Atten Percept Psychophys (2012) 74:810–815
DOI 10.3758/s13414-012-0293-0

Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli

Alice R. Albrecht · Brian J. Scholl · Marvin M. Chun

Published online: 8 May 2012

Abstract Beyond perceiving the features of individual objects, we also have the intriguing ability to efficiently perceive average values of collections of objects across various dimensions. Over what features can perceptual averaging occur? Work to date has been limited to visual properties, but perceptual experience is intrinsically multimodal. In an initial exploration of how this process operates in multimodal environments, we explored statistical summarizing in audition (averaging pitch from a sequence of tones) and vision (averaging size from a sequence of discs), and their interaction. We observed two primary results. First, not only was auditory averaging robust, but if anything, it was more accurate than visual averaging in the present study. Second, when uncorrelated visual and auditory information were simultaneously present, observers showed little cost for averaging in either modality when they did not know until the end of each trial which average they had to report. These results illustrate that perceptual averaging can span different sensory modalities, and they also illustrate how vision and audition can both cooperate and compete for resources.

Keywords Statistical summary representations · Perceptual averaging · Ensemble perception · Audiovisual interaction

The goal of perception is to represent our immediate environments so that we can more efficiently and adaptively interact with them. A central challenge of perception is that the complexity of our environments vastly outstrips the ability of the human mind to process them. For example, it is not possible for us to arrive at high-resolution representations of every part of the visual field at a glance, even when that information is not changing from moment to moment. There are at least two general strategies that the mind appears to take in order to cope with this challenge. First, we can build high-resolution representations of only some of the immediate environment, via attentional selection (for a review, see Chun, Golomb, & Turk-Browne, 2011). Second, we can build lower resolution representations of all or most of the environment—for example, by the "quick and dirty" processing of visual scene gist (for a review, see, e.g., Oliva & Torralba, 2006). Recently, this second strategy has also been recognized in a new line of research that is focused on the phenomenon of perceptual averaging and the construction of "statistical summary representations."

A. R. Albrecht (✉) · B. J. Scholl · M. M. Chun
Department of Psychology, Yale University,
Box 208205, New Haven, CT 06520-8205, USA
e-mail: alice.albrecht@yale.edu

Perceptual averaging

Beyond perceiving the features of individual objects, we also have the intriguing ability to efficiently perceive *average* values of collections of objects across various dimensions. For example, when faced with a display of discs of varying sizes, observers are able to accurately report their average size, even while unable to judge whether particular individual sizes are present (Ariely, 2001).

This ability also persists when the display is presented only briefly (down to 50 ms), and with considerable variation in the number and density of discs, and the types of distributions from which their sizes are drawn (Chong & Treisman, 2003, 2005b; but cf. Whiting & Oriet, 2011). Moreover, this ability is present even in patients who cannot consciously perceive more than one or two objects at a time (Demeyere, Rzeskiewicz, Humphreys, & Humphreys, 2008). Some of the initial studies of perceptual averaging

did not necessarily implicate a novel perceptual mechanism, since performance could be approximated by simpler conscious heuristics that sampled only a few objects (Myczek & Simons, 2008). However, more recent work has shown that perceptual averaging remains robust in contexts wherein such heuristics cannot function (e.g. Albrecht & Scholl, 2010; Alvarez & Oliva, 2008; Chong, Joo, Emmanouil, & Treisman, 2008; Haberman & Whitney, 2010).

We have learned a great deal about the nature of perceptual averaging in recent years, and work spanning several decades has shown that such statistical summary representations can occur over many different perceptual dimensions, including size, inclination, length, orientation, spatial position, speed, and motion direction (for a recent review, see Alvarez, 2011) — as well as seemingly higher level features such as facial identity (e.g., de Fockert & Wolfenstein, 2009) and emotion (e.g. Haberman & Whitney, 2007). Such averaging is also not limited to spatial arrays, but can occur just as efficiently over temporal sequences, including those that are continuous in nature, as with a single disc that constantly grows and shrinks over time (Albrecht & Scholl, 2010). Nevertheless, to our knowledge, all previous studies of perceptual averaging have limited themselves to *visual* stimuli, even though our perceptual experience of the world is intrinsically multimodal.[1]

Auditory selection and averaging

Objects in lab-based perception experiments are frequently unimodal (insofar as objects depicted on computer monitors rarely have their own smells or haptic textures), and the vast majority of work in perception in recent decades has involved vision, so it is perhaps not a surprise that work on perceptual averaging has thus far focused on visual stimuli. In the present study, we focused on perceptual averaging in the auditory modality, and in particular on the perception of pitch, but of course one could also ask similar questions about a variety of stimulus dimensions in various sensory modalities.

