A selective summary of visual averaging research and issues up to 2000

Ben Ben Bauer
Department of Psychology, Trent University, Oshawa, Ontario, Canada

Ariely's (2001) “Seeing Sets: Representation by Statistical Properties” (Psychological Science, 12, 157–162) rekindled interest in summary-value estimation for visual ensembles (groups of similar items). Revisiting and reinvigorating research on the “intuitive statistician” has prompted a new set of insights and debates concerning how and why the visual system might benefit from a compact representation of the optic array and how this might relate to crowding, sparse representation, efficiency coding, and processing limits. New research tools and imaging techniques coupled with solid psychophysical work have added substantially to the large base of work done in the 20th century. The present brief review acts as a summary of the ensemble of work prior to Ariely’s (2001) landmark paper to encourage a comprehensive continuity of knowledge and reintroduce some of the contemporaneous concerns to help inform ongoing research and modeling.

Introduction

Ariely’s (2001) “Seeing Sets: Representation by Statistical Properties” has been cited (as of February 2015, English language survey) almost 300 times across a wide range of sources: in Kahneman’s Nobel Prize lecture, in dozens of vision science judgment/decision making and experimental psychology journals, and in published proceedings, theses, and textbooks. This influential article rekindled interest in a type of visual information processing commonly known as intuitive statistics (e.g., Cosmides & Tooby, 1996; Peterson & Beach, 1967; Pollard, 1984; Slovic & Lichtenstein, 1971). The central topic is the ability of human observers to estimate various summary statistics based on properties of groups (ensembles) of visual stimuli. These statistics are summary values based on the whole ensemble (contributions of many items) rather than on a single item and often taken under conditions thought to preclude item-by-item inspection. Ariely discussed his results in terms of several interwoven research perspectives including information-processing models, perception of illusions, adaptation-level theory, information theory, visual search and attention, judgment and decision making, and social cognition. In fact, these areas plus attribution theory (see Fischhoff, 1976) and Anderson’s functional measurement framework (see Anderson, 2014) have shared methodology, analytic techniques, and insights, as can be seen in the Peterson and Beach (1967), Slovic and Lichtenstein (1971), Pollard (1984), and Cosmides and Tooby (1996) review and integration articles.

The goal of the present article is to summarize the work done prior to Ariely’s (2001) contribution to provide context for understanding the more recent results and controversies. In particular, the emphasis is on visual ensembles and estimation of central tendency because that has been the focus of much research in the recent decade or so. Occasional mention of other types of statistical judgment are given to provide context. Historically, the interest in ensemble statistics as psychological units in sensory/perceptual systems (and perhaps multimodally) was called “coalescence” by William James (1890) and “constructive combination” in Messenger (1903).

As a broad synopsis of the several review articles and a large number of research articles in scope, three compendious observations can be made. First, as Peterson and Beach (1967) and Pollard (1984) concluded, within limits people can make respectable estimates of ensemble statistics, although deviations from the formal or normative values are often observed. Cosmides and Tooby (1996) have hence revisited the inferential aspect of humans qua good intuitive statisticians in the context of adaptive behavior. Their ecological/evolutionary point of view echoes the statement made 30 years earlier by Peterson and Beach (1967), that man “survives and prospers while using the fallible information to infer the states of his uncertain environment and to predict future events” (p. 29). Weaving together evidence from 86 papers...
(from Alba to Zacks) and their eight new studies, Cosmides and Tooby (1996) conclude that we should “grant human intuition a little more respect than it has recently been receiving. The evolved mechanisms that undergird our intuitions have been subjected to millions of years of field testing against a very rich and complexly structured environment” (p. 69; see also Hogarth, 1981, and Runeson, 1977, for similar conclusions but from different points of view). Thus, despite a bias in favor of contrary reports (Christensen-Szalanski & Beach, 1984), there is plenty of evidence to support a claim that human observers are skillful at estimations of various statistics (e.g., averages, totals, variances) on numerous properties of a diverse range of stimuli, from low-level perceptual attributes such as weight, length, speed, or inclination to more high-level conceptual properties such as supermarket food prices, production costs, IQ scores, or likeableness of an individual based on symbolic stimuli such as numerals or lists of adjectives.

The second observation concerns, as Meyer, Taieb, and Flascher (1997) note, that “instead of dealing with the subject domain (e.g., statistics) when trying to understand intuitive judgments one should analyze the geometrical and perceptual properties of the displays or other information on which estimates are based” (p. 19). Wright and Murphy (1984) warn researchers to avoid availability bias in investigating the nature of the algorithm with normative presuppositions: “there is no a priori reason for using a particular statistic as the sole standard by which to evaluate subjects’ judgments” (p. 304). Clearly, the nature of the algorithm should not be prejudged; the psychophysical function and the environmental information on which it is based should be allowed to emerge without predilection, lest the true mental algorithm be obscured by expectation (Anderson, 1968; Levin, 1975). The import of this is that given reliable performance by the observers (that is, that they produce responses systematically influenced by stimulus ensemble properties), the goals should be to determine how the ensemble coding is performed, if the algorithm corresponds to known mathematical procedures, whether the algorithm acts prior to or following known nonlinearities in the visual system (e.g., cortical magnification, constancy, accommodation and assimilation, attentional demands), and whether these factors interact with the property under investigation (i.e., is the averaging mechanism similar across visual domains such as size, color, spatial/temporal frequency, speed, or are these summarized in differing ways perhaps based on dynamic range, inherent noise, etc.). As Runeson (1994) stated, “rather than prescribing the perceptual attributes, and trying to make people good meters of them, psychophysics must take as its primary goal to search for and identify the properties that are actually perceived” (p. 762). For example, Peterson and Miller (1964) showed that observers can produce mode, median, or mean (of speeds) as requested based on payoffs (see also Massaro, 1969, and Minturn & Reese, 1951, for a demonstration of observer compliance to feedback). This shows that people are adaptable but does not reveal what they do naturally in situ. It was very wise of some of the early researchers to avoid leading instructions that could have tainted the results with demand characteristics. For example, Spencer (1961) asked observers to view cards containing 20 two-digit numbers and, for each card, to “assess the single value which would best represent the series presented or to estimate ‘average value’ suggested by the series” (p. 318). Anderson (1964) instructed his observers “not to worry unduly about mathematical correctness but to give ’what your feeling about the average is’” (p. 191).

