Journal of Experimental Psychology:
Learning, Memory, and Cognition
2010, Vol. 36, No. 4, 1068–1074

0278-7393/10/$12.00 DOI: 10.1037/a0019694

Smaller Is Better (When Sampling From the Crowd Within): Low Memory-Span Individuals Benefit More From Multiple Opportunities for Estimation

Kathleen L. Hourihan and Aaron S. Benjamin
University of Illinois at Urbana-Champaign

Recently, Vul and Pashler (2008) demonstrated that the average of 2 responses from a single subject to general knowledge questions was more accurate than either single estimate. Importantly, this reveals that each guess contributes unique evidence relevant to the decision, contrary to views that eschew probabilistic representations of the evidence-gathering and decision-making processes. We tested an implication of that view by evaluating this effect separately in individuals with a range of memory spans. If memory span is the buffer in which retrieved information is assembled into an evaluation, then multiple estimates in individuals with lower memory spans should exhibit greater independence from one another than in individuals with higher spans. Our results supported this theory by showing that averaging 2 guesses from lower span individuals is more beneficial than averaging 2 guesses from higher span individuals. These results demonstrate a rare circumstance in which lower memory span confers a relative advantage on a cognitive task.

Keywords: working memory, estimation, individual differences

Several recent articles have provided strong evidence that there are probabilistic processes underlying how people estimate values (Herzog & Hertwig, 2009; Vul & Pashler, 2008). Research in the estimation literature has examined estimates about Bernoulli events (e.g., Ariely et al., 2000), estimates of values (Vul & Pashler, 2008) or of dates (Herzog & Hertwig, 2009; Yaniv & Milyavsky, 2007), and even estimations of the strength of relation between two factors (e.g., Hirt & Markman, 1995). Extensive research in these areas has focused on the fact that averaging estimates across individuals improves accuracy relative to individual estimates (e.g., Wallsten & Diederich, 2001), but only recently has research examined the potential benefits of averaging estimates within an individual or how individual differences might affect any potential benefits.

According to one view, estimates of real-world values (e.g., "The area of the USA is what percent of the area of the Pacific Ocean?"; Vul & Pashler, 2008) are made via a retrieval-inference cycle (e.g., Brown, 2002), in which three processes are executed sequentially and iteratively. First, relevant information is retrieved from memory; if the true value can be retrieved, then the process terminates. If the true value cannot be retrieved, then related information is retrieved (the approximate area of the United States, the fact that the Pacific Ocean is the largest ocean, etc.), and this information initiates a *plausible inference* of the value based on this information. Finally, the inference is evaluated for its likely accuracy, and the process either terminates with production of an estimate, or the cycle begins again with retrieval of additional information. In this view, information retrieved from memory serves to set an appropriate range and magnitude for the value to be estimated.

Similarly, real-world judgments may be made via a probabilistic mental model (PMM; Gigerenzer, Hoffrage, & Kleinbölting, 1991). When individuals are faced with a forced choice task (e.g., choosing which of two cities has a larger population), Gigerenzer et al. (1991) proposed that they first attempt to create a local mental model that contains the correct answer, retrieved from memory. If retrieval fails, then a PMM is generated to solve the task with inductive inference. The PMM consists of a reference class of objects (e.g., the two cities in question in addition to other related cities) and a variable set of cues, each with an associated validity. Cues consist of facts relevant to making the choice (e.g., whether a city is a capital, presence of an international airport) and vary in their usefulness for making the choice (cue validity). In making the choice, the individual generates cues and tests them in order of cue validity (with cues that are more likely to be useful generated first); thus, each cue has a certain probability of being tested. When a valid cue is found, the choice is made on the basis of that cue; if no valid cue is found, then the choice is considered to be random.

Research examining how individuals estimate the probability of an outcome suggests that estimates are based on a covert set of processes, yielding a latent variable that is perturbed in some subset of those processes by random error (e.g., Wallsten & Diederich, 2001; Wallsten & González-Vallejo, 1994). According

Kathleen L. Hourihan and Aaron S. Benjamin, Department of Psychology, University of Illinois at Urbana-Champaign.

