Perception & Psychophysics
2008, 70 (3), 456-464
doi: 10.3758/pp.70.3.456

Visual statistical decisions

George Fouriezos, Sara Rubenfeld, and Gary Capstick
University of Ottawa, Ottawa, Ontario, Canada

To identify variables that underlie intuitive judgments about the sizes of groups of similar objects, we asked people to judge the relative heights of vertical bars briefly shown, two groups at a time, on a computer display. Randomly selected normal deviates determined individual bar height. Average differences in height and group sizes were also randomly varied. Twenty-eight participants judged 250 differences each, which were then submitted to multiple regression analysis and psychophysical inspection. The total number of bars sharpened discrimination, whereas variance dulled it. *Critical ratio* (CR), the forerunner to the modern *t* test, emerged as the most important predictor; little additional variance was explained by other factors. The difference in the number of bars was a reliable factor, favoring the greater number of bars. Confidence limits around thresholds, defined as CRs needed to say "*possibly* greater," surrounded 1.65; as a *z* value, this corresponded to a one-tailed probability of .05. Judgments about noisy stimuli thus seem to be based on a statistical process and to employ a probability criterion similar to that used in the formal statistical evaluation of experimental findings—namely, $p < .05$.

This study owes its inception to a remark about a physical chemist overheard by one of the present authors at a social gathering of staff and students: "He just *said* the bubbles were bigger. He didn't do any stats at all." Fouriezos joined the conversation by commenting that sometimes a statistical analysis is not required, when a difference is literally visible. This led to wondering what a significant difference—say, of a *t* test—would look like if it were portrayed in a visual display by two groups of similar shapes that were varied in size. Would the eyes be more sensitive than a formal test of significance, or would it be the other way around? Those idle questions have been supplanted by the questions posed in this article: When two sets of similar objects with individual variation of dimension are judged visually, to what variables is the perceptual assessment sensitive? Might the process underlying the judgment be a statistical one?

People make accurate judgments about the average size or average position of several similar objects when the objects are portrayed graphically and when the judgments are not consciously computational. Spencer (1961, 1963) found that his participants could accurately adjust a cursor's height to match the average height of 10 or 20 dots placed at varied heights on graph paper, and that they did so more quickly and with greater precision than when the task was to call out the numeric average of 10 or 20 two-digit numbers viewed on printed cards. The arithmetic mean of the graphically portrayed heights came closer to matching the participants' judgments than did the median or the midpoint of extremes (Spencer, 1963). The arithmetic mean also seems better matched to intuitive judgments of graphic central tendency than does the geometric mean, harmonic mean, or root mean square (Bauer, 2006). Ariely (2001) demonstrated that the average diameter of briefly displayed circles is accurately judged to be greater or smaller than a test circle's diameter and that it is accurately chosen from two test circles, of which one equals the average. Yet, when asked to declare whether a test circle had been a member of the displayed set, or when asked to choose which of two test circles had been among the previously displayed circles, subjects responded at chance levels. Ariely's work suggests that working knowledge of the mean size of visible objects is a preattentive process that occurs without retention of the values of the individual elements. Extensions of Ariely's work replicate that main finding (Chong & Treisman, 2003, 2005a, 2005b). It thus seems that humans can extract average sizes of visual objects using an accurate, effortless, and automatic process.

The purpose of the present experiment was to see how well statistical rules apply to intuitive judgments of group differences in size. At least two conditions should be met for the process to qualify as a statistical one: (1) Accuracy or personal confidence in answers should grow with increases in the number of objects seen, and (2) accuracy or personal confidence in answers should diminish with increases in the apparent variability. Number and variance have been examined, but consensus about their roles has not yet been reached. Spencer (1961, 1963) found no difference in mean estimation with 20 points in view over 10. Legge, Gu, and Luebker (1989) asked participants to choose the greater average or the greater variance of two groups presented symbolically as two-digit values, as the heights of graphic symbols in a scattergram, and as patches of luminance. They found that accuracy in choosing the greater average in scattergram display, for which

G. Fouriezos, georgef@uottawa.ca

efficiency was superior to the numeric and luminance displays, improved with greater numbers of displayed elements—but not at the rate predicted ($\times \sqrt{N}$)—by a statistically ideal observer. The same conclusion about numbers of elements was reached by Sorkin, Mabry, Weldon, and Elvers (1991), who, using a signal detection task, asked their participants to state whether samples of varied vertical positions came from a signal ($\mu = 2.4$, $\sigma = 0.89$, arbitrary units) or noise ($\mu = 1.6$) distribution. Although the magnitude of the perceptual signal (d') did grow, it fell increasingly short of ideal as the number of elements increased. No effect of total number of elements, however, was seen by Chong and Treisman (2005b) when they asked participants to choose the side of circles with the greater mean diameter. The two positive studies (Legge et al., 1989; Sorkin et al., 1991) included smaller sample sizes than did the negative ones (Chong & Treisman, 2005b; Spencer 1961, 1963). If the impact of number does grow in measure to N's square root, the greatest differences will be seen at its low end (see Sorkin et al., 1991, Figures 2–4). The role played by variance has received less attention than has number. Variance in the displayed elements reduces accuracy in graphic estimation of the mean (Spencer, 1961, 1963).

Here we examine the statistical properties of snap judgments made of two groups of visual stimuli of varied size and number. The goals were to identify the implicit calculation of difference performed by the visual system and to examine the roles played by morphological factors in guiding decisions about relative average sizes. Participants were asked to judge the relative average heights of two groups of vertical bars shown beside each other (for an example, see Figure 1). With only a few constraints, the individual heights of the bars were selected at random and recorded along with the participants' judgments. A posteriori reviews of the data using multiple regression analyses and d' plots were then conducted with the aim of understanding how variables like variance, number, overall size, and aspects of the appearance of the display contribute to judgments of group differences.

