

Running head: MEAN FOR ALL SEASONS

A Mean For All Seasons

David J. Weiss

California State University, Los Angeles

Ward Edwards

University of Southern California

Abstract

Averaging scores differs from averaging numbers in that behavioral issues are built into scores. The behavioral issues are the weight attached to a score and the metric on which the scores have been gathered. A single equation is proposed, derived from Aczél's (1966) model of the quasilinear mean, that encompasses the standard measures of central tendency. The equation allows for differential weighting of scores, and also addresses the metric issue by incorporating response transformation.

A Mean For All Seasons

A frequent research agenda for psychologists is to establish the empirical connection between stimulus and response. Response variability often threatens to obscure the connection. To surmount that challenge, behavioral researchers combine scores over subjects and/or occasions. The responses to a stimulus can be represented with a single value, an average. Behavioral researchers inherited this methodology from astronomers, who began the study of statistics several hundred years ago (Stigler, 1986), and it has become standard practice in psychology.

What is the criterion by which one single-valued summary statistic better represents the scores than another? Why not simply apply the same statistic in all cases? If we were averaging abstract numbers, we would follow the standard procedure of using the arithmetic mean, and there would be little to discuss. However, behavioral issues attached to scores can make them not be comparable, so that all scores, though they should be considered in the average, ought not to contribute to it in the same way. The purpose of this essay is to show how the researcher's stance regarding those behavioral issues suggests the appropriate average.

The two behavioral issues are the weight that each score should carry, and the metric that has been used to gather the scores. All of the usual measures of centrality can be expressed in a single equation. The equation has two elements built into it, reflecting the two pertinent issues that might be attached to a score. These two elements provide the flexibility that allows one expression to describe the variety of averaging formulas in common usage. An average is a weighted sum applied to values represented on a distance metric, wherein the sum of the weights is one. The equation will be more comprehensible if we first examine how weights and distance metrics operate.

Weights

The traditional psychological counterpart of weight is importance. It is linguistically more neutral to think of weights simply as loadings, based on relative contribution of some kind. A suitable average should compensate for unequal contributions. Scores that incorporate more observations or are deemed more reliable might count more heavily. It might also be feasible to employ weights estimated within a prior study that validated a cognitive model (Zalinski & Anderson, 1991). For example, in combining predictions (Clemen, 1989), one might apply source credibility weights (Rosenbaum & Levin, 1969) estimated for forecasters with varied degrees of experience.

Consider the batting average of a baseball team or subset of its players. Computing the ordinary mean of the individual batting averages (BAs, defined as the ratio of hits to batting opportunities¹) yields an incorrect value for the team as a unit, because the players need not have batted an equal number of times. The appropriate rule that combines individual performance into a mean can represent two separate aspects of the data collection process: one is individual batting performance, and the other is number of opportunities to bat. Let's say there are five players, whose individual averages are shown in Table 1.

Insert Table 1 here

The mean of the individual BAs is $(.200 + .400 + .300 + .100 + .250) / 5 = .250$, but that is misleading because the overall BA for the team is $142/760 = .187$. A correct group result is

obtained if the individual BAs are weighted by the player's proportion of the total of 760 opportunities:

$$\frac{10}{760} \times .200 + \frac{50}{760} \times .400 + \frac{100}{760} \times .300 + \frac{400}{760} \times .100 + \frac{200}{760} \times .250 = .187$$

This average is meaningful in the sense that it informs us how well the team is batting as a whole, which is relevant to baseball success. If we want to determine whether the team's new batting coach got better performance out of the players than last year's coach, comparing via the weighted average is appropriate. Before dismissing the arithmetic mean as meaningless, though, we note that it estimates the probability that a randomly selected team member will get a hit at his next opportunity. That prediction has little value in the world of baseball, so we need not bother with the statistic. The important point is that substantive knowledge is required to dismiss it, in this case knowledge of the fact that batters are not chosen randomly from the population of team members. This illustrates the fact that different means for different purposes can be calculated from the same data. The appropriate measure of central tendency depends upon the analyst's purpose.