For the third observation, it is worth emphasizing that the tools we use to measure these intuitive central tendency values are somewhat blunt. It is well known that psychological measurement is not impervious to the effects of context, scaling, measurement error, and other issues that can hinder the revelation of the true psychophysical relationship. It is therefore possible that measurement precision will bump up against experimental error in seeking answers to many of the questions asked. In some cases, performance may reflect more on the nature of the probe (experiments) than of the probed (perceptual averaging) than desired. Errors such as those due to context and memory effects (e.g., Algom, Wolf, & Bergman, 1985; Ross & Di Lollo, 1971; Ward, 1987), end-level and contrast effects (e.g., Anderson & Jacobson, 1968; Birnbaum, Kobernick, & Veit, 1974; Birnbaum, Parducci, & Gifford, 1971; Poulton, 1975), contraction bias (e.g., Hollingworth, 1910; Poulton, 1979), primacy/recency (e.g., Anderson, 1967, 1973; Anderson & Whalen, 1960; Hendrick & Costantini, 1970; Weiss & Anderson, 1969), response frequency equalization (e.g., Baird, Lewis, & Romer, 1970; Parducci, 1965; Parducci & Haugen, 1967; Steinberg & Sekuler, 1973; Stevens & Galanter, 1957), propensity to use integer multiples of 5 or 10 in ratings (e.g., Baird & Noma, 1975; Baird et al., 1970; Krueger, 1982), or any other experimental/observer biases in the stimuli/responding (see Pearson, 1922) simply mean that we need to consider these when interpreting the accumulating evidence and to be careful and clever in investigations. To be sure, these effects are real, meaningful, and worthy of study. The enterprise is to figure out what people are doing—what is the output from the perceptual operators that calculate these intuitive statistics.

Finally, N. R. Brown and Siegler (1993) frame all three issues as the difference between metrics (plausible ranges and values for ensemble statistics) and mappings (information about members of the ensemble). Most important for the present discussion is their assertion
that this distinction and our knowledge of observers’ representation of these two classes of information are incomplete.

It raises such questions as whether people are sensitive to differences among alternate measures of central tendencies (means, medians, and modes), whether they are sensitive to different moments of the distribution (mean, variance, skewness, and kurtosis), and whether they differentiate among general shapes of distributions (normal, rectangular, bimodal, and so forth; N. R. Brown & Siegler, 1993, p. 530).

### Summation

For some mathematical central tendency algorithms (e.g., the arithmetic, quadratic, and harmonic means), there is a summation term involved. Therefore, a short and selective discussion of perceptual summation is presented while acknowledging that even if the output of one of these computations turned out to be highly similar to the output of the mental-mean operation, there is no reason to suggest that a summing process is necessarily involved in arriving at the mental mean. Summation of perceptual and symbolic material within and across sensory modalities is not treated in Peterson and Beach (1967) or Pollard (1984), although, admittedly, the area was nascent at that time primarily in terms of how clinicians integrate various diagnostics in assessment of patients. Clearly, the nature of the intuitive integration rule (e.g., additive/arithmetic averaging vs multiplicative/geometric averaging) greatly influences the final intuitive value and decisions based on that value. The related topic—intuitive averaging—is discussed in the two reviews and in detail in a following section. In a normative sense, summation and averaging are related by knowledge of numerosity, but again, this link may have no computational concomitant in the perceptual system. One might also consider numerosity as a summative operation in at least a nominal sense. In light of this “intuitive relation,” experiments in which observers were specifically asked or implicitly required to sum impressions/ perceptions are described in the following paragraphs with occasional contact with averaging research.

Anderson (1962) validated a mathematical model of impression integration in which observers were asked to give an overall assessment of likeableness for a hypothetical individual based on three adjectives. The three component adjectives in the description of the individual were selected from sets of low-, mid-, and high-likeableness adjectives (as rated by his observers) in various combinations. Anderson found that the additivity of the three likeableness values predicted the judged likeableness with correlations across observers of 0.94 to 0.99 between judged overall likeableness and summed adjective likeableness values. For the present purposes, this is an ensemble of three adjectives, each with some valence, and the intuitive integration rule is highly similar to a normative summing process. The assertion that subjective impressions of adjectives and sensory experiences are combined according to a purely additive rule (or that this can even be unambiguously determined using the current statistical tools in vogue) has been challenged on the grounds of appropriate weighting/scaling for stimuli and responses (Birnbaum, 1974), with confirmation based on supporting the null hypotheses (viz, no interactions, e.g., Gigerenzer, 2005) and range/bias effects (e.g., Poulton, 1975, 1979).