This research was supported by grant R01AG026263 to Aaron S. Benjamin from the National Institutes of Health. We are grateful to Jonathan Tullis for his assistance with programming.

Correspondence concerning this article should be addressed to Kathleen L. Hourihan, Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel Street, Champaign, IL 61820. E-mail: hourihan@illinois.edu

to Wallsten and González-Vallejo's (1994) stochastic judgment model (SJM), the accuracy with which the validity of a statement is evaluated depends on the statement itself, the strategic manner in which an individual searches memory for knowledge relevant to the statement (cf. Benjamin, 2008), and random variations in how the individual searches memory. The random element in this theory represents the fact that individuals may rely on different underlying knowledge each time a statement is judged and that the judgments will not always be consistent across estimation opportunities. In the SJM account, knowledge about the validity of a statement is represented as a set of relevant evidence and a latent distribution of corresponding values that information is sampled from in a probabilistic manner when a person judges the validity of a fact or estimates a probability. Wallsten and Diederich (2001) have extended this model to account for the improvements in accuracy when judgments of different individuals are averaged.

These theoretical perspectives, all of which emphasize probabilistic aspects of the stages involved in retrieval and decision making, suggest that the opportunity to average over multiple judges or judgments has the potential to increase accuracy. And, indeed, numerous studies have shown that averaging probability judgments or estimates across individuals is beneficial (i.e., reduces error). This result is known as the "wisdom of crowds" effect (e.g., Ariely et al., 2000; Hogarth, 1978; Surowiecki, 2004). When an individual is asked to make a guess about a particular factual value, the average error of the individual's response is greater than the error of the average of the responses of all individuals in the group (e.g., Wallsten, Budescu, Erev, & Diederich, 1997). This occurs when the error of each individual is at least partially statistically independent of the error of other individuals, which will always be the case when some of that error is purely random.

Despite the substantial evidence that averaging estimates across individuals reduces error, there was, until recently, little evidence to suggest that such a benefit could be obtained when averaging estimates within individuals. Ariely et al. (2000) proposed that an analogous *crowd within* effect should obtain and would be evidenced when the average of two estimates from a single individual was more accurate than either single estimate on its own. This would be possible only if estimates were based on information drawn probabilistically (and in a somewhat independent manner) from an internal knowledge set. Although they observed the typical benefit in accuracy when averaging probability estimates across individuals, Ariely et al. found no benefit to averaging repeated estimates within individuals.

However, Ariely et al. (2000) averaged responses within individuals who were providing probability estimates, which generally tend to be biased toward extreme values. Recent studies have detected benefits of averaging within individuals when requiring subjects to estimate non-Bernoulli events, such as values (Vul & Pashler, 2008) or dates (Herzog & Hertwig, 2009). In these tasks, averaging two estimates from the same individual should produce a more accurate value than either estimate on its own to the extent that the estimates are based on probabilistically drawn samples. However, if individuals simply use the best information (such as the most easily simulated outcome; cf. Hirt & Markman, 1995) when estimating values, then averaging two estimates should not provide an improvement over the initial (best) estimate.

Lewandowsky, Griffiths, and Kalish (2009) used an iterated-learning paradigm to demonstrate that individuals have wise crowds within. The subjects in Lewandowsky et al.'s study were required to make a series of predictions in various general-knowledge domains, for values initially chosen by the experimenter, but then randomly selected from a range based on the subject's response on the first trial. For example, subjects were asked to make multiple estimates of the lifespan (i.e., age at time of death) of a male given that that male had already reached a particular age. The set of estimates provided in a given domain was then used to construct group distributions based on the last 10 estimates in a given series. Lewandowsky et al. found that these distributions not only differed depending on the particular domain in question but also closely replicated the shape of the true distribution. Their results indicate that people can appropriately use their knowledge of prior distributions to make a reasonable estimate, given a particular value. It is important that their results also indicate that people provide their estimates based on samples drawn from that known distribution, resulting in the correct distribution shape when estimating repeatedly: wisdom of the crowd within.