Just as in vision, there is too much continuously streaming auditory information in our environment for us to process fully, and again a primary coping strategy is for auditory perception to be selective — focusing on only some especially relevant auditory features or objects at the expense of others. In addition to coping with auditory information overload by being selective, it also seems possible that perception could summarize auditory scenes — effectively producing a low-resolution representation of a full auditory landscape, rather than a high-resolution representation of only a part of it. In the present article, we address this question in the context of explicit tasks that require averaging the pitch from a sequence of tones.

[1] Perhaps the only exception is in studies of ambiguous speech perception, wherein the mean frequency of a prior tone sequence can bias the subsequent categorization of an ambiguous speech segment (e.g. Holt, 2006) — although this work was focused on mechanisms of speech perception rather than perceptual averaging per se, and hasn't yet been discussed in the literature on statistical summary representations.

Multimodal perceptual experience

Although perception can be rich and informative in several sensory modalities, these modalities do not exist in isolation. Indeed, it is common to identify features across several modalities that are all features of a single object (e.g. as one may (simultaneously or sequentially) perceive the size of a cat, the complex frequency of its purr, and the softness of its fur). Accordingly, a large body of work has focused on multimodal and crossmodal effects in perception and on how the senses cooperate to produce an integrated percept of the environment.

Audition and vision, for example, interact in several interesting ways (for reviews, see Ernst & Bülthoff, 2004; Kubovy & van Valkenburg, 2001; Spence, 2007). Beyond studies of explicit or implicit competition among stimuli from different modalities (e.g. Colavita & Weisberg, 1979; Shimojo & Shams, 2001; Shams & Kim, 2010; Welch & Warren, 1980), researchers have looked in other studies at how different modalities interact and support each other. For example, auditory information can facilitate covert spatial attention in vision (Driver & Spence, 1998), lead to inattentional blindness (Pizzighello & Bressan, 2008, Sinnett, Costa, & Soto-Faraco, 2006), facilitate learning in a visual perceptual learning task (Seitz, Kim, & Shams, 2006), and lead to the disambiguation of an otherwise ambiguous visual stimulus (Sekuler, Sekuler, & Lau, 1997).

The present study: Averaging with multimodal stimuli

In the experiment, we explored multimodal effects during perceptual averaging in the context of to-be-averaged displays that involved sequences of both visual discs and auditory tones, often presented in tandem. In particular, we explored two questions in the present experiment. First, can observers readily average pitch from a sequence of tones, just as they can average size from a sequence of discs? Second, to what degree can we effectively average simultaneously in both vision and audition?

Method

Participants

Twelve observers with normal or corrected-to-normal acuity participated in a 1-hour session in exchange for course credit.

Apparatus

The stimuli were presented using custom software written using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). Observers sat without head restraint approximately 50 cm from an 18-in. monitor in a dark room. Auditory stimuli were presented at the same moderate volume (25% of the computer's maximum output) via headphones for each observer.

Stimuli

The initial stimuli presented on each trial consisted of a 6-s sequence of eight discs (displayed in white on a dark gray background in the center of the display) and/or pure tones, each presented for 750 ms with no interstimulus interval. The diameter of each disc was randomly chosen (without replacement) from 60 possible values, uniformly distributed from 0.5° to 9°. The frequency of each tone was similarly randomly chosen (without replacement) from 60 possible values, logarithmically distributed from 65 Hz (C_3, the first note of the third octave) to 1,976 Hz (B_7, the last note of the seventh octave).

Procedure

Observers completed three blocks of trials (presented in a counterbalanced order across observers). Observers were initially informed that their task on each trial would be either to report the average pitch from a sequence of tones presented via the headphones, or to report the average size from a sequence of discs presented on the display. The single-modality auditory block and the single-modality visual block consisted of only auditory tones or visual discs, respectively. The remaining dual-modality block consisted of stimuli from both modalities presented simultaneously. However, the auditory and visual sequences in this block were uncorrelated: Each was generated independently in the manner of the single-modality sequences, such that a given disc diameter did not predict the temporally synchronized tone pitch, and vice versa.

After each trial in the dual-modality block, observers had to report *either* the average disc diameter (on dual-modality visual trials) or the average tone pitch (on dual-modality auditory trials) — but they did not find out which average they would have to report on a given trial until the stimulus sequence ended.