Additional work concerning averaging of numbers as prices, IQ, or performance scores is covered in the section “Numbers in Real-World Context” below.

Anderson has made an impressive contribution to the study of integration rules in social, cognitive, and perceptual psychology, with more than 100 articles addressing the mathematical, psychophysical, and psychological nature of information integration (Noble & Shanteau, 1999). His functional measurement model (Anderson, 1970, 1979) formally specified the mathematical signatures necessary in the behavioral data to support claims about various integration rules (i.e., cognitive algebra; Bettman, Capon, & Lutz, 1975). Essentially, he provided the litmus test for determining the integration rules that observers use (naturally) or can use (according to instructions) when estimating ensemble statistics. His work showed that observers produce, to a first approximation, sums/averages, but they also could produce estimates approaching normative multiplicative performance (Anderson & Butzin, 1974, 1978; Graesser & Anderson, 1974; Oden & Anderson, 1974). Anderson (1965) provided evidence (although not conclusive in his opinion) that the integration rule for adjectives was more similar to an averaging process than to a pure additive process because including two mid-value adjectives with two high-value adjectives resulted in a lower overall assessment than for the two high-value adjectives alone. In other words, the two mid-value adjectives pulled the overall judgment down (dilution) rather than augmenting it. At the same time, he replicated the set size effect (later known as the “numerosity heuristic”; see Pelham, Sumarta, & Myaskovsky, 1994, for a brief history) in that a greater numerosity of items biases judgment of the whole. In Anderson’s case, a set of four adjectives of identical value resulted in a more extreme judgment than did two adjectives of the same identical value. (e.g., four high-value adjectives led to a higher average rating of likeableness than did two high-value adjectives). In terms of ensemble statistics, the observation that an ensemble of two identical items leads to a different response than does an ensemble of four of
these same items is not, prima facie, consistent with an arithmetic averaging model. Anderson does provide a modification to the averaging model that can address this by assuming an anchor and adjust strategy.

Ongoing research on the effects of ensemble-member variability in summary statistics can be informed by Anderson's work. Krueger (1970), based on the established Stevens' exponent of approximately 1.0 for visual line length (see Baird, Romer, & Stein, 1970; Mashhour & Hosman, 1968; Rule, 1969; Spence, 2004; Stevens & Galanter, 1957; M. Teghtsoonian, 1965; R. Teghtsoonian & M. Teghtsoonian, 1970; Wiest & Bell, 1985) tested whether the estimated combined magnitude for line ensembles was equivalent to the sum of the physical magnitudes. His observers adjusted a test line to be equal in length to their perceptual sums of the two- or four-line ensembles (physical sums were 1.5, 3.0, and 6.0 inches). Krueger's results showed almost ubiquitous overestimates of combined size across three levels of total length, two- and four-line ensembles, and several display layout parameters. That said, the performance was very much normative with a constant error of about 1/3 inch for two-line ensembles and about double that for four-line ensembles. The numerosity heuristic (a whole is perceived as a joint function of the sum of the parts and the numerosity of the parts) would explain the overestimation in general and the greater overestimation for the four-line ensembles.

Abravanel (1971) found general overestimation for length sums of wooden bars when the two bars were presented intramodally (visual-visual or haptic-haptic) or intermodally (visual-haptic). The deviation from normative was essentially a constant error. The task again was magnitude production using a slider. Stanley (1974) found approximately normative summation of perceived angles (sums 30° to 130° in ensembles of two) with systematic overestimation of sums evident. Indow and Ida (1977) showed slight overestimation for summed numerosity with magnitude production. Zwislocki (1983) found approximate additivity of loudness for tone bursts.

In summary, when the task is summation of a property of ensemble members, the perceived sum appears to slightly overstate the normative sum.

Central tendency

Early related investigations

Plateau (1872, cited in Laming & Laming, 1996) asked painters to mix a gray that fell midway between a black and a white paint. In this early “method of production experiment,” there was remarkable consistency across the eight artists in the (bisection) gray they produced. Hollingworth (1910) noted that for many stimulus domains (e.g., size, brightness, duration), observers tend to produce judgments that “gravitate toward a mean magnitude” (p. 461). This observation is based on experiments in which observers were comparing (within seconds) the memory of a probe stimulus size against available comparison stimuli. Granted, the observer’s task is not explicit average-value estimation; nevertheless, sensitivity to the average of the stimuli was demonstrated. Hollingworth refers to the mean magnitude as the “indifference point” and notes that it approximates the midpoint of the current set of stimuli under consideration by the observer. Philip (1947, 1952) noted a similar effect. Gottsdanker (1952, 1956) conjectured that the relative insensitivity of observers to velocity changes in manual tracking tasks reflected their adherence to a running mean. That is, acceleration or deceleration is an instantaneous departure from the contemporaneous mean of the constant velocity pursuit, and averaging a small number of higher velocities with a large number of previous velocities would result in a lagged impact in both time and magnitude. A similar claim was made by Craig (1949, cited in Slack, 1953), although Slack (1953) determined that it is only very recent experience (the previous few trials) that contributes to performance on a current trial (see also Morgan, Watamaniuk, & McKee, 2000, for a similar contention).