The nature of the task demanded of participants—to judge which of two fully visible groups of bars had the greater average height—carries a minor and a major implication stemming from the fact that statistical inference to parent populations was not a task demand. The minor implication is that stimulus measures and notation are in population format. The major implication applies to theories about the underlying process. Without preclusion of others, two possibilities are suggested: a comparison of means and a statistical decision. *Comparison of means* posits that the two groups are mentally represented, each by its own average height, and the two heights are compared as two single values to inform the decision. The automatic abstraction of the mean size of each side, as described by Ariely (2001), may be the process used to generate two mental values. If the decision is based on a straight comparison of means, the difference in means would be the best predictor of decisions made. The alternative, *statistical decision*, posits that the decision is based on an involuntary statistical assessment in which the difference in mean heights, the group numbers, and the group variances are all taken into consideration in a way that mimics a t test—or critical ratio (CR), if we adhere to the minor implication. If the decision is based on a statistical evaluation, the raw difference in averages would be amplified by number and reduced by variance. Decisions would be expected to conform better to CR. While the general tone of this project was exploratory, one objective that guided this report was to determine which of these two models provided the better fit to the data.

METHOD

Participants

The participants were two of the authors, a number of staff members and students in the School of Psychology at the University of Ottawa, and some acquaintances of the authors. Of the 34 people

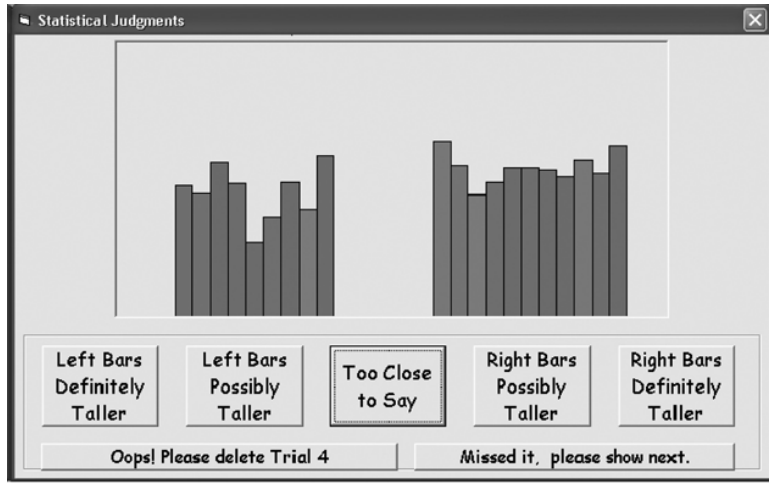

Figure 1. Sample decision trial. Bars are displayed for 1 sec, whereas the response panels are constantly in view.

who volunteered, 28 provided data that met the acceptance criteria described below. Ages ranged from 20 to 70 years. Beginning with participants in their 20s, the numbers of participants by decade were 15, 5, 2, 4, 0, and 1.

Apparatus and Stimuli

The stimuli were presented on the 15-in. (38-cm) LCD screen of a laptop personal computer set to a resolution of 1,024 × 768 pixels. Individual bars were 16 pixels (4.4 mm) wide and averaged 120 pixels (33 mm) in height. Participants were initially seated to view the display from a distance of 1.4 m. A 16 × 120 pixel bar thus corresponded to 0.18º × 1.3º of visual angle. (Stimulus sizes will be reported in pixels instead of visual angle because viewing distance was not enforced. The scale factor for visual angle at a viewing distance of 1.4 m is 0.011 degrees per pixel.) Outlined by 1-pixel-wide black lines, the interiors of the bars were filled with moderately varied hues of blue and brown. Color variety was provided partly for visual interest and partly to enhance bar-to-bar individuality. For those participants with experience in interpreting histograms, the varied color served as a reminder that these were clusters of individual bars, not summaries of data in the form of histograms. The bars were presented against a light gray background. The experiment was run with an in-house program compiled using Microsoft Visual Basic 6.

Procedure

Participants were tested individually, either in the lab or at home. Each participant tested in the lab sat on a sofa, with the computer placed on a coffee table 1.4 m away and with an optical mouse placed at his or her right or left side. For those tested at home, either the lab arrangement was replicated or the participants were tested at a dining room table. Illumination was not controlled beyond ensuring that indoor lighting was uniform and that the display was clearly visible.

Following instructions, demonstrations, and a little practice until they declared themselves ready, participants supplied demographic data on age, gender, handedness, and formal statistical training. Testing comprised two tasks: mean estimation and difference judgment. Mean estimation was a preliminary 2-min task used to ensure that the participants had a working understanding of the term *average*. They were shown only one group at a time of vertical bars of varied height that were horizontally centered and rose from the bottom edge of a 250 × 250 pixel frame (Figure 2). The number of bars in each group was selected randomly from a flat distribution between 4 and 13. The heights of the bars were selected randomly from a normal distribution ($\mu = 120$, $\sigma = 20$ pixels, range $\pm 5\sigma$).

The task was similar to Spencer's adjustment of a cursor to the estimated average of graphically displayed data. Participants in the present study had 5 sec to drag a horizontal line vertically and to register, by clicking the mouse button, the position corresponding to what they judged to be the average height of the bars. When the mouse button was pressed, the vertical position of the cursor was recorded, along with the heights of the displayed bars. One half-second later, a new set of bars was displayed for the next trial. A display was replaced with a new set of bars if no answer was given within the 5-sec time limit. This procedure ran until 25 estimates were obtained.

The main task was difference judgment. On each of 250 trials, participants were shown two sets of vertical bars centered in the left and right halves of a 500 × 250 pixel frame for 1 sec (Figure 1). Five response panels with labels "Left Bars Definitely Taller," "Left Bars Possibly Taller," "Too Close to Say," "Right Bars Possibly Taller," and "Right Bars Definitely Taller" were aligned horizontally below the bars. The response panels were always in view. After each presentation, the participant had unlimited time to click on one of the response panels using the mouse-controlled cursor. An integer between −2 and +2, inclusive, was recorded as the response, corresponding, respectively to the labels above, along with the heights of the bars. The participant could also indicate having missed the stimulus or that he or she wanted the just-registered response to be deleted. The next trial's bars were displayed 0.5 ec after the response. Halfway into the session, after Trial 125, the participant was invited to rest. Sessions took about 20 min to complete.