In their experiment examining the crowd within, Vul and Pashler (2008) asked subjects to make a guess about a series of world knowledge facts (e.g., "What percent of the world's roads are in India?"), and then unexpectedly asked them to make a second guess about the same facts either immediately or following a 3-week delay. They found that the error of the average of individuals' two guesses was significantly lower than the error of either guess on its own. This finding suggests that the act of estimation involves probabilistic sampling of remembered facts from long-term memory and that repeated estimates each contribute independent information (i.e., unique facts not previously retrieved) to the average. The value of averaging two guesses from the same individual was equivalent to approximately 1.11 (when the second guess was made immediately following the first) or 1.32 (when the second guess was delayed by 3 weeks) guesses from independent estimators.

Working Memory and the Estimation Process

One view of how the estimation process takes place is that individuals sample relevant bits of information (e.g., world knowledge facts) from long-term memory and then compile that information into a single estimate. By this view, the probabilistic aspect of the process lies in the sampling of evidence (i.e., which of many relevant facts are retrieved) rather than noise in the translation between evidence and estimate (cf. Wallsten & Diederich, 2001). Given a particular distribution of knowledge, a single large sample of information will obviously contain more information than a single small sample of information. However, when a second sample is drawn from the same distribution, a small sample is less likely than a large sample to contain bits of information that overlap with the information drawn in the first sample, simply because more information was already sampled the first time with a large sample. Therefore, two smaller samples should be more likely to contain information that is independent from one sample to another, relative to two larger samples, particularly when the distribution of relevant knowledge is relatively small. We reasoned that working memory span (e.g., Conway et al., 2005) could be a

relevant factor in determining the size of the memory sample and consequently influence the relative independence of an individual's two guesses in this task. We propose that individuals with lower memory span should show a greater benefit of averaging, relative to their population, than individuals with a higher memory span. Because low-span individuals' samples should be relatively smaller, they should be less likely to overlap and therefore contribute relatively more independent knowledge to the average of the estimates.

Although it is almost invariably the case that "more is better" in terms of information-processing capacity, we are proposing that low memory-span individuals benefit more from averaging multiple estimates than do high memory-span individuals. Benefits for lower memory span have been proposed in other domains as well, such as language acquisition, decision making, and covariation detection (Hertwig & Todd, 2003). In the realm of covariation detection, Kareev (2000; Kareev, Lieberman, & Lev, 1997) proposed that low-span individuals are able to detect large correlations faster than high-span individuals and to make use of that knowledge sooner. This effect is also based on logic regarding how memory span affects the size of samples one can consider: Because the sampling distributions of correlations are skewed, smaller samples from the population will lead to a more skewed (more extreme) sample correlation than will larger samples. Low-span individuals therefore perceive the correlation as more extreme and are thus able to detect it sooner than high-span individuals. However, this effect and the theoretical motivation underlying it are under debate (e.g., Cahan & Mor, 2007; Juslin & Olsson, 2005; see also Gaissmaier, Schooler, & Rieskamp, 2006, for evidence that low- and high-span individuals may be employing different strategies in correlation detection), so it is uncertain exactly how memory span might affect covariation detection.

Nevertheless, in a similarly counterintuitive vein, we propose that low-span individuals will benefit more from averaging two guesses from the same individual (relative to two guesses from different individuals) than will high-span individuals. It is important to note that we are not predicting any memory span differences in overall accuracy in estimating the values. If one group has privileged knowledge that the other group does not, then we would clearly expect that group to exhibit more accurate estimates in the first place, but such an effect would be outside the purview of our theoretical claim. We make no assumptions about the actual knowledge that each population possesses and can sample from when collecting relevant knowledge to form an estimate, only that the amount of knowledge that can be sampled at one time differs with memory span. We consider below the effects of memory span on overall estimation accuracy, but we focus on the relation between memory span and the relative benefit of averaging two guesses.