An Illustration – Weighting Self-Reports

Suppose a researcher has conducted a study whose aim is to contrast two alternative diet and exercise regimens designed to reduce obesity. At the beginning of the study, volunteers come to the laboratory. Body Mass Indices (BMIs) are measured prior to assignment to the experimental conditions. At the end of the study period, the patients are supposed to return to the laboratory to have BMIs assessed again.

However, a substantial number of the patients in both conditions are unable to return to the laboratory and instead phone in their current BMIs, determined using their home scales. The

researcher is aware from the literature that people are not always accurate in reporting BMI under these circumstances.

One may, of course, simply accept the self-reports at face value. If the researcher is willing to do so, then no analytic adjustment is in order; the ordinary, equal-weight mean is appropriate. A cynical researcher might simply add a fixed percentage to the reported values, but there is no clear-cut justification for assuming that all patients under-report. A more defensible policy is to give less credibility to the phone reports via weighting. For example, one might decide that a phone report is half as credible as a direct observation.

With equal weighting, each score within a group is implicitly assigned a weight of $1/n$, where n is the number of scores in the group. To assign unequal weights in accord with a predetermined ratio (2 to 1 for direct observations vs. phone reports in this example), count the number of scores in each subgroup. For example, suppose there are six directly observed scores (25, 25, 25, 30, 30, 30) and four scores reported by phone (23, 23, 29, 29). Multiply the numbers of scores by 2 and 1 respectively, so that the total is $(2 \times 6) + (1 \times 4) = 16$. Divide the multiplier by this sum to get the weight for each score. Thus, each of the six directly observed scores is assigned a weight of $2/16$, and each of the four phone reports is assigned a weight of $1/16$. Accordingly, the weighted mean BMI for the ten scores is 27.125. Note that the weighted mean is slightly higher than the unweighted mean BMI of 26.9, reflecting the lower credibility given to the phone scores.

The weighting scheme, assigning half as much weight to scores reported by phone, was adopted arbitrarily in the example. We are not very knowledgeable about translating the construct of credibility into specific weights. If we had instead presumed that phone reports merit one-third the weight of directly observed scores (the calculated weights would be $3/22$ for

direct observations and 1/22 for phone reports), the resulting weighted mean BMI would have been 27.23. If the phone reports had been given zero weight, so that they were completely ignored, then the weighted mean BMI would have been 27.5.

Our example illustrates that weights are often determined subjectively, although there are obvious risks. We might get better estimates of the appropriate weights if we could elicit opinions from experts (Von Winterfeldt & Edwards, 1986) in the telephone survey domain. Researchers who employ experts for this purpose should speak to their accreditation when reporting the results. The alternative, objective approach of determining weights using regression analysis allows variability, in the form of intersubject agreement, to contribute to the weighting. Our view is that extent of consensus is not pertinent to the importance of a score.

Distance metric

The problem of the distance metric arises when the scores we observe are not linearly related to those on the respondent's internal scale. That is, something about the judgmental task or elicitation technique distorts the respondent's translation of interpoint distances, so that equal intervals on the overt response scale do not reflect equal subjective intervals. Systematic distortion may result from a behavioral process that we can understand. Analysis of that process can suggest a transformation that will undo the distortion. Rather than seeking a specialized statistic, we can instead use that transformation and average in a familiar way. The procedure is to transform the scores, compute a (possibly weighted) mean, and then apply the inverse transformation to the result.

Transformations have long been advocated for various statistical purposes, particularly to stabilize variance and to reduce non-normality. In fact, transforming and weighting have both been proposed in specific averaging contexts, with similar statistical motivation. For the

averaging of correlation coefficients, Fisher's z transformation has been a standard recommendation. Silver and Dunlap (1987) and Strube (1988) reported that backtransformed z transformations of r were less biased than simple averages of r . Stanley and Porter (1972) added weighting, in this case by sub-sample size, to the discussion. Because correlation is not a direct function of a psychological distance, the present approach does not suggest a particular transformation.