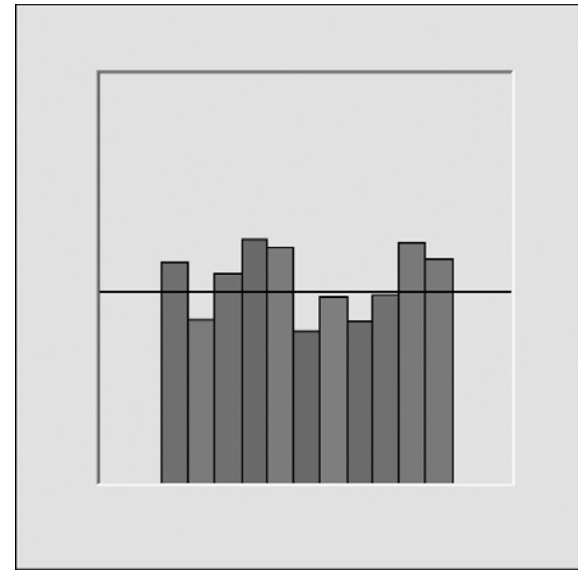

Figure 2. Mean estimation task. Used to ensure that participants appreciated the meaning of "average," the task was to align the horizontal line with the average bar height.

Independent random variables. The values of three variables were chosen pseudorandomly by the software: (1) the number of bars assigned to the left and right, (2) the individual heights of the bars, and (3) the difference in mean height as defined by "Student's" (1908) *t* ratio. The number of bars was chosen separately per side from a flat probability distribution between 2 and 14. To keep group sizes similar, the absolute difference in bar number was constrained to be within 25% of the total number of bars. Bar height was initially assigned to each bar as a random normal deviate ranging between −5 and +5 inclusive, multiplied by 20 pixels, and added to 120 pixels. A randomly selected *t* value from a rectangular distribution of ±3.1 was chosen by the software. A constant number of pixels was then added to or subtracted from each bar height in one of the groups to make the bar heights conform to the selected *t* value.[1] The two sets of bars were then randomly assigned to either the left or right side of the display.

RESULTS

The data from 6 of the original 34 volunteers were dropped from analysis, 2 for failing the mean estimation screen and 4 for responding too timidly in the face of extreme differences. In mean estimation, a judgment was accepted if its height was somewhere between the tallest and shortest members of the set. Participants were allowed a couple of slips, so only 23 of the 25 answers had to be within range. Two participants who scored only 10 and 16 within-range answers out of 25 were excluded by this criterion. In difference judgments, we wanted participants to choose "definitely" answers at least when the two sets did not overlap. Trials in which none of the bars on one side were taller than the shortest bar on the other were examined to see whether the response was definite. Even

then, we insisted on definite judgments on only a simple majority of such trials (median = 7 trials administered per participant). The data from 4 participants were dropped for answers no stronger than "possibly" on at least half of their extreme-difference trials.

The results of the mean estimation task are not reported in detail because of a suspected artifact. Participants' responses—the cursor heights—seemed to underestimate the arithmetic mean by approximately 4 pixels (3%). Because the response was registered by clicking the mouse button, and because the optical mouse was often at the participant's side on the seat of a sofa, we cannot rule out the possibility that the underestimate was an artifact of slight downward motion of the mouse when the responses were made. Accordingly, the data from our mean estimation task cannot be used to infer the sort of mean (arithmetic, geometric, etc.) that corresponds best to the subjective average. For this we rely instead on Spencer's (1963) and on Bauer's (2006) findings that the arithmetic average corresponds best.

Each participant contributed 250 difference judgments. A median of 2.5 trials (1%) were marked as having been missed by the participants. At the high end, 11, 11, 12 and 20 (the 70-year-old) trials were marked as having been missed. The median number of trials deleted by the participants was 0. Six participants deleted 1 of their trials, 1 participant deleted 4, and 1 participant deleted 8. When trials were missed or deleted, new trials were administered in order to bring the total to 250.

Multiple Regression Analyses

The participants' judgments, recorded as integers from −2 to +2 inclusive, were the sole dependent variable. Measures calculated from the heights of the displayed bars were treated as independent variables. Their roles were examined in three multiple regression analyses, each performed on 7,000 judgments (n = 28 participants × 250 trials). The first analysis was conducted on elementary statistical measures alone. The second analysis was a moderation analysis to help define the nature of the internal statistical judgment. In the third analysis, statistical measures were combined with factors related to the appearance of the bar groups.

Elementary statistical measures. Basic statistical computations derived directly from the displayed heights were submitted as independent variables in the first stepwise multiple regression. The six measures were $\Delta\mu$, ΔN, $\Delta\sigma$, $\Sigma\mu$, ΣN, and $\Sigma\sigma$, where Δ stands for a right-side minus left-side difference in the measure and Σ stands for the sum of right- and left-hand sides. N is the number of bars, μ represents the group mean height, and σ stands for the standard deviation of group height. As can be seen from the correlation matrix of Table 1, these six measures were generally independent of each other. Of the 15 correlations, 14 were close to 0. The exception of a +.30 correlation between ΣN and $\Sigma\sigma$ may have been an artifact of the stimulus selection process: Drawing from a flat distribution of t values between −3.1 and +3.1 inclusive, a stimulus set would be rejected if its bars did not fit within the limits of the display area. Thus, the selection process may have exerted a bias against the combination of a large absolute mean difference, a large variance, and a small number. Since this correlation was well below that at which collinearity begins to be a problem (r = .7–.8; Meyers, Gamst, & Guarino, 2006), it was not taken as a threat to the interpretation of the results. The results of the regression analysis are summarized in Table 2. Four of the measures were found to be significant predictors of participants' responses. The most important factor was $\Delta\mu$, accounting for 60% of the variance. Corresponding precisely to the task demanded of the participant, $\Delta\mu$'s prominence was expected. Highly significant contributions of ΔN and $\Delta\sigma$ were also detected, but, by accounting for a little more and a little less than 1% of the variance, respectively, these variables seem to play minor roles in comparison with $\Delta\mu$. The overall height ($\Sigma\mu$) was found to be marginally significant ($p < .05$). Its marginal significance combined with its accounting for less than

Table 1
Correlation Matrix of Predictors: Statistical Elements

Variable	M	SD	$\Delta\mu$	$\Sigma\mu$	$\Delta\sigma$	$\Sigma\sigma$	ΔN
$\Delta\mu$	−0.46	19.2					
$\Sigma\mu$	240	24.0	.01				
$\Delta\sigma$	0.13	7.27	.01	−.03*			
$\Sigma\sigma$	35.5	7.85	−.01	.01	.00		
ΔN	0.01	2.37	−.01	−.02	.07*	−.01	
ΣN	17.6	6.02	.02	−.02	.00	.30*	−.01

*$p < .001$.