We conducted a computer simulation of the effects of memory span size on averaging two guesses. As detailed in Appendix A, this simulation generated two guesses based on sampling a finite memory store. We assumed for the sake of parsimony that the validity of knowledge does not vary systematically with memory span (though it does randomly, across individuals) and that the distributions that govern values drawn from that information base center approximately on the true (to-be-predicted) values. Given these assumptions, it is to be expected that larger samples from that population of knowledge should lead to more accurate single estimates and that lower estimation error should be associated with higher memory span. However, as we demonstrate shortly, the magnitude of such an effect is predicted to be quite small relative to the relation between memory span and the benefit provided by averaging two guesses. This disparity arises because whereas the first correlation includes the noise introduced by individual differences in knowledge, the second correlation—which is based on a within-subject difference score—does not.

The simulation successfully demonstrates that smaller samples do indeed lead to a relatively greater advantage of averaging of two guesses, and this result does not depend on any particular set of assumptions about the relative accuracy of knowledge in individuals with differing memory spans. Specifically, we found that the benefit of averaging (relative to the best single guess) correlated negatively with sample size. Thus, our simulation produced results in line with our prediction: Low-span individuals benefit more from averaging their own two guesses than do high-span individuals.

Method

We asked 181 undergraduates at the University of Illinois to make their best guesses on the same general knowledge questions used by Vul and Pashler (2008). They were unexpectedly asked to "make a second, different guess" for each question during the same 1-hr session. Subjects also completed the Automated Operation Span task (OSPAN; Unsworth, Heitz, Schrock, & Engle, 2005) as a measure of memory span during the same session. Most ($n = 130$) of the subjects made the second guess immediately following the first guess, whereas some subjects ($n = 51$) made the first guess at the beginning of the session and the second guess at the end of the session (with the OSPAN task and another, unrelated task taking place between the guesses). This procedural difference did not have any impact on the data, so it is not examined further. Five subjects were removed because they failed to follow directions; data from a further six subjects were removed because their errors on either Guess 1 or Guess 2 were more than three standard deviations from the mean. The mean square error of the average was computed by first averaging an individual's two raw guesses for a particular question and then computing their mean square error for the averaged guesses across the eight questions (i.e., the mean square error of the average is not simply the average of mean square error for Guess 1 and mean square error for Guess 2).

Results

We replicated the result of Vul and Pashler (2008) that the mean square error of the average of the two guesses from one individual ($M = 484$, $SE = 18$) was lower than the error for either Guess 1 ($M = 502$, $SE = 20$), $t(169) = 2.15$, $p = .03$, or Guess 2 ($M = 565$, $SE = 21$), $t(169) = 7.89$, $p < .001$. Next, on a per-subject basis, we selected the lower mean square error of either Guess 1 or Guess 2 to be that subject's best guess. We then computed the benefit of averaging by subtracting the mean square error of the average of Guess 1 and Guess 2 from the mean square error of the best guess and correlated this benefit with individuals' OSPAN scores. The result was a significant negative correlation ($r = -.16$, $p < .05$), indicating that as memory span increases, the benefit of averaging decreases.

For illustrative purposes, we also conducted a median split based on OSPAN scores to create low- and high-span groups, and calculated the number of between-person guesses to which the average was equivalent for each group (see Appendix B for details). We found that averaging two guesses within a low-span individual is equivalent to averaging approximately 1.07 guesses from different low-span individuals; the same benefit was significantly lower for high-span individuals: only 1.04 guesses.

Although we made no assumptions about whether knowledge might differ at various memory span levels, we previously mentioned the possibility of a relation between memory span and estimation accuracy. That is, larger samples are more likely to yield a mean close to the average of a common distribution than are small samples. This view thus predicts a negative correlation between memory span and mean square error on a single guess. To investigate this possibility in our data, we computed the correlation between memory span and mean square error on a per question basis for Guess 1. This correlation was indeed negative for four of the eight questions, but it was positive for the other four questions, with none of the correlations exceeding .09 in magnitude. The average correlation was slightly negative but not significantly different from 0 ($r = -.01$).