Generalized Averaging Equation

Weiss (1975) discussed a class of models for averaging derived from what Aczél (1966) calls the quasilinear weighted mean. This class includes the geometric mean, harmonic mean, and root-mean-power. If we extend Aczél's bisection equation to accommodate n scores, it is the answer to our quest for the general expression for the average response:

$$\bar{R} \equiv f^{-1} \left[\frac{w_1 f(R_1) + w_2 f(R_2) + \dots + w_n f(R_n)}{w_1 + w_2 + \dots + w_n} \right], \text{ where all } w_i \geq 0 \text{ and } \sum w_i = 1 \quad (1)$$

The equation incorporates possible differential weighting via the w_i . It also provides for a possible nonlinear distance metric via the transformation function f . Application of the inverse transformation f^{-1} ensures that the measure of central tendency has the same units as the original responses.

Examples of f that lead to well-known measures are $f(R) = \log(R)$, the logarithmic transformation, and $f(R) = R^2$, the root-mean-square. Whenever f is a continuous, strictly monotonic function that has an inverse, an average can be defined, though only a few have specific names. Psychologically, f^{-1} is the function that the task induces the subject to apply in translating from internal representation to overt response. Thus, both of the adjusting elements in the equation, the w_i and f , have behavioral meaning.

There are three elements to the equation – weights, transformation function and observations. Nothing else enters into the averaging formulation. All of the usual measures of central tendency, including the median, trimmed mean and mode, can be seen as instances of the equation; the latter statistics assign zero weight to observations whose identity can be determined only after the data are examined.

To Transform or Not To Transform

How is the researcher to know whether transformation is needed? Our statistical colleagues have argued that data themselves proclaim their need to be transformed (Box & Cox, 1964; Hinkley & Runger, 1984). According to this perspective, ideal data have three properties: normal distribution of errors, homogeneity of variance, and simplicity of interpretation (e.g., main effects without interaction) (Bartlett, 1947). The analyst transforms the observations in the hope of obtaining a good fit to a linear model. Box and Cox (1964) illustrated the efficacy of a particular family of transformations, indexed by a small number of parameters, that did remarkably well in obtaining the desired results. There is no guarantee that all three characteristics of an ideal representation will be achieved, but it is not entirely coincidental that the same transformation fulfills multiple objectives. If the data have the hypothesized underlying structure and the observations have been distorted in a way that the transformation can undo, then the approach will achieve its goals.

Underlying our perspective is a belief that the respondents' experience and the demands imposed by the task combine to generate an internal continuum along which values are located. An observed response is a projection from the internal value, the "true" response, inspired by the stimulus onto the scale imposed by the researcher. The unstated contract, assumed by both researcher and respondent, is for the projection to be as close to linear as possible. If the

projection is linear, then the researcher can analyze the responses as the respondent intended to express them. The researcher's goal in averaging observations across either trials or participants is to overcome the variability inherent in both the selection of the internal value and the projection process, and thereby to estimate the typical projected response to each stimulus.

Instruction and training may help the respondent to achieve the desired linear projection. However, some instruments or tasks may consistently induce nonlinear projections; in effect, the transformation function f is applied to the true responses. If the observed responses are not linearly related to the internal responses, it is appropriate to undo that distortion. But since the researcher does not have direct access to the true values buried inside the respondents' heads, how does one know whether distortion has occurred?

There are two plausible reasons to decide that the data to be collected will need to be transformed. First, the researcher may anticipate that a task will impose systematic distortion. A transformation that will undo the distortion can be specified in advance. For example, Weiss and Gardner (1979) squared responses in a study of subjective hypotenuse estimation, in accord with the normative view that the judgment calls for internal computation of the square root of apparent lengths. Similarly, Tversky and Russo (1969) analyzed logarithms of judgments of apparent size, reasoning that people respond using distance terms but envision areas as products of distances.