Table 2
Stepwise Multiple Regression: Statistical Elements

Variable	B	SE	β	ΔR^2
$\Delta\mu$	059	.001	.778**	.601
ΔN	.079	.005	.129**	.016
$\Delta\sigma$	−.014	.001	−.071**	.005
$\Sigma\mu$	−.001	.000	−.017*	.000

Note—ΣN and $\Sigma\sigma$ did not meet the entry criterion ($p_{IN} < .05$). *$p < .05$. **$p < .001$.

Table 3
Moderation Analyses of Statistical Elements in Interaction With $\Delta\mu$

Step	Variable	B	SE	β	ΔR^2
Step 1	$\Delta\mu$	.059	.001	.776**	
	ΣN	.000	.002	.002	.601**
Step 2	$\Delta\mu \cdot \Sigma N$	.001	.000	.275**	.013**
Step 1	$\Delta\mu$	.059	.001	.776**	
	$\Sigma\sigma$	.002	.001	.010	.602**
Step 2	$\Delta\mu \cdot \Sigma\sigma$	−.001	.000	−.509**	.012**
Step 1	$\Delta\mu$	.059	.001	.776**	
	$\Sigma\mu$	−.001	.000	−.017*	.602**
Step 2	$\Delta\mu \cdot \Sigma\mu$	.000	.000	−.124	.000*
Step 1	$\Delta\mu$	.059	.001	.777**	
	ΔN	.076	.005	.124**	.617**
Step 2	$\Delta\mu \cdot \Delta N$	.000	.000	.009	.000
Step 1	$\Delta\mu$	.059	.001	.776**	
	$\Delta\sigma$	−.012	.002	−.061**	.605**
Step 2	$\Delta\mu \cdot \Delta\sigma$	.019	.009	.017*	.000*

*$p < .05$. **$p < .001$.

0.1% of the variance suggests that it was an unimportant factor in guiding participants' responses. As contributors on their own, the sum of the left and right standard deviations ($\Sigma\sigma$) and the total number of bars (ΣN) failed to meet the entry criterion ($p_{IN} < .05$).

Moderation analysis. The purpose of the moderation analyses was to see whether any of the statistical measures with null effects, given the scaling used, were contributing to the responses. For example, if overall variance were to reduce apparent differences, responses would be pushed toward 0 from positive differences above and from negative ones below. Thus, the slope of the response–$\Delta\mu$ function might rotate with changes in variance without an appreciable difference in its general elevation.

The elementary statistical measures other than $\Delta\mu$ were tested for status as moderators of $\Delta\mu$ in five analyses (Table 3). Each analysis first examined $\Delta\mu$ with one of the other five elementary measures as separate predictors, then examined each pair of predictors with its interaction product. Both ΣN and $\Sigma\sigma$, neither of which was found to be significant on its own, were found to be highly significant factors in interaction with $\Delta\mu$. Each of these interactions added about 1% explained variance. The moderating effect of these two variables is also illustrated in Figure 3, where the data were divided into high and low ΣN groups or split into high and low $\Sigma\sigma$ halves. Separate plots of d' versus $\Delta\mu$ showed steeper discrimination functions when the total variance ($\Sigma\sigma$) was low and when the total bar number (ΣN) was high. (The d' computations are detailed below in conjunction with Figure 4.) The other three variables that were tested for moderation were either marginally significant in interaction with $\Delta\mu$—but added less than 1% to R^2 ($\Sigma\mu$, $\Delta\sigma$)—or were not significant in interaction (ΔN). As expected, d' plotted against high-versus-low halves of these marginally significant and nonsignificant variables resulted in indistinguishable line pairs (not illustrated). Thus, two variables, ΣN and $\Sigma\sigma$, were found to possess a moderating influence over $\Delta\mu$. The impact of ΣN was positive, in that greater numbers of bars strengthened discrimination, whereas that of $\Sigma\sigma$ was negative, in that greater variance overall weakened discrimination.

CR. The moderation analyses pointed to CR as a plausible representation of the implicit statistical calculation at the core of the decision about which side had the greater average. CR was the precursor to the modern *t* test. The *t* test progressively replaced CR during the second quarter of the last century (Rucci & Tweney, 1980), as researchers gradually implemented Student's (1908) solution to the problem of underestimating population variance from samples. Like *t*, CR is a ratio of the difference in means to a standard error based on both groups. It is calculated thus:

$$\mathrm{CR} = \frac{\mu_R - \mu_L}{\sqrt{\frac{\sigma_R^2}{N_R} + \frac{\sigma_L^2}{N_L}}},$$

where μ represents the mean bar height, σ^2 represents the variance, N represents the number of bars, and the subscripts R and L designate right and left sides. The positively moderating effect of ΣN on $\Delta\mu$ in the regression analysis and the steeper slope of the discrimination function at the greater group sizes (Figure 3) agree with each other in recommending that $\Delta\mu$ is enhanced by greater numbers of bars. Similarly, the negative moderation of $\Delta\mu$