However, it is important to note that the correlation between memory span and mean square error includes relatively large between-subject variance. In contrast, the correlation between memory span and the benefit of averaging capitalizes on relatively small within-subject variance because the benefit is a difference score computed on a per-subject basis. Thus, the observed null correlation may reflect the differential influence of between-subject variance on these two measures. To assess the validity of these ideas, we evaluated the two correlations for the simulated data presented in Appendix A. There one can see that this explanation has merit: correlations between memory span and mean square error are much lower in magnitude than correlations between span and the benefit score. In fact, the actual data values both fall within approximately one standard deviation of the means of their respective simulated distributions.

Finally, we wanted to rule out the possibility that the likelihood of remembering and using the first guess while making the second guess was increased for higher memory-span individuals. We calculated the number of questions for which an individual provided the same response for both Guess 1 and Guess 2 ($M = 1.67$, $SE = 0.15$), and this value did not correlate significantly with OSPAN scores ($r = .08$, $p > .1$). This number also did not differ between span groups based on a median split of OSPAN scores, $t(168) = -0.16$, $p > .1$, suggesting that low- and high-span individuals were equally likely to remember and re-use their first estimate when providing the second estimate.

Discussion

If individuals do possess a crowd within and make estimates of values based on samples probabilistically drawn from a large knowledge base, then the samples are likely to be drawn somewhat independently. It is for this reason that either delaying the second estimate (Vul & Pashler, 2008) or instructing subjects to retrieve alternative information (Herzog & Hertwig, 2009) increases the degree to which the average of two guesses is superior to the first guess. Replicating Vul and Pashler (2008), we have shown this to be the case. Moreover, we predicted that individuals with a lower memory span would enjoy more independent samples across multiple opportunities because the smaller sample size would decrease the amount of overlap across samples.

This hypothesis was supported: Averaging two guesses from individuals with lower memory span was relatively more beneficial than averaging two guesses from individuals with higher memory span. These results fit well with an interpretation of the SJM (Wallsten & González-Vallejo, 1994) in which working memory span controls the size of the sample of information that can be obtained from long-term memory. Assuming that long-term memory contains some finite quantity of relevant facts that is sampled from in a probabilistic manner, larger samples from memory are more likely than smaller samples to contain bits of information that overlap with one another. The second guess should therefore provide less independent information, relative to a single sample, to the estimation process. Note that a straightforward prediction of this view is that a secondary task requirement that burdens working memory somewhat should have a similar effect.

One alternative possibility is that subjects with a higher memory span are more likely to remember their first estimate when making the second estimate. When the individual estimates a second time, the estimation is simply made on the basis of the value provided in the first estimate, rather than resampling relevant information from memory. However, subjects did not know in advance that they would be making a second guess, and it seems unlikely that they were attempting to remember their responses in this manner. Critically, we found that there was no difference between low- and high-span individuals in the likelihood of providing the same value for both guesses (nor did the number of questions for which this occurred correlate with OSPAN score), so it seems highly unlikely that our results can be explained by differential remembering of the first guess.

It is also possible that low- and high-span groups in general do differ in the amount of relevant knowledge from which they can sample when estimating these values. It may be that high-span individuals have superior knowledge (i.e., use different, more valid cues; Gigerenzer et al., 1991), but averaging their estimates suffers from redundancy in sampling caused by the larger samples they may use. In contrast, low-span individuals may have less knowledge, but averaging their estimates benefits from the relatively greater independence in information sampling and therefore results in the observed advantage. It is not possible to conclude definitively from our data whether the two groups share the same quantity or quality of relevant knowledge or whether one group possesses superior knowledge.

Our results support the claim that information gathering from long-term memory is indeed a probabilistic process. Moreover, we have revealed a circumstance in which the generally negative effects of low memory span are partially mitigated by the greater independence of memory sampling that it affords.

References

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., . . . Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6,* 130–147. doi:10.1037/1076-898X.6.2.130

Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (Vol. 48, pp. 175–223). Lon-

don, England: Academic Press. doi:10.1016/S0079-7421(07)48005-7 UID 2008-16272-005

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 41, pp. 321–359). New York, NY: Academic Press.