An empirical inference that distortion has occurred is available when raw data exploring the same underlying construct, but collected using different response tasks, do not exhibit the expected congruence (Birnbau & Veit, 1974). The researcher can decide, on extra-statistical grounds or perhaps even arbitrarily, which response mode will be considered the standard. Then

when another mode is used for data collection (perhaps because respondents find it comfortable), the transformation that has united the results from previous studies can be specified in advance.

The alternative justification for transformation is the researcher's conviction that a particular process governs generation of the responses at an internal level, but the elicitation method may distort the internal responses in an unknown, albeit systematic, way. A transformation is sought that provides agreement with data patterns predicted by the behavioral model, given the usual restriction that the transformation be monotone (Krantz, Luce, Suppes, & Tversky, 1971). If the model adequately describes the transformed data, then the obtained transformation may be applied routinely when similar stimuli and response procedures are used in future studies. This approach may be traced to seminal papers by Anderson (1962) and by Luce and Tukey (1964). An example illustrating transformation to additivity for a bisection task is given in Weiss (1975). Additivity was sought not to find a structurally simple way to describe the data, but because an additive model was theorized to be an appropriate description of the behavior at the internal level. The model provides the leverage to derive the transformation connecting internal responses to their observed counterparts. The raw data were not additive, and had not been expected to be. If no monotone transformation could bring about additivity, then the model would be falsified. Rather than being specified in advance, the form of the successful transformation was an experimental outcome. Because the transformation is data-dependent, it is subject to sampling error and therefore reproducibility is a concern.

These arguments for transformation are psychological rather than statistical. We acknowledge circumstances where transformation can be justified statistically, with an eye toward increasing power when comparisons are to be made (Anderson, 2001; Levine & Dunlap, 1983; Ratcliff, 1993). Typically, such cases arise when the response is a physical measure rather

than an opinion expressed by a subject. We see no behavioral reason to prefer measuring response time rather than its reciprocal, response speed (Anderson, 1961), so transforming to the scale that yields greater power is not objectionable. A physical scale is “essentially an arbitrary choice of the scientist or instrument maker” (Mandel, 1976). There are grounds in the literature for making a prior determination in favor of speed measures, which generate more power (Levine & Dunlap, 1982). Power is affected because transformations alter variances as well as means.

Example of Transformation

The geometric mean (the n -th root of the product of n numbers) is the classically recommended (Stevens, 1955) way to average magnitude estimates. Magnitude estimation is a method for eliciting subjective magnitude championed by Stevens (1958). The respondent is instructed to assign numbers to stimuli such that the numbers are proportional to apparent magnitude. Stevens’s rationale for using the geometric mean is that responses are often observed to be approximately normally distributed when plotted on a logarithmic scale. It would seem equally plausible to apply the transformational approach, in this case computing the mean of the logarithms of the responses. The (distance metric) rationale for transformation is the behavioral theory that the magnitude estimation task induces a power response function on true subjective magnitude (Birnbaum & Veit, 1974). These two ways of handling the data, either by computing the geometric mean or by computing the mean of the logarithms, are mathematically equivalent. If the responses to a stimulus are 16, 24, and 36, then the geometric mean, the cube root of $16 \times 24 \times 36$, is 24. Equivalently, e to the $(\ln 16 + \ln 24 + \ln 36)/3$ power is 24.

The Median

The median is the appropriate measure of central tendency when there is no applicable distance metric, so that validity attaches only to the ordering of the scores and not to their

magnitude. Indeed, for ordinal data, only order statistics such as the median satisfy the plausible criterion of “comparison meaningfulness” (Ovchinnikov, 1996; Ovchinnikov & Dukhovny, 2002). However, use of the median has also been recommended when numerical data come from a skewed distribution (Stevens, 1955).