Helson’s adaptation level theory (Helson, 1947; Helson, Michels, & Sturgeon, 1954; Michels & Helson, 1949; see also, D. R. Brown, 1953; Parducci, Calfee, Marshall, & Davidson, 1960) proposed that judgments for a given stimulus are a combination (weighted geometric mean) of three types of stimuli. The first is the stimulus currently being judged. The second type (spatially removed) includes the ensemble of nontarget stimuli supposedly extraneous to the judgment (flankers, surrounds, peripheral items, etc.). The third type (temporally removed) consists of all stimuli experienced prior, not just the stimulus array on a given trial, but also the frequency/probability of that stimulus over the series of trials and even the observer’s responses, which can also be treated as stimuli. An important aspect of this taxonomy for the present review is to understand that the sampling interval (temporal or spatial) on which a perceptual statistic is calculated can be a significant variable. This variable should be controlled where possible and its performance contribution investigated. In other words, although a dependent measure such as a numerical estimation or reaction time might be recorded and attributed to the stimulus presented on a given trial, the functional stimulus set is far more complex and extended in space and time than is commonly assumed. These examples suggest that the computation of temporal or spatial ensemble averages...
may occur irrespective of instructions to do so, and this computation affects behavior on an ongoing basis.

Wallin (1912) investigated the ability of observers to judge the mid-tempo for two tick rates produced by a metronome (rates of 40 to 208 beats per minute were used). Perceptual averages of the tempos approximated the arithmetic mean in many cases and the geometric mean in others. Accuracy decreased as the difference between the to-be-averaged tempos increased. Relative error ranged from 5% to 10%, with perceptual midrates generally below the arithmetic mean of the two rates. Pearson (1922) found slight underestimation (about 2%) in his line bisection task using a 5.94-inch line bisected 1,260 times. H. K. Wolfe (1923) conducted a study on estimation of the midpoint of lines (bisection of lines between 5 cm and 1 m viewed at “reading distance”); for both vertical and horizontal lines, errors in midpoint estimation were on the order of 1% to 2%. Chidester (1935) replicated H. K. Wolfe’s findings with lengths of 5 cm to 50 cm, again at reading distance. Her observers produced errors on the order of a few percentage points. Both H. K. Wolfe’s and Chidester’s papers include discussion of several German-language papers from the late 19th and early 20th century, generally corroborating the small percentage error but with inconsistent results of right versus left overestimation of length. Attnave (1954) suggested that by “averaging out particulars and abstracting certain statistical homogeneities” (p. 188), observers can avoid overloading the finite perceptual capacity. In particular, Attnave implicated intuitive measures of central tendency and variance.

**Time/frequency**

Deese (1955) made a claim about mental averaging similar to the early claims of Hasher and Zacks (1979, 1984) regarding the incidental encoding of frequency of occurrence. Deese’s (1955) expectancy hypothesis claimed that “the detection for a given signal should be determined by the average of all signal intervals before the one in question” (p. 363). His focus was the effects usually attributed to the putative mechanisms of attention/fatigue/vigilance in a task requiring observers to indicate the occurrence of a signal with low temporal probability (on the order of tens per hour). Rather than inferring the existence of these three mechanisms and their impact, he preferred an explanation based on the expectation by the sensory system, that a signal should occur when elapsed time since the most recent signal is equal to the mean intersignal duration. This is an implementation of what Estes (1955) called “shift[ing] the burden of explanation from hypothesized processes in the organism to statistical properties of environmental events” (p. 145). Baker (1959) formalized Deese’s hypothesis and tested it experimentally (Baker, 1962). Briefly, observers detected 12 flashes of light (the signal) that occurred at various programmed times over 20 to 30 min. For example, in one session, the intersignal intervals ranged from 30 s to 400 s. Following the detection session, the observers were asked to extend the series by manually flashing the light two more times when they expect it should occur. For the sessions with a 150-s mean intersignal interval, the first manual signal occurred within 161 s and the second manual signal within 153 s. For the sessions with a 100-s mean intersignal interval, the manual signals were given within between 84 and 90 s. This represents ensembles of 12 intersignal intervals whose (highly variable) durations were averaged by observers to within ±10% over a period of 20 to 30 min with no explicit instructions to do so (see also Eggleton, 1982; Spencer, 1961, for use of extrapolation as method of eliciting intuitive estimates of averages). A post hoc test ruled out an explanation based on the use of the duration of the final intersignal interval as the benchmark for the manual productions.

These studies predate Peterson and Beach’s (1967) review but were not included. The second major subsection of their review and Pollard’s (1984) treatment of descriptive intuitive statistics covers judgments of means and variances, although both are dominated by the latter. They summarize evidence for several key points: Observers can produce a reasonable estimate of averages, and these estimates deviate from normative as the ensembles increase in skew or variability and are prone to recency effects for sequential presentation of ensemble members. The classic articles are Beach and Swenson (1966),1 Spencer (1961, 1963), and Peterson and Miller (1964).

These basic findings are developed and expanded in the next few paragraphs, with a treatment of other relevant research in a somewhat chronological order.