Cahan, S., & Mor, Y. (2007). The effect of working memory capacity limitations on the intuitive assessment of correlation: Amplification, attenuation, or both? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 438–442. doi:10.1037/0278-7393.33.2.438

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12,* 769–786.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180. doi:10.1037/0003-066X.60.2.170

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *SIAM CBMS-NSF Monograph, 38.*

Gaissmaier, W., Schooler, L. J., & Rieskamp, J. (2006). Simple predictions fueled by capacity limitations: When are they successful? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 966–982. doi:10.1037/0278-7393.32.5.966

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528. doi:10.1037/0033-295X.98.4.506

Hertwig, R., & Todd, P. M. (2003). More is not always better: The benefits of cognitive limits. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making* (pp. 213–231). Chichester, England: Wiley. doi:10.1002/047001332X.ch11

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

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. doi:10.1037/0022-3514.69.6.1069

Hogarth, R. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21,* 40–46. doi:10.1016/0030-5073(78)90037-5

Juslin, P., & Olsson, H. (2005). Capacity limitations and the detection of correlations: Comment on Kareev (2000). *Psychological Review, 112,* 256–267. doi:10.1037/0033-295X.112.1.256

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review, 107,* 397–402. doi:10.1037/0033-295X.107.2.397

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126,* 278–287. doi:10.1037/0096-3445.126.3.278

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science, 33,* 969–998. doi:10.1111/j.1551-6709.2009.01045.x

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.* New York, NY: Random House.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37,* 498–505.

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

Wallsten, T. S., Budescu, D. V., Erev, L., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 243–268. doi:10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences, 41,* 1–18. doi:10.1016/S0165-4896(00)00053-6

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101,* 490–504. doi:10.1037/0033-295X.101.3.490

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes, 103,* 104–120. doi:10.1016/j.obhdp.2006.05.006

Appendix A

Simulation of the Estimation Process

The process by which we simulated the sampling and estimation procedure has three distinct parts. Each simulated individual (total $N = 170$; 34 per span group) i had a finite distribution of knowledge (20 values) sampled from a Gaussian distribution with mean μ_i and standard deviation σ_i. Values in this information set were capped at 0 and 100. The parameters μ_i and σ_i for an individual were sampled from two Gaussian parent distributions with means of 50 and 30 and standard deviations of 36 and 10, respectively.

Once an individual's set of knowledge was established, an estimate of the true value, 50, was provided by sampling s values from that set of knowledge, where s was set to 2, 3, 4, 5, or 6 to represent a range of working memory spans. These values were selected to approximate the range of error-free set sizes completed on the OSPAN task (Unsworth et al., 2005) by our real subjects. The average of the sampled values provided the first estimate. Seven additional samples were drawn to simulate estimating eight values. The sampling procedure was then repeated for the second estimate (with replacement). The entire simulation process was replicated 1,000 times to obtain a distribution of simulated data. We then compared our real data against this distribution to assess its likelihood of having been generated under the assumptions of our simulation. Values reported are the average value across these 1,000 simulations.

As predicted, the mean square error of the average ($M = 766$) was lower than that for either Guess 1 ($M = 824$) or Guess 2 ($M = 825$). Next, the benefit of averaging relative to an individual subject's best guess (either Guess 1 or Guess 2, whichever produced the lowest mean square error) was correlated with the sample size. A negative correlation ($r = -.12$) was obtained, indicating that higher spans benefit less from averaging than do lower spans.

Finally, we compared our real data to the distributions of simulated data. We first assessed the likelihood of our observed correlation between span and benefit of averaging of $-.16$ (see the Results section) being generated by the simulation. In the distribution of 1,000 correlations simulated, $-.16$ falls approximately in the 30th percentile (see Figure A1, left panel). This indicates that our simulation reasonably replicates the observed relation between benefit of averaging and memory span.