After each trial ended, observers then reported the average disc size or tone pitch via a blue judgment slider that appeared in the bottom third of the display. The judgment slider consisted of a stationary blue rectangle (15° × 2°) with a 15° two-pixel-wide horizontal black line centered on it. The observers manipulated an additional 2° two-pixel-wide vertical black line (the "slider"), which they could move to the right or left using the computer mouse to indicate their responses.

In the single-modality auditory block and the dual-modality auditory trials, observers heard a single judgment tone (that onset simultaneously with the judgment slider), and the pitch of this tone changed as observers moved the slider to the right (causing the pitch to become higher) or to the left (causing the pitch to become lower). The initial pitch of the judgment tone (and its associated slider position) was randomly selected from the full range of possible average tone pitches (from 83 to 1,585 Hz). As observers moved the slider to the left or the right, the judgment tone changed according to the slider's position on the scale (as a percentage of the length of the slider from left to right).

In the single-modality visual blocks and in the dual-modality visual trials, observers saw a blue judgment disc near the center of the display, and the size of this disc changed as observers moved the slider to the right (causing the disc to become larger) or to the left (causing the disc to become smaller). The initial diameter of the judgment disc (and its associated slider position) was randomly selected from the full range of possible average disc diameters (from 1.09–8.51°). The size of the judgment disc changed in the same way as did the judgment tone in the auditory judgment trials.

After observers indicated that they were satisfied with the resulting tone pitch or disc size (via a keypress), observers received feedback on the display in the form of (a) their error on that trial (expressed as a percentage, of the difference between the correct and reported average value, relative to the correct value), (b) their running average error percentage across trials, (c) the correct slider location on the judgment slider, and (d) either the correct average pitch (played through the headphones) or the correct average disc size (presented as a red disc near the center of the display). Observers then pressed a key to move on to the next trial, which began after a 750-ms interval containing only a blank gray screen with no auditory stimulus. Figure 1 summarizes this procedure for the Attend Visual task.

Design

Each observer completed 192 trials, consisting of four modified repetitions of each of 48 base trials: a single-modality visual trial, a single-modality auditory trial, a dual-modality visual trial, and a dual-modality auditory trial. These trials were presented in three separate blocks, with the dual-modality block being twice as long as either of the single-modality blocks. The order of the blocks was counterbalanced across observers, and the trial order within each block was randomized separately for each observer.

Results

Averaging performance on each trial was calculated as an error percentage, defined as the absolute value of the difference

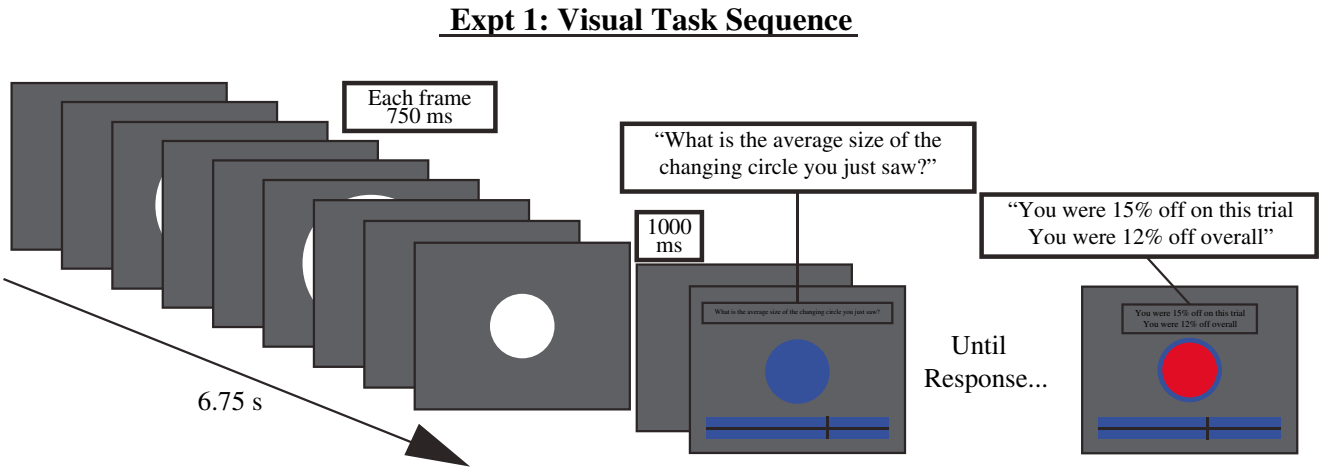

Fig. 1 Depiction of a sample trial sequence for visual averaging. Nine discs of different diameters are sequentially presented in the same location for 750 ms each, followed by a 1,000-ms blank screen. To report judgments of the average disc size from the sequence, observers adjusted a blue "slider" at the bottom of the screen, which manipulated the size of the judgment disc. Once observers indicated that they were done adjusting the slider, the correct answer was overlaid (or underlaid) as a red disc, along with the error % on that trial and also the running error %