What is the average of a set of salaries? Suppose the employees of a department within a corporation are paid as follows:

\$14,400; \$14,400; \$14,400; \$14,400; \$16,800; \$16,800; \$16,800; \$18,000; \$19,200; \$19,200; \$19,200; \$24,000; \$26,400; \$54,000; \$72,000. All but the last two employees are clerks; the last two are Assistant Manager and Manager.

The first thought is to calculate the arithmetic mean of the 15 salaries - \$24,000. How can we judge the appropriateness of that number? Appropriateness for what purpose? Someone who wanted to argue for increasing those salaries might argue that inclusion of the last two figures makes the mean too large. That person might prefer the median, which is \$18,000. The company’s management might prefer the mean. Either number is technically acceptable.

Other technically acceptable numbers can be calculated for other purposes. Suppose we argue that the way one thinks about a salary should depend on its utility. If we accept Galanter’s (1990) conclusion that utility is approximately the square root of dollars, then the mean salary to which that conclusion leads is \$22,237. If instead we use the utility function proposed by Breault (1983), a power function with exponent .43, the mean salary is \$22,027. In our view, all of these means are acceptable. If we were consultants, we might well choose the one that seems most defensible and still serves our purpose best.

A mean represents a “typical” score – but what does “typical” mean? The answer is not obvious. Different criteria of typicality lead to different calculations. We are proposing lines of

thought to guide choice among them. But the lines do not provide unequivocal answers.

Calculation of means should be based on Equation 1 and an underlying theoretical structure.

The fact that two utility functions were deployed in the example highlights the limitation of the proposed weighted mean/distance metric approach, namely that the analyst needs to know the appropriate weighting and transformation. Incorporating prior knowledge into a statistical formulation is a concept with which Bayesians (Edwards, Lindman, & Savage, 1963) are already comfortable. Here we extend that idea from the world of inferential statistics to that of descriptive statistics.

Behavioral Indices

The impetus for this re-examination of averaging was a need to average observed CWS scores (Weiss & Shanteau, 2003). The CWS index is an empirical measure of judgmental expertise that does not depend on knowing correct answers. Instead, the index captures a candidate's ability to assign different ratings to different stimuli and similar ratings to similar stimuli. With an experimental design comprised of repeated presentations of a set of various stimuli to an individual, the CWS index may be computed as the ratio of between-stimulus variance to within-stimulus variance. When we first began representing the typical expertise of a group of professionals using the new index, we found that the arithmetic mean of a set of CWS scores could be heavily influenced by extreme values, just as in the salary example above. We then thought we might borrow a better method for averaging CWS scores from the standard procedure for averaging F-ratios. A search of the literature revealed no extant recommendation.

We illustrate how the generalized averaging equation can be used to derive an appropriate mean. In the case of F-ratios, the scores do not comprise a collection of individual responses, but values of an index derived from a set of responses. The transformational approach we espouse

calls for use of the square root transformation, because the values to be averaged involve squared original distances. The key constituent of an F-ratio, the sum of squares for an effect, is related to the square of a psychological distance as expressed by the respondent (SS_A is proportional to $\sum \Delta_A^2$, where the Δ_A 's are the differences between marginal means for the levels of the variable A). Therefore, we propose that the average F-ratio is the square of the mean of the square roots of the individual F's. Our aim is to average on the continuum on which lie the behavioral values to be summarized.

Choosing a Transformation

Mellers and Hartka (1989) reanalysed and replicated a study by Anderson (1976), in which participants were given information about how well two people had worked on a job and how much one of them had been paid. The participants were asked to determine the “fair” payment for the other worker. Both sets of data could be monotonically transformed to additivity, a result that Mellers and Hartka interpreted as supporting a subtractive model (which despite its name is an instance of additivity, because the model implies no interaction) with a nonlinear response function. Mellers and Hartka noted that a doubly bounded response scale, such as a standard rating scale, could be expected to show floor and ceiling effects that could be transformed away.