**Numbers/spatial and effects of ensemble variance**

Spencer (1961, 1963) presented symbolic stimuli (two-digit numbers in lists of 10 or 20 per card) and asked observers to “assess the single value which would best represent the series presented or to estimate ‘the average value’ suggested by the series” (Spencer, 1961, p. 318). Observers were given 5 or 10 s to view each card, and no knowledge of results was given to an observer prior to completion of the estimates. The numbers on the cards were selected to create (arithmetic) means of between 20 and 80, with standard deviations of 2.0 to 19.0. The mean error (deviation of the estimated mean from the arithmetic mean of the numbers on the card) was between −0.1 and +2.9.
might worry that this impressive result arises from large errors of opposite sign canceling each other out in the arithmetic mean. However, the mean absolute value of the estimation error ranged from 0.44 to 7.28, with shorter viewing time and more variability in the values on a given card resulting in larger error (see also M. L. Wolfe, 1975). It is interesting to note that observers in this type of task often feel as if they are responding randomly at worst or guessing at best. Spencer (1961) reports that after having explained the experimental task to his observers, they would express skepticism that such a task could be performed at all. In other words, they had little confidence in their ability to estimate averages with any precision. He furthers that in contrast to observers’ dire predictions, he enjoyed “the sheepish grins of pride and surprise on the subjects’ faces when, at the conclusion of their judgments, they were shown how well they performed” (p. 318).

Spencer next investigated the effect that extreme or “rogue” numeric values would have on estimates. He found that the rogue value pulled the estimate in the expected direction but that its subjective influence was larger than would be predicted by its numerical value. As an extension, Spencer then took the same sets of numbers that had been presented on the cards and used them to create spatial stimuli for averaging. Essentially, the numbers were used as heights from the x-axis in a bar graph–like layout, and the observer was asked to indicate the average position (i.e., the one-dimensional centroid). Again, the average estimates were quite accurate, and the amount of absolute error increased with the standard deviation of the distances. Interestingly, in the case of these spatial stimuli, the effect of rogue items was opposite that encountered with the symbolic stimuli. In effect, rogue stimuli (deviant spatial locations) undercontributed to the average estimate. Nevertheless, observers are capable of determining averages with great accuracy under viewing times unlikely to permit the conventional serial sum and divide calculation (e.g., for arithmetic means, \[\frac{x_1 + x_2 + \ldots + x_n}{n}\]. Anderson (1968) replicated the differential contributions of outliers in spatial versus numerical averaging. Laestadius (1970) used ensembles of 15 numbers between 0 and 99 and found that observers’ estimates of the means were quite accurate (mean absolute deviations of about 5% of the true mean) but that accuracy was lower for high-variance ensembles (see also Bulger, Hiles, & Lowe, 1969). Hendrick and Costantini (1970) report a similar result using a wider range of numbers (presented in six-item spoken sequential ensembles) with as much as 20% error in the estimated mean (see also Lowe, 1969).

Irwin and Smith (1956) and, Irwin, Smith, and Mayfield (1956) asked observers to view a series of numbers (positive and negative) presented one at a time on cards and to predict with as few cards as possible (out of a deck of 500) whether the expected mean of the deck would be greater than or less than zero. This is sometimes called a “stopping task.” On average, observers stopped the deck at 20 to 40 cards and reported the correct sign of the mean. The number of cards required increased with the absolute value of the deviation of the deck mean from zero (to ±1.5) and with the standard deviation of the numbers (2.0 vs. 7.5).

Spencer (1963) classified the average-value estimates made by his observers to ensembles of two-digit numbers according to their proximity to three averages: the arithmetic mean, the median, and the maximin (defined by Spencer as the average of a small quantity of the largest and smallest values). Skewed distributions of items (numbers or locations) were used so that the arithmetic mean, median, and maximin were numerically distinct. The arithmetic mean and maximin accounted for about 36% each of the classifiable responses, with the median accounting for 18% (about 9% of responses did not fit in any these three categories). He repeated this analysis for his spatial stimulus data and found that 57% of responses were consistent with the arithmetic mean. Of the remaining 41% of classifiable responses, 14% were median-like and 27% were maximin-like (less than 3% were unclassifiable). Introspectively, observers reported several strategies such as subsampling the items, performing maximin calculations, and a “synoptic” or molar approach. These results did not pinpoint a clear winner.

Beach and Swenson (1966) also asked observers to estimate means for lists of one- and two-digit numbers. They found that the mean absolute deviation between the subjective estimates and the arithmetic mean was smaller than the deviation from the median or maximin. They also found that greater variability in the ensemble members reduced the accuracy of estimation. Lovie and Lovie (1976) found essentially perfect arithmetic mean estimation for 8, 16, or 24 numerals, although accuracy decreased with list variance. Peterson and Miller (1964) used a different approach. Using a biasing payoff schedule, their observers could be compelled to produce a mode, median, or mean (of speeds) to maximize return. This underscores the importance of not providing knowledge of results if one wishes to explore which measure of central tendency observers produce naturally.