between the correct average and the reported average, divided by the correct average. Error rates for each condition are depicted in Fig. 2. An inspection of this figure suggests two primary results. First, there is a large and clear main effect of modality, such that observers' reported averages were more accurate for auditory stimuli (6.78%) than for visual stimuli (13.19%). Second, there appears to be a very small but statistically reliable cost for averaging in the dual-modality condition, but only for visual stimuli. These impressions were verified as follows. A two-factor repeated measures ANOVA revealed a main effect of modality, $F(1, 11) = 118.042$, $p < .0001$, a main effect of single vs. dual presentation, $F(1, 11) = 8.790$, $p = .013$), and a significant interaction, $F(1, 11) = 5.679$, $p = .036$). Specific comparisons then revealed that performance in the dual-modality visual condition was worse than in the single-modality visual condition [14.02% vs. 12.36%, $t(11) = 3.26$, $p = .010$], but no such difference was present for single vs. dual auditory conditions [6.62% vs. 6.94%; $t(11) = 1.03$, $p = .330$].

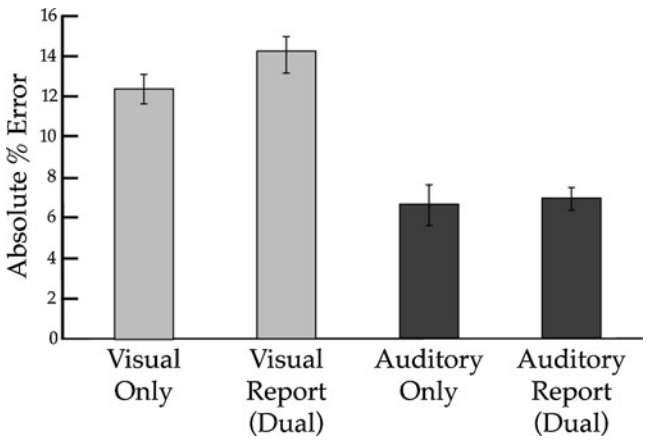

Fig. 2 Results, depicting absolute error percentages for performance in the four within-subjects conditions. Error bars represent 95% confidence intervals

Discussion

The first result of this experiment — and the primary finding of the project as a whole —was that observers were able to efficiently average sequences of pitches, in much the same way as they can efficiently average visual stimuli. This result suggests that perceptual averaging is not a process that is specific to vision, and that statistical summary representations can be computed from at least two modalities. To our knowledge, this is the first demonstration of auditory perceptual averaging.

In fact, auditory averaging performance in this experiment was better than the matched case of visual averaging. This could be because of more precise individual estimates in the auditory modality, although these were not measured here. For example, it may be that auditory processing (and/or auditory memory) is more precise than visual processing/memory in the temporal domain (as tested here), whereas visual processing is more precise in the spatial domain (e.g., Shams, Kamitani, & Shimojo, 2000). (Note, though, that visual size averaging is no better or worse when the stimuli are presented simultaneously vs. sequentially; Albrecht & Scholl, 2010; Chong & Treisman, 2005a). In any case, note that this result may not reflect anything general about perceptual averaging, since it could be specific to the temporal sequences employed here, or to the specific features tested

in each modality (viz. size and pitch), or just to intrinsic differences in the relative salience of these stimuli (perhaps because the auditory stimuli as presented through the headphones may have seemed to be closer to the observers). In addition, although we controlled for the relative differences in each of the individual sizes and tones chosen (by having 60 evenly spaced possibilities in each domain), it may be that the use of distinct and discrete notes for the individual tones biased observers to perform better on the auditory task, given that the visual sizes were not tested explicitly beforehand for their differentiability. Similarly, perceived size has been shown to follow a power function with an exponent of 0.76 (Teghtsoonian, 1965), whereas we selected visual sizes based on a linear scale. This may also explain why the individual tones (chosen based on typical Western musical notes, which follow a logarithmic scale) may have been easier for observers to discriminate than the individual visual sizes. For these reasons, we consider the comparison of auditory and visual averaging in the present study to be largely an informal case study, with our primary result simply being that auditory pitch averaging is robust in absolute terms in at least some contexts. And, of course, we cannot draw any sweeping conclusions about multimodal averaging in general, since the present results could also be specific to audition.[2]