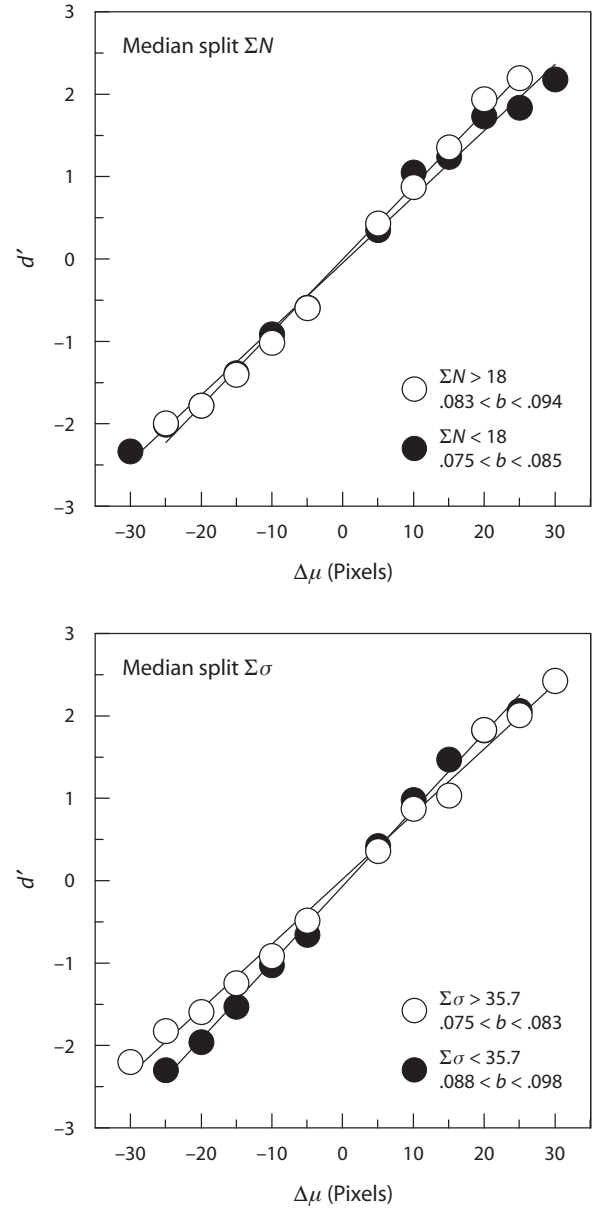

Figure 3. Discrimination as a function of number and variance. Median split of number (upper) shows a steeper d' function of $\Delta\mu$ when the total number of bars is greater than 18 than when fewer than 18 bars are displayed. Median split of the sum of left and right standard deviations, $\Sigma\sigma$ (lower), shows a steeper discrimination function when the variance is lower. Printed limits are 95% confidence intervals (CIs) around slope, *b*. Omitted for clarity, 95% CIs are approximately as big as the symbols, comparable to the CIs in Figure 4.

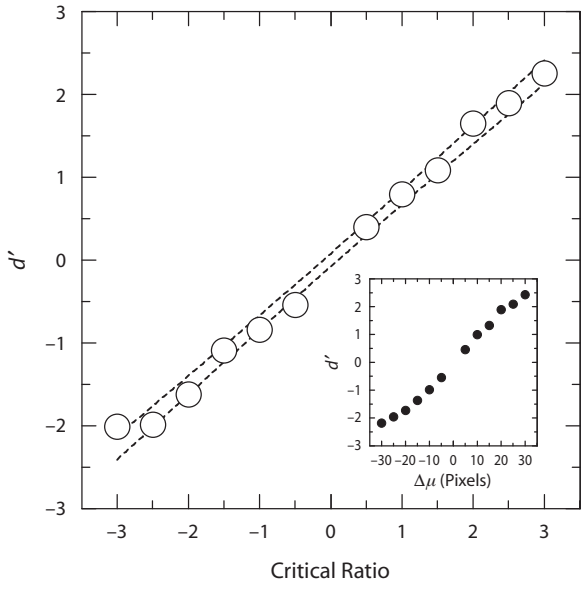

Figure 4. Discrimination plots of *d′* as a function of critical ratio, with 95% confidence corridor as dashed lines. Inset shows similar plot of *d′* as a function of the right- minus left-hand difference in mean height.

by $\Sigma\sigma$ and the steeper discrimination function at lower values of overall variability are consistent with $\Delta\mu$ being perceptually diminished by the overall variance. CR, as does Student's *t* ratio, incorporates these variables with $\Delta\mu$: CR grows with greater number and shrinks with greater variance. CR was thus included as a potential computation to represent the mental summary extracted from the display by visual cognition.

The other computations that were considered as potential representatives were the difference in means ($\Delta\mu$), Cohen's *d*, and Student's *t*. Only one measure could be submitted to the final regression because the correlations between them were close to 1. The difference in means ($\Delta\mu$) measures precisely the feature the participants were asked to judge. Cohen's *d* is sensitive to variance, but not to overall number. Between *t* and CR, which differ from each other only in whether one uses degrees of freedom or actual numbers to calculate variance, CR is the theoretically justified choice by virtue of the lack of inference demanded by the task. In testing for empirical support for the choice of CR, individual ($n = 28$) correlations were determined between raw responses and four measures calculated from 250 displays: $\Delta\mu$, a population version of Cohen's *d*, Student's *t*, and CR. Mean Pearson's correlations with responses were $\Delta\mu$, .788; Cohen's *d*, .794; Student's *t*, .803; and CR, .806. The difference between *t* and CR was not only tiny, but was also the smallest of all paired comparisons. Nevertheless, a repeated measures *t* test found it to be statistically robust [$t(27) = 6.1$, $p < 10^{-5}$]. CR thus seemed—both on theoretical grounds and with empirical justification—to be the choice to represent the internal metric, incorporating into one calculation $\Delta\mu$, the sum of standard deviations ($\Sigma\sigma$), and the total number of displayed bars (ΣN).

