity. This hyperbola fits the across-person averages perfectly ($R^2 = 1$). However, N guesses from one person are not as beneficial as N guesses from N people. The reduction in mean squared error from averaging N guesses from one person can be described as $1/[1 + \lambda(N - 1)]$, where λ is the proportion of an additional guess from another person that an additional guess from the same person is worth; when λ is 1, averaging in a second guess from the same person confers the same benefit as averaging in a second guess from a different person; when λ is 0, averaging in a second guess from the same person confers no benefit at all. The value of λ can be estimated by interpolating the benefit of within-person averaging onto the hyperbola representing the benefit of across-person averaging. Thus, we computed how many different-person guesses one would need to average together to attain the same error as in the average of two guesses from one person. This value is 1.11 ($\lambda = 0.11$) for two immediate guesses and 1.32 ($\lambda = 0.32$) for two delayed guesses.

Simply put, you can gain about 1/10th as much from asking yourself the same question twice as you can from getting a second opinion from someone else, but if you wait 3 weeks, the benefit of reasking yourself the same question rises to 1/3 the value of a second opinion. One potential explanation of the cost of immediacy is that subjects are biased by their first response to produce less independent samples (a delay mitigates this anchoring effect).

DISCUSSION

Although people assume that their first guess about a matter of fact exhausts the best information available to them, a forced second guess contributes additional information, such that the average of two guesses is better than either guess alone. This observed benefit of averaging multiple responses from the same person suggests that responses made by a subject are sampled from an internal probability distribution, rather than deterministically selected on the basis of all the knowledge a subject has.

Temporal separation of guesses increases the benefit of within-person averaging by increasing the independence of guesses, thus making a second guess from the same person more like a guess from a completely different individual. Beyond having theoretical implications about the probabilistic nature of knowledge, these results suggest that the benefit of averaging two guesses from one individual can serve as a quantitative measure of the benefit of "sleeping on it."

Acknowledgments—This work was supported by the Institute of Education Sciences, U.S. Department of Education (Grants R305H020061 and R305H040108 to H. Pashler) and by the National Science Foundation (Grant BCS-0720375 to H. Pashler; Grant SBE-0542013 to G. Cottrell).

REFERENCES

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

Central Intelligence Agency. (2007). *The world factbook*. Retrieved January 2007 from https://www.cia.gov/library/publications/the-world-factbook/

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

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

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.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*, 150–158.

Laplace, P.A. (1878). Théorie analytique des probabilitiés, Section 2. In *Oeuvres de Laplace* (Vol. 7, pp. 9–18). Paris: Imprimerie Royale. (Original work published 1812)

Ma, W.J., Beck, J.M., Latham, P.E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432–1438.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*, 1142–1150.

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). New York: Springer-Science+Business Media.

Steyvers, M., Griffiths, T.L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*, 327–334.

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

Wallsten, T.S., Budescu, D.V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*, 243–268.

(Received 9/17/07; Revision accepted 1/7/08)