Weights, lines, tilt, and order effects

In a heaviness averaging task (Anderson, 1967), observers hefted six weights sequentially and then rated the average from 1 (very light) to 20 (very heavy). The lifting sequences were chosen beforehand to allow tests
for recency and order effects. These effects were obvious in the mean heaviness judgments (e.g., the same ensemble of weights with the lighter weight lifted last was deemed to have a lower average than when the last weight was heavier). Nevertheless, using the results given in Anderson’s table 1, one can compute the Stevens exponent for the average heaviness judgments as a function of average weight. The exponent is 1.49, which compares favorably with the exponent for magnitude estimation of weights hefted singly (1.47; Stevens & Galanter, 1957). This demonstrates that bias effects can be offset by counterbalancing. Sequence effects were replicated in Anderson and Jacobsen (1968). These artifacts aside, the intuitive estimate of central tendency is closest to the arithmetic mean in these cases. Weiss and Anderson (1969) provided further support for an intuitive arithmetic mean for line lengths using stimuli 14 to 26 cm long viewed at a distance of 168 cm (4.8° to 8.8° visual angle). There were three presentation conditions. Ensembles of three or six lines were presented in sequential ensembles (observers produced the mean after the last item). Ensembles of 10 items were also presented in sequential ensembles but with a running mean produced by observers after every item. Finally, spatiotemporal ensembles of six items were tested in which one to five items were presented simultaneously with the balance presented sequentially within a trial (produce the mean after the last item). After weighting items to remove serial position effects, the arithmetic mean turned out to be the best normative analog to observers’ productions. Of course, given that the Stevens exponent for judged length is near 1.0, no scaling was required. Note that Weiss and Anderson were careful not to enforce a particular type of averaging on their observers; they simply included the word average in their instructions with no interpretation. Further discussion of which measure of central tendency is computed is presented in the summary of modern ensemble statistic work in a later section. To anticipate, there is some evidence that the scaling of the stimulus range (linear vs. logarithmic) can influence which mean is calculated and that for some properties, performance may indicate estimates smaller than the arithmetic mean therefore approaching the geometric mean.

Groman and Worsham (1970) presented observers with two cards slanted differently in depth (one to each eye) to determine the nature of the integrated perception of slant. The perceived slant was very close to the arithmetic mean of the slant to each eye. Averaging of inclination was also investigated in Miller and Sheldon (1969) along with intuitive averaging of line length. Their six-item ensembles (simultaneous presentation) contained lengths suitable to produce means of 20 to 185 cm viewed from approximately 5.2 m, resulting in means of about 2° to 20° visual angle. In contrast to Weiss and Anderson (1969), instructions to the observers made reference to the method of computing an arithmetic mean. Three different comparison lines (moduli) were available during judgment (30 cm, 91 cm, 152 cm: between observers), and group Stevens exponents were 1.14, 1.00, and 1.04, respectively (as expected, the moduli at the extremes of the range produced exponents that deviated from unity; see Engen & Ross, 1966). Individual observer exponents ranged from 0.84 to 1.19. For inclination, ensembles of six lines with angles between 0° and 90° were used (0° = horizontal). Mean inclinations of the ensembles were between 10° and 80°. The individual exponents were more variable than the length exponents (range = 0.68–1.30), with a group exponent of 1.0, which is consistent with the linearity of inclination judgments for single items reported by Stevens and Galanter (1957) when taken over a range less than 90°. The scaling of physical-to-perceptual angle has a history of inconsistent results. Some of these can be traced to stimulus presentation confounds such as orientation of the presented angles or alignment with context edges, reference/canonical angles, and anchoring effects (see Beery, 1968; Fischer, 1969; Jastrow, 1892; Judd, 1899; Pratt, 1926; Smith, 1962; Weene & Held, 1966, for history and development). Stanley (1974) found a tendency toward overestimation of the average of angle pairs containing a small angle and underestimation of averages as the size of the individual angles increased (component angles ranged from 10°–70°). This is consistent with Maclean and Stacey’s (1971) result for angles judged singly and recent work by Nundy, Lotto, Coppola, Shimp, and Purves (2000), who also found under-then overestimation in three psychophysical tasks. Note that such a nonveridical perception seems curiously nonnormative; however, the goal of the Nundy et al. article was to show that perception is in fact statistically valid. That is, the physical angle in the environment that likely resulted in a small perceived angle was likely larger than its projection on the retina (or, mutatis mutandis, the reverse for large perceived angles). This is (par excellence) one of Gigerenzer’s “good errors.” Notably, the perceived angle represents something akin to the average value of the physical angles in the environment that gave rise to that perception (see figure 6 in Nundy et al., 2000, for an amazing correspondence between behavior and natural scene statistics). This is an intuitive average value on an extended temporal ensemble and, although nonnormative, full of survival value.

Returning to line-length averaging, Miller, Pedersen, and Sheldon (1970), replicated the line-averaging part of the Miller and Sheldon (1969) study (which had used 18 military men as observers) with a different sample of observers (80 university students) and presentation of modulus during or only before the ensemble. Explo-
ponents ranged between 0.92 and 1.03, in line with the earlier experiment. Miller and Sheldon (1969, p. 12) suggest that “from these considerations the 'power law’ might be expected to apply to the SA [subjective average] when it holds for the unitary continuum.” This is a profound statement that surely requires additional evidence.

**Color and scaling**

Weiss (1972) compared judgments of grayness for stimuli taken one at a time or averaged over an ensemble of two items. The judgments were provided by observers as magnitude estimates relative to a modulus of 100 (the modulus was available during judgments) or by marking a position on a line (graphic rating) to represent the judgment. In both cases, a higher rating was represented darker. The graphic rating procedure produced results indicating a linear averaging procedure. Magnitude estimation results indicated some nonlinearity, especially at the dark end of the scale. There are several possibilities for this discrepancy. Observers may have been averaging lightness rather than brightness (see Landauer & Rodger, 1964), which would likely invalidate the perceptual uniformity of the gray-scale values for the stimuli used (Munsell Neutral Chps). In addition, the graphic scale was bounded and the magnitude estimation scale unbounded (the largest discrepancies from additivity are at the dark end (high-magnitude estimates), which might indicate context/contrast effects. It is noteworthy that as the range between the two grays to be averaged increased, accuracy decreased (see Wallin, 1912). Further work on averaging of brightness would elucidate the relation between single-item rating and average rating. In a related study (Kuriki, 2004), observers were asked to produce by adjustment a color in a uniform patch so as to match the average color of tessellated pattern containing several other colors. The color set in the uniform patch was close to the arithmetic mean of the chromaticities of the component colors. The colors were measured in CIE 1976 UCS coordinates, which provided an approximately rectilinear uniform perceptual color space. This is akin to scaling other perceptual dimensions by the appropriate Stevens exponent in order to make sure that departures from normative in the average are not the result of an inappropriate application of a linear physical scale onto a logarithmic psychological scale.