The second primary result from this experiment was that simultaneous averaging in vision and audition was possible either without a cost (for dual-modality trials that ended up asking for an auditory average) or with a small (<2%) cost (for dual-modality trials that ended up asking for a visual average).[3] This difference may again reflect a relative benefit for auditory processing in temporal sequences, such that auditory averaging occurs more automatically. In any case, this pattern of results is reminiscent of the visual results of Emmanouil and Treisman (2008): When averaging size and speed with the dimension to be averaged cued either before or after the displays, their observers showed only a small (often < 1%) postcue versus precue cost for size averaging, and no statistically reliable difference for speed averaging.

It appears that auditory averaging, even when observers are asked to compute averages over multiple modalities simultaneously, provides no special challenge to perception. Averaging in the more difficult visual modality was only slightly more susceptible to task load in our study. These results are thus consistent with the possibility of distinct resources for averaging in vision and audition, or with shared resources whose limits are generous enough to encompass both tasks simultaneously. For now, in any case, these results make clear that perceptual averaging can occur in multiple modalities, which in turn stresses the idea that perceptual averaging may transcend visual processing.

Author Note For helpful conversation and/or comments on previous drafts, we thank Sang-Chul Chong, Karla Evans, Todd Horowitz, Megan Long, and one anonymous reviewers. For more information, visit http://www.yale.edu/perception.

References

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science, 21,* 560–567.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Science, 15,* 122–131.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*, 392–8.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12,* 157–162.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Chong, S. C., Joo, S. J., Emmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics, 70,* 1327–1334.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393–404.

Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics, 67,* 1–13.

Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45,* 891–900.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62,* 73–101.

Colavita, F. B., & Weisberg, D. (1979). A further investigation of visual dominance. *Perception & Psychophysics, 25,* 345–347.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology, 62,* 1716–1722.

Demeyere, N., Rzeskiewicz, A., Humphreys, K. A., & Humphreys, G. W. (2008). Automatic statistical processing of visual properties in simultanagnosia. *Neuropsychologia, 46,* 2861–2864.

Driver, J., & Spence, C. (1998). Cross-modal links in spatial attention. *Philosophical Transactions of the Royal Society London Biological Sciences, 353,* 1319–1331.

[2] Designing analogous experiments in the context of olfactory or gustatory averaging would be challenging, and such studies might end up having to employ very different experimental designs. However, it does seem possible to study haptic averaging in an analogous fashion—just asking about the ability to report average size from a set of carefully constructed marbles while blindfolded. We are asking such questions in present research.

[3] Although errors were small in this experiment, they were nevertheless systematic in at least one way: On dual-modality trials, observers had a relatively small but statistically reliable trend to bias their judgments toward the location of the correct average of the irrelevant modality on the judgment slider. This was true for both visual responses (of which 68% were biased toward the judgment slider location of the correct but task-irrelevant auditory average; $t(11) = 10.68$, $p < .001$) and auditory responses (of which 55% were biased toward the judgment slider location of the correct but task-irrelevant visual average; $t(11) = 2.91$, $p = .01$), with the former being larger, $t(11) = 4.92$, $p < .001$.

Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Perception & Psychophysics, 70,* 946–954.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Science, 8,* 162–169.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17,* R751–R753.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72,* 1825–1838.

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America, 120,* 2801–2817.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition, 80,* 97–126.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70,* 772–788.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Resesarch, 155,* 23–36.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Pizzighello, S., & Bressan, P. (2008). Auditory attention causes visual inattentional blindness. *Perception, 37,* 859–866.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*, 308.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature, 408,* 788.

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews, 7,* 269–284.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Current Opinion in Neurobiology, 11,* 505–509.

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology, 16,* 1422–1427.

Sinnett, S., Costa, A., & Soto-Faraco, S. (2006). Manipulating inattentional blindness within and across sensory modalities. *Quarterly Journal of Experimental Psychology, 59,* 1425–1442.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Society of Japan, 28,* 61–70.

Teghtsoonian, M. (1965). The judgment of size. *The American Journal of Psychology, 78,* 392–402.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin & Review, 88,* 638–667.

Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review, 18,* 484–489.