We also examined the correlation between sample size (memory span) and mean square error, averaged across the eight questions. As stated previously, one might expect higher span individuals to have access to more relevant knowledge than lower span individuals, and therefore memory span should correlate negatively with estimation error (i.e., higher spans tend to be associated with lower error). In our simulated data, the correlation between sample size and mean square error was indeed found to be negative ($r = -.08$). In the distribution of 1,000 correlations simulated, our observed correlation of $-.01$ (see the Results section) falls approximately in the 85th percentile (see Figure A1, right panel), again indicating that our simulation reasonably replicated the observed data.

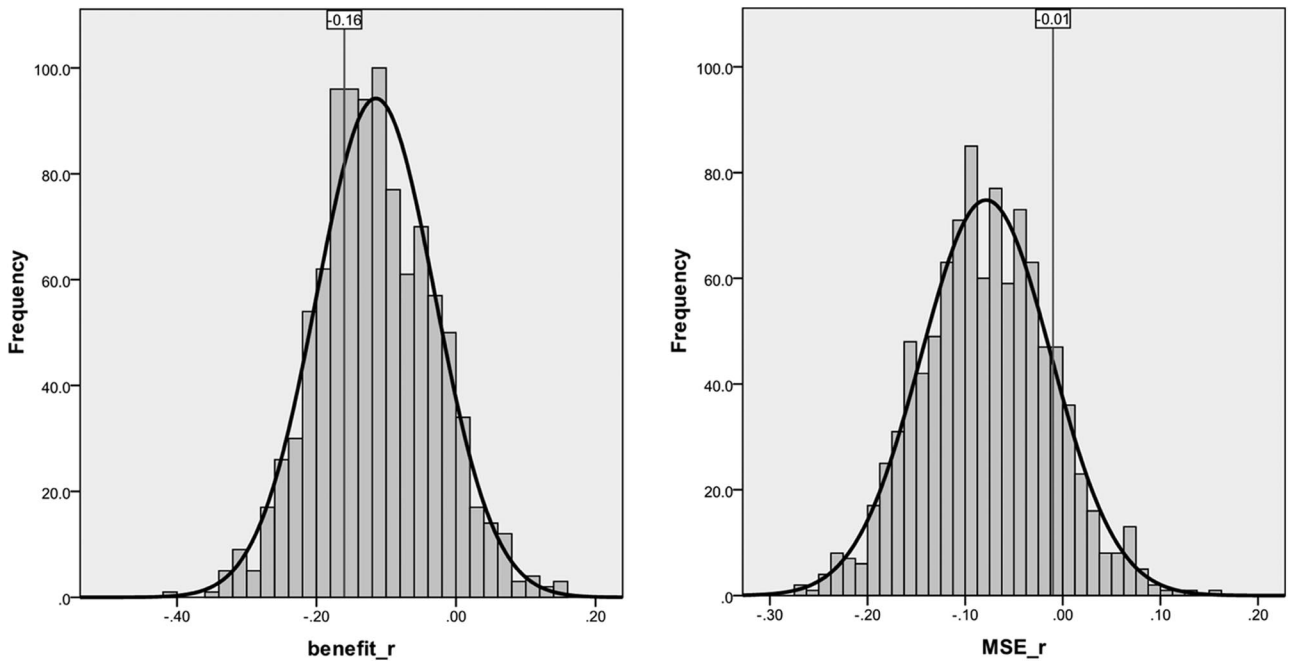

Figure A1. Histograms of the distributions of simulated correlations between sample size and the benefit of averaging (left panel) and between sample size and mean square error (right panel). The correlations obtained from the real data are displayed as reference points.

(*Appendices continue*)

Appendix B

Value of Averaged Guesses for Low- and High-Span Groups

Table B1
Mean Square Errors (MSEs) for Guess 1, Guess 2, and Averaged Guesses, and Derived Value of Averaged Guesses, for Low and High Memory Span Groups