**Numbers in real-world context**

Pollard (1984) provides extensive treatment of the contributions of Levin (1974a, 1974b, 1975) and Levin, Ims, Simpson, and Kim (1977), so only the highlights of those articles are given here. Levin’s approach was to use symbolic stimuli (numerals) but to provide a context by instructing the observers to think of the numbers as, for example, price increases, IQ scores, or academic test results (see also Malmi & Samson, 1983). Ensembles of numbers were presented to observers who were to integrate the values toward deciding the desirability of one store versus another based on price increases, the likelihood of academic success based on test scores, and so forth. Levin motivated his series of investigations as extensions of Peterson and Beach’s “statistical man.” Levin (1974a) required observers to view two cards containing 20 IQ scores each for 10 s and from these they were to estimate (oral report) the mean IQ of the school from which the two samples had been drawn. Inferred school IQs were very close to the arithmetic mean of the card means (or of the 40 values) within 1% to 2%. He repeated the experiment including a group that was just supposed to report a descriptive mean rather than an inferential mean and using cards with unequal numerosity of IQ scores. Again, performance was well predicted by the arithmetic average in all cases. More important than just the approximate normative values in averaging is the fact that observers weighted the card with the greater numerosity of IQ scores more heavily than the card with fewer values. This demonstrates an intuitive understanding of the value of increased sample size. Levin (1975) replicated these effects using rating scales (favorability of a store based on price increases) and demonstrated subjective weighting of the components in the average according to their variability in intuitive averaging of simulated test scores but only when advised to do so. An unbiased group of observers failed to do this spontaneously. Troutman and Shanteau (1976) found that observers (expectant couples) averaged (as opposed to summed) simulated expert ratings for durability/absorbency of overall quality assessment of diapers and design/convenience ratings for car seats.

Simms (1978) reports a similar result for judged overall favorability impressions of outstanding women using verbal descriptions containing various numbers of low-, medium-, and high-favorability statements. Dougherty and Shanteau (1999) found that subjective assessments of quality for consumer goods (including corn chips, hand lotion, and men’s cologne) are based on averaging biasing quality information on the product label with assessments of quality based on actually testing (taste, smell, etc.) the product.

The lessons learned in general from the earlier work were that to a first approximation, observers can compute ensemble statistics and that the ensemble is
treated as an item as far as the psychophysical assessment of a statistic is concerned. A reasonable assumption is that if proper scaling is taken into consideration, the arithmetic mean is a good candidate for the perceptual average. Thus, a whole new psychophysics of combination is not required. This should be taken as tentative because the parametric exploration of various property integration has not been done.

**Speed/depth/direction/spatial frequency**

Morgan (1977) challenged the common explanation for the Pulfrich Phenomenon (a pendulum oscillating in a plane appears to sweep an elliptical pattern if the view to one eye is darkened with a filter) with a spatial direction-averaging model. In this case, the ensemble contains the multiple views/locations occupied by the stimulus within an eye over some temporal interval. Because the mean visual direction of one view is delayed, a depth interpretation occurs in the visual system. Neill (1981) investigated the mechanism behind another binocular depth illusion: the dynamic visual noise stereophenomenon. Briefly, the stimulus is a pattern of random dots such as might be seen on an analog television with no station tuned (occasionally called “snow”). The pattern is shown with a slight time lag between the view that each eye receives. The visual experience is of two or more sheets of dots sliding in different directions. Neill describes a model in which the perceived location of an object (perhaps a dot or dot cluster) is at the average of the locations it occupied over a small window of time. Thus, there is a spatiotemporal ensemble whose average value corresponds to the perceived location. Because the two eyes are integrating the same image at slightly different times, the two averages differ. Neill reasoned that if the averaging window were the same in both eyes, then it should not matter at what frame rate the patterns were presented: The average location would be constant. He found no systematic effect of frame rates across 20 to 120 Hz for velocity matches to the perceived velocity of motion. Verghese and Stone (1995) found that more samples (one to six spatially discrete moving Gabors) reduced speed discrimination thresholds. Although the integration function was not definitive, the average was suggested as a candidate. More importantly, the finding that increasing the number of property samples available effectively reduces thresholds (here for speed) is consistent with Heelley’s result for spatial-frequency discrimination. Heeley (1987) found that increasing the number of cycles of a sinewave grating that were visible to an observer allowed for better discrimination (detection of smaller spatial frequency differences between two gratings). Thus, discrimination was improved by spatially averaging multiple periods across the stimuli. Verghese and Stone (1995, 1996, 1997) suggest that the visual system needs to interpret the samples as discrete entities rather than just as an extended sample in order to benefit from the redundancy.