Main regression. CR was entered into the main stepwise regression analysis to represent the implicit statistical calculation. Because CR incorporates $\Delta\mu$, ΣN, and $\Sigma\sigma$, these elements were omitted. The difference in bar number (ΔN) and the difference in standard deviations ($\Delta\sigma$) were entered into the analysis because they were found to be highly significant factors on their own in the first regression analysis. The overall height ($\Sigma\mu$) was also entered into the final analysis because it attained marginal significance in the first analysis. The roles of apparent disorder and outliers were also examined. Apparent disorder was calculated using Lathrop's λ. Having originally developed this index to measure subjective variability when stimuli were delivered sequentially (Lathrop, 1966), Lathrop (1967) applied this index to presentations of adjacent bars viewed all at once. If tall and short bars are positioned right beside each other, the variability will seem great. If, instead, the bars are more or less in progressive order by height, with small adjacent-bar differences, the same set of bars might seem more uniform. The measure λ is the square root of the ratio of mean vertical distance between adjacent bars to the population standard deviation of the group, defined thus by Lathrop (1967):

$$\lambda = \sqrt{\frac{\sum_{i=1}^{N-1} \left| x_i - x_{i+1} \right|}{(N-1)\sigma}}.$$

Typically, λs were $M = 1.12 \pm SD = 0.12$. They were greater when short and tall bars were intermingled, but were well under 1 when a lengthy series of bars occurred in nearly perfect ascending or descending order. Finally, four variables representing outliers were included: These were the tallest and shortest bars, on the left and on the right. To properly scale these as outliers, we calculated their standard scores from the heights of bars within their own displayed sets.

Table 4 shows that, whereas many predictor correlations were highly significant, their actual values were relatively low. Correlations among the variables sensitive to appearance—Lathrop's λ and the outliers—were generally higher. As can be seen in Table 5, CR dominated the analysis by accounting for 63% of the variance on its own. The only other variable to increase the explained variance by more than 1% was ΔN. The remaining factors, with the exception of the marginally significant $\Sigma\mu$, were statistically highly reliable, but they each accounted for less than 1% of the variance. Display disorderliness, as measured by Lathrop's λ, was a member of this group, as were the tall outliers, designated *z*HighL and *z*HighR. Given the scaling of the response, disorder on the left (λL) was expected to contribute positively, whereas tall outliers on the left (*z*HighL) were expected to contribute negatively, with opposite predictions for their right-hand counterparts. Short outliers (*z*LowL, *z*LowR) were expected to perform in the direction opposite to their tall counterparts. The short outliers were not correlated well enough with responses to

Table 4
Correlation Matrix of Final Predictors

Variable	*M*	*SD*	CR	ΔN	$\Delta\sigma$	$\Sigma\mu$	λR	λL	*z*HighR	*z*HighL	*z*LowL
CR	−0.04	1.45									
ΔN	0.01	2.38	−.01								
$\Delta\sigma$	0.13	7.26	.01	.07**							
$\Sigma\mu$	240	24.0	.01	−.02*	−.03*						
λR	1.13	0.12	.00	−.07**	−.01	.01					
λL	1.12	0.12	−.01	.08**	.01	.00	.17**				
*z*HighR	1.58	0.38	.00	.18**	.02*	−.01	−.22**	−.16**			
*z*HighL	1.58	0.38	−.04**	−.16**	−.01	−.01	−.17**	−.23**	.18**		
*z*LowL	−1.58	0.38	.02	.14**	.02*	.00	.18**	.23**	−.20**	.21**	
*z*LowR	−1.58	0.37	.00	−.12**	−.02*	.01	.23**	.17**	.20**	−.19**	.20**

*$p < .05$. **$p < .001$.

be admitted into the analysis. As was the case with the regression on statistical elements, $\Sigma\mu$ attained marginal significance but evidently contributed a proportion below .0005 to the explained variance.

If d' represents psychological magnitude, and if the internal computation resembles CR, then, as Sorkin et al. (1991) have suggested, d' might be predicted from CR calculated from the displayed heights. In the main portion of Figure 4, d' is plotted as a function of CR using the response distribution (Gescheider, 1997, p. 120). The 7,000 observations were sorted into an array of 13 CR levels (−3 to +3 inclusive at 0.5 CR increments) by five response categories ("Left Bars Definitely Taller" through "Right Bars Definitely Taller"). Hits and false alarms were unweighted counts of possibly and definitely right responses for all positive CRs. The false alarm count was taken from the response bins at CR = 0 (i.e., $-0.25 \leq CR < 0.25$). Misses, therefore, were the sums of too close, possibly left, and definitely left responses. Similarly, negative CR hits and false alarms were the total number of possibly and definitely left responses taken from trials in which CR was substantially negative. Hit and false-alarm proportions were expressed in ratios to total responses at each CR level ($M = 511$ trials per CR level, $SD = 25$; 358 trials beyond $|CR| = 3.25$ were disregarded). The linear d'–CR relation accounted for .9943 of the variance. The 95% confidence corridor around the 12 points is shown as a dashed line. The inset in Figure 4 represents a similar exercise performed with the data sorted by $\Delta\mu$ instead of CR. With .9922 explained variance, the d'–$\Delta\mu$ relationship was also a good fit to a straight line. Visually, the unexplained variance around CR (Figure 4) seems to be due to random error around a straight line (note the point at CR = −3). In the case of $\Delta\mu$ (Figure 4, inset), however, the points seem to systematically follow a cubic trend. Fitting the data to a third-order polynomial helped CR a little (R^2 increased to .9973 with the quadratic and cubic components), whereas, in the case of $\Delta\mu$, the Order 3 polynomial helped a great deal ($R^2 = .9993$). Thus, whereas adding a second- and third-order term to CR explained an additional 53% of the .0057 residual variance, adding the two terms to $\Delta\mu$ accounted for nearly all (91%) of its .0078 unexplained variance.

Table 5
Stepwise Multiple Regression of Critical Ratio (CR), Statistical Elements, and Appearance Variables

Variable	*B*	*SE*	β	ΔR^2
CR	.59	.01	.80**	.631**
ΔN	.06	.00	.10**	.015**
$\Delta\sigma$	−.01	.00	−.07**	.005**
λR	−.64	.09	−.05**	.002**
λL	.58	.09	.05**	.002**
*z*HighL	−.18	.03	−.04**	.001**
*z*HighR	.17	.03	.04**	.002**
$\Sigma\mu$	.00	.00	−.02*	.000*

Note—The variables *z*LowR and *z*LowL did not meet inclusion criterion ($p_{IN} < .05$). *$p < .05$. **$p < .001$.