Value	Low-span group	High-span group
Guess 1 *MSE*	483	521
Lower bound 95% CI	430	464
Upper bound 95% CI	535	578
Guess 2 *MSE*	540	591
Lower bound 95% CI	485	530
Upper bound 95% CI	594	653
Average *MSE*	460	508
Lower bound 95% CI	411	455
Upper bound 95% CI	509	561
Estimated value of average	1.07	1.04
Lower bound 95% CI	1.05	1.02
Upper bound 95% CI	1.09	1.06
Function parameters		
C	152.14	173.04
a	-6.30×10^{-5}	-1.95×10^{-4}
b	3.09×10^{-3}	3.06×10^{-3}

We based our analytic procedure on that used by Vul and Pashler (2008), which acts to reparameterize the mean square error for two within-person guesses as an equivalent number of between-person guesses. Low-span ($n = 86$) and high-span ($n = 84$) groups were based on a median split of absolute OSPAN score. These two groups did not differ significantly on Guess 1 or Guess 2, or on the average (all ps $> .18$). The group mean square errors are displayed in Table B1.

First, values of average mean square error across a range of between-person guesses were computed by taking all n-tuples ($n = 1$ to 5) from the Guess 1 population and computing the mean square error for the average of those guesses. The average for all n-tuples for a given n provides the estimate for the mean square error for the average of n independent guesses. As argued by Vul and Pashler (2008), the relation between n and MSE_n is assumed to be hyperbolic with a nonzero asymptote:

$$MSE_n = C + \frac{1}{bn_g + a}, \quad (1)$$

where a and b are free parameters and C was fixed for each memory span group by averaging Guess 1 across all individuals in that group (i.e., the mean square error across, rather than within, individuals). This function was estimated separately for each memory span group using least squares estimation.

Algebraic manipulation yields the equation to solve for the number of between-person guesses (n_g) to which the average of two within-person guesses is equivalent:

$$n_g = \frac{1}{b\,(MSE_{average\_within} - C)} - \frac{a}{b} \quad (2)$$

This value was estimated for each group on the basis of the parameters estimated for the two memory span groups. As described above, the value of two within-person guesses was estimated as the number of between-person guesses that would yield an equivalent mean square error for that population. Data points for the function were generated by taking all possible pairs, triplets, quadruplets, and quintuplets of subjects and evaluating the average error displayed by this combination of one to five subjects (using their Guess 1 values). Hyperbolic functions fit the data (i.e., the estimated mean square errors for the across-subject averaging of samples of size 1 to 5 for the two populations) nearly perfectly, as they had in the Vul and Pashler (2008) data ($R^2 = 1.00$ and .998 for the low- and high-span groups, respectively). Parameters for the functions are indicated in Table B1. The resulting values were 1.07 and 1.05 for the low- and high-span groups, respectively.[1] These values represent the number of guesses from different people in that population that would need to be averaged together to obtain the same error as the averaged within-person guesses.

To determine whether the observed value of averaging two guesses was reliably larger for the low-span group than for the high-span group, we derived sampling distributions for the two groups' value scores by bootstrapping (Efron, 1982). In our procedure, we drew 1,000 random samples, with replacement, of size n (where n is equal to the number of subjects in that group), of subjects from each memory span group. For each of the 1,000 samples, we computed that sample's estimated value of averaging (n_g in Equation 2) using the appropriate parameters (a, b, and C; see Table B1) for that group and the sampled subjects' actual mean square errors for Guess 1, Guess 2, and the average, resulting in a sampling distribution of the estimated value of averaging within-person (relative to averaging across-person) for both of the memory-span groups. We then computed 95% confidence intervals for the estimated value of average scores using the variance of this bootstrapped distribution. As can be seen in Table B1, the two confidence intervals overlap by only 16%. Viewed in terms of significance testing, the two means can be considered to be significantly different, with $p < .05$ (cf. Cumming & Finch, 2005). Individuals in the low-span group benefited more from adding a second guess to their average, relative to their population, than did individuals in the high-span group.

[1] We also computed the value of averaging two guesses from the same individual using Vul and Pashler's (2008) linear interpolation method. We found estimated values to be 1.17 for the low-span group (95% CI [1.15, 1.19]) and 1.14 for the high-span group (95% CI [1.12, 1.16]).

Received August 31, 2009
Revision received March 15, 2010
Accepted March 22, 2010 ■