Rather than allowing the MONANOVA algorithm (Kruskal and Carmone, 1969) to search for a transformation that would work, we would propose that an arcsin transformation applied to the responses might have brought the means into conformity with the proposed subtractive model. That particular transformation has long been used with sigmoidal data (Cochran, 1940), because the arcsin has little effect on intermediate values, but stretches out the ends of the scale. Specifying the transformation in advance is a more definitive statement, one

justified by Mellers and Hartka's (1989) account of the distortion induced by the response scale. If this hypothetical operation had been able to make the transformed means plot in accord with the parallel pattern called for by the model, then subsequent research using tasks in which respondents make similar judgments (e.g. Singh, 1995) might fruitfully apply the same transformation. Note that additivity is sought in this setting because the model is held to describe the psychological process by which people judge equity, not because an additive representation is structurally simple.

Concluding Remarks

In 1955, Stevens published a landmark paper that discussed several alternative measures of central tendency. Stevens (1955) provided a theoretically grounded taxonomy of the measures, but at the same time conveyed the sense that one must look at the data to make the best choice for a particular application. We certainly do not wish to speak against the idea of scrutinizing one's data, but that guideline has always seemed insufficient. Why, for example, should the harmonic mean be used to average speeds, or the geometric mean to average magnitude estimates? Sometime a median was deemed best. It all seemed so arbitrary, as though one needed to have absorbed the appropriate scientific folklore to choose correctly among competing descriptive statistics.

Equation 1 formalizes the folklore. The researcher's duty includes specification of the behavioral theory that justifies a particular weighting scheme or transformation. We link the mathematical machinery, the weighting and distance metrics that distinguish one summary statistic from another, to behavioral constructs. In doing so, we are following a tradition that stresses the importance of substantive theory in resolving measurement problems (Anderson, 1979; Luce, 1996).

In proposing substitutes for the arithmetic mean, Stevens's (1955, p. 113) expressed goal was to undo the bias that results from skewed distributions. Skewness was viewed as a sign of a nonlinear relation between observed values and the underlying variable of interest. The strategy appears predicated on a belief that nature provides normal distributions when unobservable psychological quantities are measured properly. Stevens intuitively grasped the importance of the distance metric when he argued that it is "more sensible to average loudness than to average decibels". Our interpretation of that statement is that Stevens appreciated that averaging should take place on the psychological continuum, even when responses are expressed in physical units.

Our recommendations for averaging data are not inconsistent with those of Stevens, though our justification is quite different. It is the generalized averaging equation, a model for averaging, that supports transformation and unequal weighting. We might choose to apply a logarithmic transformation to magnitude estimates of individual stimuli (or equivalently, to present geometric means as average responses), just as Stevens would, but not in order to undo skewness or to stabilize variance. In selecting a prior logarithmic transformation, we would be accepting Birnbaum and Veit's (1974) contention that the task itself induces a transformation on the internal responses. Our objective is to average those internal responses, a goal Stevens understood but did not regard as primary. Alternatively, if we employed an integration task in which respondents provided magnitude estimates of the combined value of two stimuli, the additional leverage provided by a cognitive model would allow us empirically to estimate the transformation that maximized agreement with the model (Weiss, 1975). That transformation might turn out to be logarithmic, though it need not. In fact, for magnitude estimates of the average darkness of two gray chips, geometric means did not yield results consistent with an additive model (Weiss, 1972).

We also differ from Stevens in that we do not advocate presenting a transformed version of the data. The inverse transformation is applied to the average of the transformed scores (Edwards, 1966), so that a mean obtained from the general averaging equation maintains the units of the original responses. When we plot mean responses against the independent variable, we employ the units in which the data were collected. If individual responses were expressed as probability estimates or as ratings on a particular scale, then so ought to be the mean that describes a typical response across subjects or trials.