Watamaniuk and Duchon (1992) used random-dot cinematograms whose motion properties were carefully controlled for mean or modal speed. Observers viewed pairs of cinematograms and decided which one of the two had the higher overall speed. The mean speed of dots determined the difference thresholds. Further, the variance of the component speeds had little effect on thresholds. Note that this does not mean that variance is not coded or represented; rather, it may not have been relevant in this task, or the variances used were too small to affect performance. Watamaniuk and Sekuler (1992) used similar displays in a direction discrimination experiment. They report that global direction is perceived as approximating the mean of the component directions and that increased variance in the component directions did elevate thresholds. Interestingly, the temporal integration interval for the averages estimated from their results was about 500 ms, and the threshold (especially for high-variance displays) decreased as more area of the display was visible. Atchley and Andersen (1995) showed that observers were able to detect differences in the mean (first moment) and variance (second moment) but not higher moments in fields of dots drifting in the same direction with carefully controlled underlying speed distributions. Stoloff (1969) had also found sensitivity to the second moment in an experiment measuring the ability of observers to detect differences between two patterns of elements (textures). Curran and Braddick (2000) demonstrated that both speed judgments and direction judgments for yoked pairs of dots were the vector average of the two velocities.

Dakin and Watt (1997) and Dakin (1997) report sensitivity to mean orientation values when observers estimate global orientation for textures containing heterogeneous orientations. In fact, Dakin (1997) provides evidence that the mean of interest is the arithmetic mean rather than other possible means, such as the geometric or harmonic mean.

Another interesting spatial averaging result is presented in Cheng, Spetch, and Miceli (1996). Observers tried to maximize points earned by pressing a button according to a spatiotemporal rule that they were to intuit from training sessions. The time-space rule was defined by travel time or travel distance (redundant) of an item that moved across a computer monitor. For example, in one training condition, the speed of the moving item was 1 cm/s. If a response was given at 10 s (time rule) or 10 cm of travel (position rule), the points increased. During the test phase, the
speed of travel was faster or slower than trained on some trials. On these exception trials, observers used a compromise between the time and distance rule. The compromise decision rule was a weighted average of both the time and distance criteria.

Allan and Gibbon (1991) found evidence for the geometric mean as the bisection point for tone pairs with tone durations on the order of just under 1 s to 6 s. Wearden (1991), Wearden and Ferrara (1996), and Wearden, Rogers, and Thomas (1997) also found evidence that in some cases, the bisection point was below the arithmetic mean and approaching the geometric mean. There was a tendency for this to occur when the intermediate stimuli were logarithmically spaced whereas linearly spaced stimuli produced bisections that were more arithmetic mean–like. Other stimulus effects such as range or ratio of the endpoints (large versus small) were also evident.

Conclusion

To summarize, the effects of various stimulus parameters on average estimation had been formally investigated numerous times in the 50 years leading up to Ariely’s (2001) article. The lessons to be learned from the various parameter manipulations over that period are summarized above, with the caveat that symbolic materials may be of a different kind than are perceptual attributes. Further, the ensuing work (that is, following Ariely) has more densely sampled the parameter spaces for some of the spatiotemporal properties under earlier investigation. One can hope that the combination of newer presentation technologies, especially for nonstatic displays, more sophisticated eye-tracking and imaging techniques, and insights gained in looking at more than a century of intuitive statistician research have conveyed their benefits in leading toward elucidation of the averaging algorithm(s) and the mechanism that performs this computation. Some lively debates regarding the scope of the putative sampler/averager and even its existence have sharpened the questions being asked and reminded researchers to consider scaling issues, experimenter and observer bias, and the multidimensional nature of stimuli and ensembles.

In the Introduction, several paragraphs were devoted to expounding the ecological perspective (which, according to Slovic, Fischhoff, & Lichtenstein, 1977, was emerging out of the prurient normative focus) as a good starting point for ensemble averaging research. It is telling that Barlow (2001) has switched horses from the compression theories based on Shannon’s theory to a more ecological perspective based on computation and combination of statistics gleaned from experience in the environment. His thesis is that “natural stimuli have to be classified according to their statistics in a way that allows the resulting items to be separately counted and have their probabilities estimated.” (p. 248). These sound remarkably like ensemble statistic computation and frequency encoding. He closes his article by observing that perception is a learning and learned process whose input is regularities in the environment that researchers should ask, “What probabilities are needed? How are they represented? How are they estimated? How are they modified? How are they transmitted to other places in the brain? And how are they combined for making the moderately rational decisions that we observe brains making?” (p. 250). To summarize his question, “How does the brain do these things?” One answer is, on average, “very well.”

Keywords: ensemble processing, intuitive statistics, perceptual averaging

Acknowledgments

Commercial relationships: none.
Corresponding author: Ben Bauer.
Email: benbauer@trentu.ca.
Address: Department of Psychology, Trent University, Oshawa, Ontario, Canada.

Footnotes

1 Richard G. Swenson’s surname is printed as Swenson in Beach and Swenson (1966) and in Pollard’s (1984) review. Peterson and Beach (1967) and Beach and Swensson (1967) use Swensson, which is the correct spelling.

2 Irwin and Smith (1956) contains two errors addressed in a later volume. Specifically, in the Results and Discussion section and in the Summary section, the statements about the number of cards needed for stopping should declare an inverse rather than direct relationship to the average value of the mean.

References

Algom, D., Wolf, Y., & Bergman, B. (1985). Integra-


Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-


Laming, J., & Laming, D. (1996). *J. Plateau: On the measurement of physical sensations and on the law which links the intensity of these sensations to the intensity of the source.* *Psychological Research,* 59, 134–144.