Individual raw data averaged into half-CR-unit bins were fitted using Campbell, Evans, and Gallistel's (1985) broken-line fit method. This is a conventional least-squares fit of growth data to (1) an optional horizontal, lower plateau; (2) a positively sloped line; and (3) an optional upper plateau, with breakpoints forced to occur at abscissa test points. Figure 5 shows mean judgments, standard errors, and fitted lines for the participant who ranked 14th (thus deemed to be typical) according to the proportion of variance explained by the fit (R^2 in Figure 5). The proportion of explained variance for these 28 participants was not correlated with any of the basic demographic measures: age, $r = -.11$, n.s.; sex, $r = .09$, n.s. (14 males); handedness, $r = .02$, n.s. (only 1 left-handed person); and statistics training, $r = .27$, n.s. (mode = one or more undergraduate courses). The central line of the broken-line fit was used to interpolate CR individually at response levels of −1 and +1, corresponding to answers at the "possibly" criterion. For bars taller on the left, 95% confidence limits for the 28 participants were $-1.72 < CR < -1.39$, and for bars taller on the right, $1.37 < CR < 1.67$. Respectively, these confidence limits would correspond to one-tailed personal probabilities of .043 to .082 and .085 to .047.

DISCUSSION

CR was the most important predictor of difference judgments, accounting for the greatest share (63%) of the variance. Thus, confidence in the decision about which group of bars has the greater average is enhanced by the total number of bars and diminished by the variance. To a first approximation, intuitive decisions about relative averages seem to follow rules of statistical inference. It is

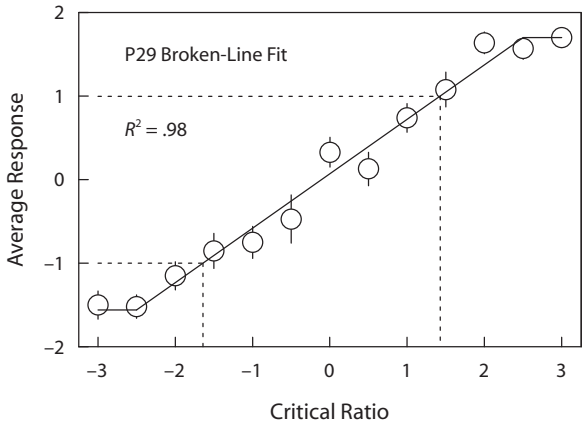

Figure 5. Raw responses from 1 participant sorted into 0.5-unit-wide critical ratio (CR) bins, averaged, and fitted to up to three line segments. Participant 29 illustrates median performance, as she ranked 14th best fit of $n = 28$. Dashed lines represent interpolated CRs corresponding to a judgment of "Possibly Taller" on the left (at ordinate $= -1$) or right ($+1$).

important to recognize that, had the data been related to the *t* statistic, the results would have been nearly identical. The correlation between *t* and CR was almost perfect, so we do not wish to place emphasis on the fact that CR was chosen over *t*; *t* would have done a fine job, had it been used. The point to emphasize is how well the general form of CR or *t* as a class accounts for statistical judgments. Having said this, we confess curiosity that, when correlated to responses, CR (mean Pearson's $r = .806$) edged out *t* (.803) by a tiny, but very consistent degree ($p < 10^{-5}$) in individual comparisons. Since the task was not an inference to hidden populations, but a decision about groups with all elements visible, CR is theoretically the appropriate choice over *t*. Could the slight superiority of CR represent empirical confirmation of a statistical processor at the heart of visual judgments of average size that distinguishes sample from population? It would be very interesting to design an experiment in which inferences from samples were pitted against judgments about groups in order to see whether *t* would be used for inferences from samples and CR for judgments about small populations.

Normatively, the correct answer for the task was $\Delta\mu$, but raw responses and regression analyses on one hand, and d' plots on the other, were consistent with a statistical personal assessment over the normative difference in means. Despite the nonstatistical nature of the task, judgments followed CR more closely than they did $\Delta\mu$. An informal inspection of the behavior of ΣN and $\Sigma\sigma$ in accounting for residue remaining after CR or $\Delta\mu$ showed $+\Sigma N$ and $-\Sigma\sigma$ regression weightings, with $\Delta\mu$ as the main factor tested, but showed opposite weightings—and smaller coefficients—with CR as the main factor. Thus, whereas CR provided a more accurate description over $\Delta\mu$, too much weight was accorded to ΣN and $\Sigma\sigma$ by the actual computation of CR; its computation should be adjusted a little in the direction of $\Delta\mu$. Both Legge et al. (1989) and Sorkin et al. (1991) came to the same conclusion about number—it is a factor, but not quite by the $n^{1/2}$ recommended by statistical orthodoxy.

Why would a statistical evaluation underlie the decision when a simple difference gives the right answer? The question invites speculation: As illustrated by examples cited by Chong and Treisman (2003, 2005a, 2005b), the natural world contains sets of varied sizes of the same object (flowers, exposed river stones, animal herds, etc.). Natural sets of identical dimension are relatively uncommon. Assuming a selection pressure to make accurate decisions about where better to forage or to hunt, and granting that life forms are characterized by temporal fluctuation (e.g., daily, seasonal, annual), identifying where to find the richer resource seems better suited to an inferential process that takes variability and number into account. In other words, since the natural world has both static and dynamic variability, trustworthy knowledge is the statistical generality, not the particular instance of which herd has smaller members or of which patch has larger berries.

After CR, the statistical variable to contribute something to the decisions was ΔN, the difference in the number of bars. Increasing explained variance by only 1.5%, ΔN seems to be a minor contributor. Recall, however, that ΔN was restricted in range: Differences in number had to be within 25% of the total number of bars, so differences ranged from 0 to 5. Had ΔN's range been unrestricted, it undoubtedly would have yielded a better result. Nevertheless, although its contribution was shortchanged by half, its explained variance still falls more than one order of magnitude below CR's explained variance. It is difficult to determine whether the contribution of ΔN represents an artifact of the display or a general tendency among participants to choose the greater number. Since the bars in our display had definite width, greater numbers occupied greater areas. The artifact might be the error of partially confounding greater area with greater average height. On the other hand, perhaps there is some ecological sense to confounding greater number with greater size: In choosing which patch of berries to approach, a large number of ordinary berries in one patch may be just as attractive as a few large berries in another. Without studies deliberately designed to test ΔN's role, we defer dismissing it as a stimulus artifact or treating it as a variable with underlying significance.