How one chooses to integrate the scores makes a difference. At the heart of science are the empirical laws that researchers discover. An empirical law is the observed functional relationship between a set of stimulus objects and the typical response inspired by each of those objects. However, participants provide individual responses; it is the researcher who decides how to represent those responses. Any average described by Equation 1, including the customary default choice, the arithmetic mean, will yield an interpretable value.

If the researcher makes a good case that a particular law ought to connect the values of a quantitative independent variable and the typical responses, the fact that the data are consistent with that hypothesized connection when one mean is employed, and are not when a another mean is employed, suggests that the former is a preferable summary statistic. The enduring controversy over the form of the psychophysical function (Krueger, 1989) may be in part attributable to variations among the averaging methods used by researchers (Myung, Kim, & Pitt, 2000).

Our take-home message is that we ourselves do not hesitate to depart from the routine practice of employing arithmetic means when describing data. At the same time, we consider it our responsibility to justify the departure not merely by citing tradition, but by articulating a theoretical stance. Just as researchers use prior knowledge and theory to choose the behavior to

observe, so too must they exercise judgment in choosing the statistic that most accurately exemplifies the pertinent group of responses. As consumers of research, we tend to be sympathetic to any well-reasoned choice.

The present discussion has been concerned with descriptive rather than inferential statistics. If weighting or transformation is needed to depict a typical value in a given context, then one might consider the same operation to be appropriate for scores submitted to significance tests. The effect of transformation on comparisons of means has been examined for many years (Bartlett, 1947; Doksum & Wong, 1983), but the study of unequal weighting is only getting started (Wilcox, 2003).

References

- Aczél, J. (1966). *Lectures on functional equations and their applications*. New York: Academic Press.
- Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin*, 58, 305-316.
- Anderson, N. H. (1962). On the quantification of Miller's conflict theory. *Psychological Review*, 69, 400-414.
- Anderson, N. H. (1976). Equity judgments in information integration theory. *Journal of Personality and Social Psychology*, 33, 291-299.
- Anderson, N. H. (1979). Algebraic rules in psychological measurement. *American Scientist*, 67, 555-563.
- Anderson, N. H. (2001). *Empirical direction in design and analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bartlett, M. S. (1947). The use of transformations. *Biometrics*, 3, 39-52.
- Birnbaum, M. H., & Veit, C. T. (1974). Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. *Perception & Psychophysics*, 15, 7-15.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- Breault, K. D. (1983). Psychophysical measurement and the validity of the modern economic approach: A presentation of methods and preliminary experiments. *Social Science Research*, 12, 187-203.

- Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, 5, 559-583.
- Cochran, W. G. (1940). The analysis of variance when experimental errors follow the Poisson or binomial laws. *Annals of Mathematical Statistics*, 14, 335-347.
- Doksum, K. A., & Wong, C. W. (1983). Statistical tests based on transformed data. *Journal of the American Statistical Association*, 78, 411-417.
- Edwards, W. (1966). Introduction. *IEEE Transactions on Human Factors in Electronics*, 7, 1-6.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Galanter, E. (1990). Utility functions for nonmonetary events. *American Journal of Psychology*, 103, 449-470.
- Hinkley, D. V., & Runger, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association*, 79, 302-309.
- Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement* (Vol. 1). New York: Academic Press.
- Krueger, L. E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, 12, 251-329.
- Kruskal, J. B., & Carmone, F. L. (1969). MONANOVA: A FORTRAN IV program for monotone analysis of variance. *Behavioral Science*, 14, 165-166.
- Levine, D. W., & Dunlap, W. P. (1982). Power of the F test with skewed data: Should one transform or not? *Psychological Bulletin*, 92, 272-280.
- Levine, D. W., & Dunlap, W. P. (1983). Data transformations, power, and skew: A rejoinder to Games. *Psychological Bulletin*, 93, 596-599.

- Luce, R. D. (1996). The ongoing dialog between empirical science and measurement theory. *Journal of Mathematical Psychology, 40*, 78-98.
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.
- Mandel, J. (1976). Models, transformations of scale, and weighting. *Journal of Quality Technology, 8*, 86-97.
- Mellers, B., & Hartka, E. (1989). Test of a subtractive model of “fair” allocations. *Journal of Personality and Social Psychology, 56*, 691-697.
- Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory and Cognition, 28*, 832-840.
- Ovchinnikov, S. (1996). Means on ordered sets. *Mathematical Social Sciences, 32*, 39-56.
- Ovchinnikov, S., & Dukhovny, A. (2002). On order invariant aggregation functionals. *Journal of Mathematical Psychology, 46*, 12-18.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510-532.
- Rosenbaum, M. E., & Levin, I. P. (1969). Impression formation as a function of source credibility and the polarity of information. *Journal of Personality and Social Psychology, 12*, 34-37.
- Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology, 72*, 146-148.
- Singh, R. (1995). “Fair” allocations of pay and workload: Tests of a subtractive model with nonlinear judgment function. *Organizational Behavior & Human Decision Processes, 62*, 70-78.

- Stanley, J. C., & Porter, A. C. (1972). ANOVA analysis of unweighted and weighted Fisher z's. *Social Science Research, 1*, 237-241.
- Stevens, S. S. (1955). On the averaging of data. *Science, 121*, 113-116.
- Stevens, S. S. (1958). Problems and methods of psychophysics. *Psychological Bulletin, 54*, 177-196.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.
- Strube, M. J. (1988). Averaging correlation coefficients: Influence of heterogeneity and set size. *Journal of Applied Psychology, 73*, 550-568.
- Tversky, A., & Russo, J. E. (1969). Substitutability and similarity in binary choices. *Journal of Mathematical Psychology, 6*, 1-12.
- Von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. New York: Cambridge University Press.
- Weiss, D. J. (1972). Averaging: An empirical validity criterion for magnitude estimation. *Perception and Psychophysics, 12*, 385-388.
- Weiss, D. J. (1975). Quantifying private events: A functional measurement analysis of equisection. *Perception and Psychophysics, 17*, 351-357.
- Weiss, D. J., & Gardner, G. S. (1979). Subjective hypotenuse estimation: A test of the Pythagorean theorem. *Perceptual and Motor Skills, 48*, 607-615.
- Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors, 45*, 104-116.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.

Zalinski, J., & Anderson, N. H. (1991). Parameter estimation for averaging theory. In N. H. Anderson (Ed.), *Contributions to information integration theory. Vol. I: Cognition* (pp. 353-394). Hillsdale, NJ: Lawrence Erlbaum Associates.

Author's Note

David J. Weiss, Department of Psychology, California State University, Los Angeles.
Ward Edwards, Professor Emeritus, Department of Psychology, University of Southern
California, Los Angeles, California.

Preparation of this manuscript was partially supported by grants 98-G-026 from the
Federal Aviation Administration in the Department of Transportation and FA9550-04-1-0230
from the U. S. Air Force Office of Scientific Research. We wish to thank James Shanteau and
Rick P. Thomas for comments on the manuscript. We also wish to thank an anonymous reviewer
for enhancing our historical perspective.

Correspondence concerning this article should be directed to David J. Weiss, Department
of Psychology, California State University, Los Angeles, 5151 State University Drive, Los
Angeles, CA 90032. email: dweiss@calstatela.edu.

Footnote

¹Batting average as defined by organized baseball omits certain batting opportunities that do not yield hits; among them are walks and sacrifices. This inaccuracy is an attempt to avoid charging the player for a non-hit that may contribute positively to the team's performance. Therefore, batting average does not quite reflect the proportion of hits. Baseball statistics also include another index that weights for the differential value of hits, the slugging percentage, in which home runs count more heavily than singles. The batting average is the figure that leads to an annual individual award.

Table 1

Individual Batting Averages for Five Players

Player	A	B	C	D	E
Record	2 for 10	20 for 50	30 for 100	40 for 400	50 for 200
Individual BA	.200	.400	.300	.100	.250