The appearance variables (numbered 5–10 in Table 4) made a weak showing. Their small role is surprising because these visual features are readily seen; note the salience of the short outlier in Figure 1. Lathrop's λ coefficient was significant for both sides. Disorder on the right was signed negatively, whereas disorder on the left exhibited a positive slope. Thus, disorder, like $\Delta\sigma$, acts to obscure the perceived difference in average height. Tall outliers (*z*HighL and *z*HighR) were significant variables, but short outliers were not found to be significant. Tall outliers on the right contributed positively, whereas tall outliers on the left were negatively signed; given the scaling, this meant that tall outliers added to the strength of the response. The overall weak contribution of appearance variables as a group—combined, they explained less than 1% of the variance—suggests that the process that evalu-

ates averages and differences in averages is effective at seeing through display noise.

Snap judgments about differences in average size thus seem to depend on an imperfect statistical process. It is statistical because of its sensitivity to number and variance. It is imperfect for two reasons: The perceptual magnitude does not grow in precise measure with the square root of the number of elements, nor does it shrink in ideal correspondence to standard error; and it tends to confound greater number with greater size. Nevertheless, the process is described better by CR, and almost equally well by t, than it is by $\Delta\mu$. Is this statistical process sensory, perceptual, or cognitive? In their review of statistical reasoning about verbally presented problems, Sedlmeier and Gigerenzer (1997) concluded that people accurately integrate the law of large numbers in making an intuitive assessment about the probable value of one sample, but that they are inaccurate when it comes to assessing the sampling fluctuations of an average. Failure to make use of sample size and variance continues to be seen when verbal reasoning is probed (Obrecht, Chapman, & Gelman, 2007). The fidelity to CR demonstrated by judgments of group differences in our data show that sampling dispersion *is* inherently understood by the mechanisms underlying these visual assessments. Rosenholtz (2000) has recently modeled the detection of boundaries between different visual textures using statistical algorithms. Her simulations seem adept at locating transitions in pattern in quilt-like visual stimuli. Perhaps statistical algorithms are employed widely in early stages of perceptual processing; it will be interesting to extend the present findings to other visual judgments, and to other sensory channels as well. For now, we take the view that the accurate use of statistical methods takes place early in the sensory–perceptual–cognitive chain, yielding statistical summaries that percolate up to later stages.

What does it take to get a person to declare a possible difference? The analysis illustrated in Figure 5, in which CR was interpolated using response criteria of −1 (possibly left) and +1 (possibly right), yielded 95% confidence limits of 1.5 to 1.7 around CR, corresponding to a one-tailed probability of .04 to .08. In guiding the decision about magnitude judgments in the face of noisy information, the visual system evidently uses a criterion that matches the minimal level (i.e., $p < .05$) customarily used in psychological research. It is not quite tempting enough to wonder whether this is more than an amusing coincidence.

AUTHOR NOTE

We thank Ben Bauer of Trent University and Ruth Rosenholtz at MIT for comments on this work and for their stimulating correspondence. We also appreciate the time and patience devoted to earlier versions of this article by the anonymous reviewers. Correspondence concerning this article should be addressed to G. Fouriezos, School of Psychology, University of Ottawa, 145 Rue Jean-Jacques Lussier, Ottawa, ON, K1N 6N5 Canada (e-mail: georgef@uottawa.ca).

REFERENCES

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

BAUER, B. (2006, June). *Which mean do we mean? Rapid extraction of summary values from visual displays*. Paper presented at the 16th Annual Meeting of the Canadian Society for Brain, Behaviour, & Cognitive Science, Saskatoon, Saskatchewan, Canada.

CAMPBELL, K. A., EVANS, G., & GALLISTEL, C. R. (1985). A microcomputer-based method for physiologically interpretable measurement of the rewarding efficacy of brain stimulation. *Physiology & Behavior*, **35**, 395-403.

CHONG, S. C., & TREISMAN, A. (2003). Representation of statistical properties. *Vision Research*, **43**, 393-404.

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

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

GESCHEIDER, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Mahwah, NJ: Erlbaum.

LATHROP, R. G. (1966). First-order response dependencies at a differential brightness threshold. *Journal of Experimental Psychology*, **72**, 120-124.

LATHROP, R. G. (1967). Perceived variability. *Journal of Experimental Psychology*, **73**, 498-502.

LEGGE, G. E., GU, Y., & LUEBKER, A. (1989). Efficiency of graphical perception. *Perception & Psychophysics*, **46**, 365-374.

MEYERS, L. S., GAMST, G., & GUARINO, A. J. (2006). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage.

OBRECHT, N. A., CHAPMAN, G. B., & GELMAN, R. (2007). Intuitive *t* tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, **14**, 1147-1152.

ROSENHOLTZ, R. (2000). Significantly different textures: A computational model of pre-attentive texture segmentation. In D. Vernon (Ed.), *Proceedings of the 6th European Conference on Computer Vision* (pp. 197-211). Berlin: Springer.

RUCCI, A. J., & TWENEY, R. D. (1980). Analysis of variance and the "second discipline" of scientific psychology: A historical account. *Psychological Bulletin*, **87**, 166-184.

SEDLMEIER, P., & GIGERENZER, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, **10**, 33-51.

SORKIN, R. D., MABRY, T. R., WELDON, M. S., & ELVERS, G. (1991). Integration of information from multiple element displays. *Organizational Behavior & Human Decision Processes*, **49**, 167-187.

SPENCER, J. (1961). Estimating averages. *Ergonomics*, **4**, 317-328.

SPENCER, J. (1963). A further study of estimating averages. *Ergonomics*, **6**, 255-265.

"STUDENT" [GOSSET, W. S.] (1908). The probable error of a mean. *Biometrika*, **6**, 1-25.

NOTE

1. Since inferences to parent populations were not requested of the participants, the z distribution, as the software's display engine, would have been more in keeping with the task than was the t distribution. But because this was just a mechanism to ensure that larger differences were frequently displayed, we decided that it would make no difference to the participants' judgments whether t or z drove the relative frequencies of displayed differences.

(Manuscript received May 31, 2006;
revision accepted for publication October 1, 2007.)